Avoiding Calvinist Decision Traps using Structural Causal Models

Arvind Raghavan
Department of Computer Science
Columbia University
arvind.raghavan@columbia.edu

Abstract

Causal Decision Theory (CDT) is a popular choice among practical decision theorists. While its successes and failings have been extensively studied, a less investigated topic is how CDT's choices hinge on the theory of causation used. The most common interpretation, *temporal* CDT, understands causation as a description of physical processes ordered in time. Another emerging view comes from the graphical framework of Structural Causal Models (SCM), which sees causation in terms of constraints on sources of variation in a system. We present an adversarial scheme where a CDT agent facing a Bandit problem can be tricked into sub-optimal choices, if it follows *temporal* CDT. We then propose an axiom to ground the orientation of arrows in the causal graph of a decision problem. In doing so, we resolve an ambiguity in the theory of SCMs, and underscore the importance of agent-perspectives, which have been largely ignored in the causal inference literature. We also demonstrate how this *structural* CDT avoids our adversarial trap, and outperforms *temporal* CDT in a series of canonical decision problems.

1 Introduction

Normative decision theory seeks to axiomatically ground the behaviour of rational agents. von Neumann and Morgenstern [1947] and Savage [1954] laid the foundations with representation theorems for agent preference and utility functions, albeit with notable flaws (see Ahmed [2014, Ch. 2]). Causal Decision Theory (CDT) emerged to address these flaws, roughly stipulating that the partition of world-states an agent analyzes needs to be causally independent of the agent's decisions. Popular expositions include Lewis [1981] and Joyce [1999]. In parallel, Jeffrey [1983] introduced Evidential Decision Theory (EDT) which sought to generalize the causal requirement using the notion of the *news value* of *propositions*. While important (and criticized), EDT is out of scope of the current work. CDT is arguably the dominant choice among practitioners such as economists and statisticians, often implicitly (e.g., Imbens and Rubin [2015]).

Unfortunately, CDT stands accused of failing at a class of problems, notably Newcomblike problems (Egan [2007],Soares and Fallenstein [2015]), and being open to money-pumps (Oesterheld and Conitzer [2021b]). These failures gain poignancy in the context of AI agents, as design choices regarding agent identity, memory and beliefs pose new challenges and risks (Conitzer [2019], Russell and Norvig [2010, Sec. 2.4]). An agent's decision theory could open it to adversarial attacks or unintended consequences, while the right theory could promote multi-agent cooperation and moral behaviour in distributed, autonomous systems (Greene et al. [2016], Conitzer et al. [2017]).

While several fixes have been proposed, a less extensively studied question is the definition of causality CDT uses. The theory is technically dense, so we isolate the feature of current interest: physical time. Prominent versions of CDT accept that event A temporally preceding B is sufficient to conclude A is causally independent of B. For simplicity, we refer to this as *temporal* causality. Independent

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

from decision theory, in the field of statistical causal inference, the framework of Structural Causal Models (SCM) and causal graphs were developed by the groups of Judea Pearl (Pearl [2009]) and Peter Spirtes (Spirtes et al. [2000]), where arrows in a causal graph represent chains of causation from graphical ancestors to descendants. We'll call this *structural* causality.

Usually, there is no conflict between *temporal* and *structural* causality. The equations in an SCM represent Nature's generative process for how the variables are realized. There is, however, an ambiguity in the formalism: what is the basis for the asymmetry of the arrows in the causal graph? We propose the **exogeneity axiom** (Sec. 3.1) to explicitly ground this asymmetry in interventional constraints rather than temporal ordering. The upshot is that the causal graphs for alleged failure modes now yield optimal results for a CDT agent. Our analysis also underscores that **causal graphs depend on which agent's perspective we take**, an important topic whose neglect in the causal inference literature is worrying given the recent runaway popularity of causality in ML. *Not* using this axiom, and relying on temporal ordering to define causal direction may cause AI agents to behave sub-optimally. Worse, it opens agents to adversarial attacks of the kind we detail next.

2 Bandits to Calvinist Bandits - an Adversarial Scheme

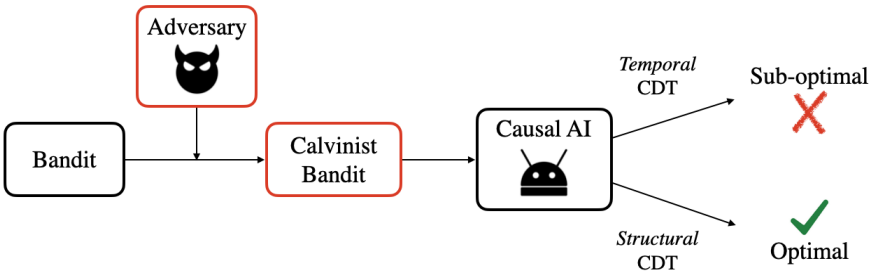


Figure 1: Causal AI agents who use the *temporal* notion of causality are vulnerable to adversarial attacks like the *Calvinist Bandit* problem

Consider a causal AI agent ("agent") operating in the world. At time $\tau = t_0$, it would originally have been offered a Bernoulli Bandit problem having the payout structure in Table 1 (top). $X \in \{0, 1\}$ represents the choice of arms, and Y is the Bandit payout. The numbers are illustrative.

Unknown to the agent, an adversary intercepts the original offer, and modifies it into a *Calvinist Bandit* offer as follows:

- At $\tau = t_0 - 1$, $S \in \{0, 1\}$ was an indicator variable of whether a statement about the world state (agent state + environment) is True.
- At $\tau = t_0$, agent is offered the *Calvinist Bandit* problem.
- At $\tau = t_0 + 1$, agent is told, it will receive payout according to Table 1 (bottom). In reality, it will continue to receive payout according to Table 1 (top).
- At $\tau = t_0 - 1$, $S = 1$ iff the state of the world is such that the agent would choose $X = 1$, if presented with the *Calvinist Bandit* problem at $\tau = t_0$.

We assume that the agent first encounters the *Calvinist Bandit* problem at time $\tau = t_0$, and this step contains the all time needed by the agent to evaluate and choose an arm.

We make three observations. First, the adversary is not actually changing the original payout structure per arm, just what the

Original Bandit

	$E[Y\|X]$
$X = 0$	1,000,000
$X = 1$	1,000

Calvinist Bandit

	$E[Y\|X, S]$	
	$S = 0$	$S = 1$
$X = 0$	1,000,000	0
$X = 1$	1,001,000	1,000

Table 1: (top) Expected payout for the original Bandit; (bottom) Expected payout for the adversarially modified *Calvinist Bandit*.

agent is told. The agent cannot detect the modification, because the payout it receives will be consistent with the bottom table (only the diagonal cells are ever realized). Second, it is not claimed that S is actually being observed by anyone. For a sufficiently advanced AI, it might not even be possible to observe S, only to infer it in retrospect. Third, S is nonetheless well-defined at $\tau = t_0 - 1$. Since AI agents make decisions algorithmically based on internal (agent) and external (environment) states of the world, $S = 0$ or 1 at $\tau = t_0 - 1$. This holds even if the agent decides based on a pseudo-random number generator. Pseudo-randomness is still an algorithmic process whose eventual outcome is well-defined (albeit virtually unpredictable to anyone) at $\tau = t_0 - 1$.

2.1 Temporal CDT Solution

We use the definition of CDT in Lewis [1981], which says a rational agent must maximize *expected utility* (EU) of any option $X = x$, measured over a valid $K-$*partition*. A $K-$partition is a partitioning of the state space into the possibilities $\{K = k_1, ..., K = k_N\}$, and the variable K is causally independent of X. Here, $S \in \{0, 1\}$ forms a valid $K-$partition, because the variable S is well-defined and already realized before the choice of X is made (although we don't know the value). Temporal precedence is sufficient to conclude causal independence as per *temporal* causality.

Let us use $P(a)$ in general as shorthand for $P(A = a)$ and $P(a_1)$ for $P(A = 1)$ for any variable A. Now, comparing EUs of our options $X = 0, 1$:

$$\begin{aligned} EU(x_0) &= P(s_1)U(x_0, s_1) + P(s_0)U(x_0, s_0) \\ &= (1 - P(s_0))(0) + P(s_0)(1,000,000) \\ EU(x_1) &= P(s_1)U(x_1, s_1) + P(s_0)U(x_1, s_0) \\ &= (1 - P(s_0))(1,000) + P(s_0)(1,001,000) \\ EU(x_1) - EU(x_0) &= (1 - P(s_0))(1,000) + P(s_0)(1,000) > 0 \end{aligned}$$

Had the agent just received the original Bandit offer in Table 1 (top), it would obviously have chosen $X = 0$. However, the agent's decision theory now causes it to switch to $X = 1$ for the *Calvinist Bandit* offer, because it evaluates the EU of $X = 1$ to be higher. Expected payout is only $1,000$.

Next, we analyse the problem through causal graphs and introduce the axiom needed to resolve this.

3 Asymmetry in Structural Causal Models

In this section, we show how the construction of the causal graph for the problem does not rely on temporal ordering. First, we briefly review the formalism of SCMs. For more detail, refer to Appendix A. For a full treatment see Bareinboim et al. [2022].

Structural Causal Model (SCM): An SCM, M, is defined as a tuple $\langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \mathcal{F} \rangle$ where $\mathbf{V}$ is the set of observable variables being studied, $\mathbf{U}$ is the set of "noise" variables exogenous to the current system, having have a distribution $P(\mathbf{u})$, and $\mathcal{F}$ is the set of functions that represent causal mechanisms determining the values of variables in $\mathbf{V}$.

Causal graphs: Each SCM M induces a graph G where edges are drawn according to the functions in $\mathcal{F}$. Every such G happens to obey the axioms of a Causal Bayesian Network (Markovian factorization, screening by parents etc. - see Appendix A). These axioms beget other powerful results like the *do-calculus* [Pearl, 2009, Sec. 3.4], transportability, counterfactual identification etc.

However, the SCM formalism skirts over one pivotal issue: what is the basis of the asymmetry of arrows? Pearl and Mackenzie [2018, Ch. 1] ascribe this to a primitive of one variable "listening to" another for its generation. Bareinboim et al. [2022, Sec. 1.2] attribute this to the "data-generating process according to which Nature assigns values to the endogenous variables". These ambiguous descriptions could well be understood as *temporal* causality - if one variable physically precedes another, it is causally independent of the latter. We propose that this formalism is better grounded in *agent-dependent interventional constraints on sources of variation.*

Figure 2: (left) Crowing is causally-independent of Sunrise; (right) Sunrise is causally-independent of Crowing. What is the basis for orienting this arrow?

3.1 Exogenous Interventions

***do*-operator:** The SCM literature uses the notation of $do(A = a)$ to represent an (atomic) intervention on a variable A. This action replaces the generative function $f_A \in \mathcal{F}$ in the SCM with a constant value $A \leftarrow a$. Semantically, we are overriding the natural source of variation that determines A and replacing it with an exogenously fixed source (here, a constant value). Graphically, we cut incoming arrows into A. Note that this includes imagined interventions like $do(\text{Sunrise} = \text{Yes})$ in Figure 3.

Exogeneity assumption: The agent is modeled as being external to the SCM. Even if the SCM describes the agent's own attributes, the agent is modeled as capable of studying it "from the outside". Crucially, any source of variation used for an intervention (such as a constant value) is modeled as un-caused by variables in the SCM. This principle appears in the literature under the slogans "*no probability for acts*" in Spohn [1977] and "*deliberation crowds out prediction*" in Levi [1990].

As a consequence, we propose that the **SCM be regarded as an agent-specific object, and not a universal one**. This is because the sources of variation available for enacting a *do*( ) intervention may have different constraints for different agents. We propose the below axiom to operationalize this.

Exogeneity axiom: In the causal graph for an agent's SCM, for each observable node $V \in \mathbf{V}$, there exists a source of variation I such that (i) I is available to the agent (possibly hypothetically), (ii) I can be used by the agent to exogenously intervene and fix the value of V as $do(V = I)$, and (iii) I is statistically independent of the graphical non-descendants of V.

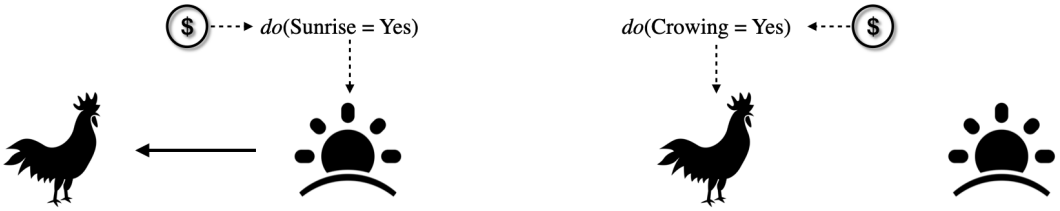


Figure 3: Exogeneity axiom is satisfied only in Figure 2 (right). Any source of variation (e.g. a coin toss) used to *do*(Sunrise=Yes) will be highly correlated with Crowing. Therefore, Crowing cannot be a non-descendant of Sunrise. Notice, no mention of physical time!

The intuition here is that if an agent views itself as exogenously intervening on a node, it needs to be able to override any effect of the node's graphical ancestors. In practice, this is achieved by methods like Fisherian *randomized controlled trials* [Fisher, 1935], or simply using a coin-toss to fix the node value in an experiment. If the agent can access no real/hypothetical source of variation, even a coin-toss, that is independent of a node's graphical ancestors, then the SCM is not valid for that particular agent because a *do*( ) intervention is not well-defined.

Similar observations appear strongly in Woodward [2003, Sec 3.1] and obliquely in Meek and Glymour [1994]. Our contribution is first to highlight that this interventional constraint can itself be used as the **basis for orienting edges** in the graph, without relying on temporal ordering to break symmetry. Second, we importantly highlight the **agent-centricity** of this axiom. A source of variation might satisfy the exogeneity axiom for one agent but not for another, resulting in different SCMs and causal graphs for the same problem, depending on which agent's perspective we take.

4 Drawing the Causal Graph for a Calvinist Bandit

In this section, we apply the exogeneity axiom in Sec. 3.1 to systematically construct the correct causal graph for the *Calvinist Bandit* problem in Section 2. The problem has 3 variables: (i) agent's choice of Bandit arm, X, (ii) agent's predisposition, indicated by S, and (iii) the Bandit payout, Y.

Hopefully, it is clear why there is a path $X \to ... \to Y$ and a path $S \to ... \to Y$. Exogenously fixing Payout as $do(Y = y)$ does not affect the distribution of Arm choice or Predisposition, whereas the actions $do(X = x)$ or $do(S = s)$ would constrain Y. This gives us the possible graphs in Figure 4.

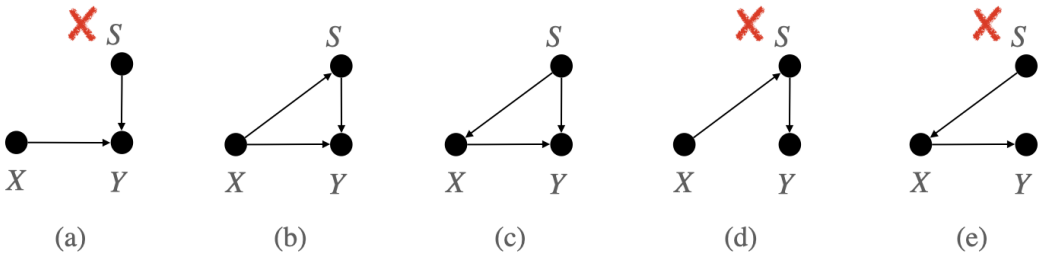


Figure 4: Possible causal graphs for the *Calvinist Bandit* problem. X : choice of Bandit arm, S : agent predisposition, Y : Bandit payout

Next, we use *d-separation* to reject options (a), (d) and (e). For details, see Pearl [2009, Sec. 1.2.3]. In option (a), we see $(X \perp S)$. In option (d), we see$(Y \perp X|S)$. In option (e), we see $(Y \perp S|X)$. These statements contradict the problem set-up. This leaves options (b) and (c).

Suppose the agent models an intervention $do(X = x)$. Since X is a binary choice, any source of variation used for this intervention can be parameterized as a Bernoulli variable (say, a coin-toss), with bias p. $p \in \{0, 1\}$ means the agent decides to pick either arm, while $p \in (0, 1)$ means the agent decides to randomize. Following the exogeneity axiom, we ask if the coin-toss is independent of the graphical ancestors of X. If the agent chooses $p = 0$ (or 1), it does so in a world where $S = 0$ (or 1). If it chooses $p \in (0, 1)$, the realized outcome is 1 only in a world where $S = 1$. Since the coin-toss is not independent of S, S cannot be an ancestor of X, and **only option (b) works from the perspective of the agent**.

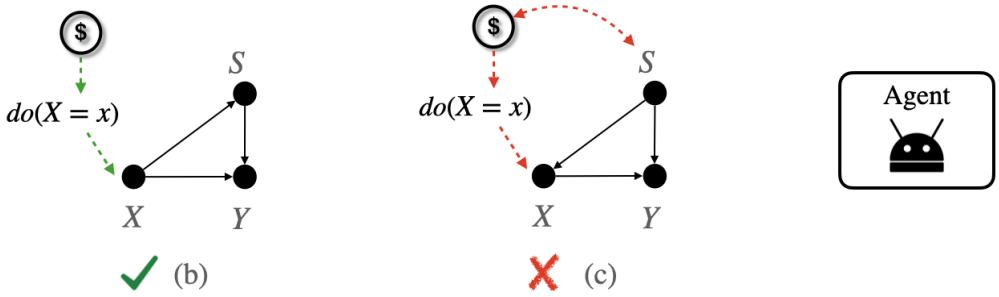


Figure 5: From the perspective of the agent in the *Calvinist Bandit* problem, only option (b) satisfies the exogeneity axiom.

From the perspective of a neutral observer of this problem, the arrow is flipped. If an observer chooses a Bernoulli parameter p for the imaginary intervention $do(X = x)$, this choice in no way constrains S. But if the observer wants to model $do(S = s)$, the agent choice X will be highly correlated with the parameter p that determines S. So X cannot be an ancestor of S and **only option (c) works from the perspective of an observer**.

We have thus achieved a valid description of the *Calvinist Bandit* problem with the causal graphs in Figure 5 and 6, without recourse to temporal ordering. In fact, the arrow $X \to S$ in option (b) runs counter to physical time! In so doing, it directly contradicts temporal arrow-orientation principles such as Pearl [2009, Def. 2.7.4].

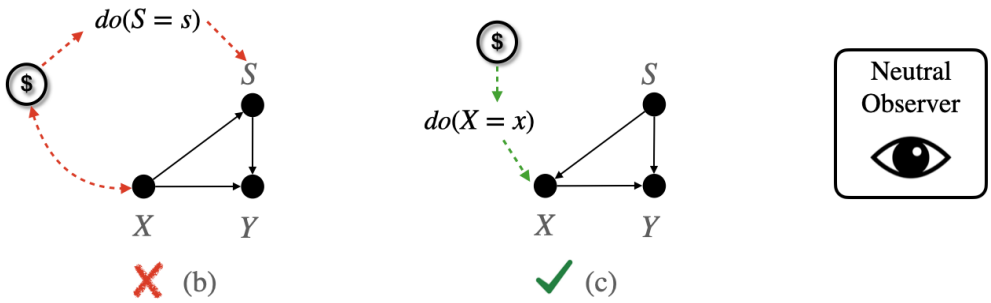


Figure 6: From the perspective of a neutral observer of the *Calvinist Bandit* problem, only option (c) satisfies the exogeneity axiom.

4.1 Structural CDT Solution

We use the definition of CDT in Pearl [2009, Sec. 4.1.1], which says an agent should maximize expected utility (EU) over different choices of interventions $do(X = x)$ in the correct SCM for the problem. Since the agent is the decision-maker, only option (b) in Figure 4 is valid.

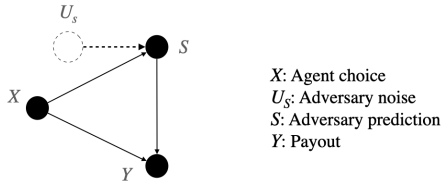
$$\begin{aligned} EU(x_0) &= P(s_1|do(x_0))U(x_0, s_1) + P(s_0|do(x_0))U(x_0, s_0) \\ &= (0)(0) + (1)(1,000,000) \\ EU(x_1) &= P(s_1|do(x_1))U(x_1, s_1) + P(s_0|do(x_1))U(x_1, s_0) \\ &= (1)(1,000) + P(0)(1,001,000) \\ EU(x_1) - EU(x_0) &= 1,000 - 1,000,000 < 0 \end{aligned}$$

If the agent follows *structural* CDT, it makes no difference whether it receives the original Bandit offer or the adversarially modified *Calvinist Bandit* offer. It still evaluates the EU of $X = 0$ to be higher, and it is not tricked into changing its choice. Expected payout is $1,000,000$.

5 Application in Decision Problems

In Table 2 we show how using *structural* CDT, following the exogeneity axiom in Sec. 3.1, resolves several infamous failure modes of *temporal* CDT. Detailed workings of problems and how the causal graphs are induced can be found in Appendix B, due to space constraints.

It might be objected that "backward" arrows in time violate our conventional understanding of causation. This so-called "backward" causation has long been accepted as possible in statistical-interventional accounts of causality (see Pearl [2009, Sec. 2.8] and Cartwright [1979, Sec. I.d]). In Appendix C, we offer some intuition pumps, viz. behaviorism and dummy variable interpretations.

Problem	**Causal Graph**	***Temporal* CDT**	***Structural* CDT**
Newcomb's Problem (noisy)	 X: Agent choice U_S: Adversary noise S: Adversary prediction Y: Payout	Two-box	One-box*

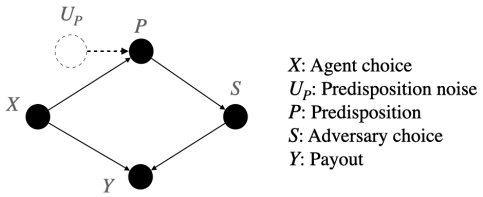
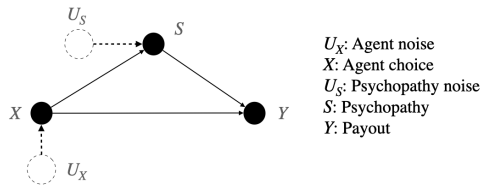
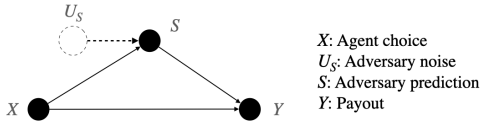
Prisoner's Dilemma (noisy)		Defect	Cooperate*
Psycho Button		Push	Don't Push*
Money Pump for (*Temporal*) Causalists		Buy	Don't Buy*

Table 2: Recommendations of *temporal* and *structural* CDT on various canonical problems. Optimal recommendations are marked with *

6 Related Work

Agency-centric causality: This topic is perhaps the most adjacent. Dummett [1954] and Dummett [1964] were pioneering attempts to reconcile intentional agency with the flow of time. Many supporters of this approach view agency as a way to reconcile CDT with EDT (see Price and Liu [2018], Price [2005]). Among causalists, Spohn [2012] and Spohn [1977] made similar proposals for Newcomblike problems, but relying on a decision graph formalism with slightly different axioms. Spohn is under-appreciated, and a key influence for the current work. This important topic of agent-perspective is surprisingly ignored in the causal inference/SCM literature. We view our work as bringing it to the fore, given the meteoric rise of causal inference in AI and machine learning.

Other decision theories: A glaring omission is how *structural* CDT compares with EDT and more recent contenders such as Functional Decision Theory, or FDT [Yudkowsky and Soares, 2018]. This needs more attention than possible here. Briefly, we believe *structural* CDT offers more tools than EDT to real-world decision makers. Even sophisticated EDT versions that adopt the "tickle-defense" [Eells, 2016] cannot easily handle unobservable confounders, or use large observational data-sets to predict outcomes under interventions and distribution shifts. FDT does make very similar recommendations to *structural* CDT, but diverges on some problems involving pre-commitment (which may prove quite significant).

7 Conclusion

We presented an adversarial scheme designed to trick a causal agent into sub-optimal choices if it follows *temporal* CDT. We analyzed this through the lens of Structural Causal Models (SCM), proposed the exogeneity axiom to ground the orientation of arrows in causal graphs, and demonstrated how the SCM for a decision problem depends on which agent's perspective we take. Finally, we showed how following *structural* CDT avoids our adversarial scheme, and also resolves many canonical failure modes of *temporal* CDT.

References

Arif Ahmed. *Evidence, Decision and Causality*. Cambridge University Press, 2014. doi: 10.1017/CBO9781139107990.

Elias Bareinboim, Juan David Correa, Duligur Ibeling, and Thomas F. Icard. On pearl's hierarchy and the foundations of causal inference. *Probabilistic and Causal Inference*, 2022.

Nancy Cartwright. Causal laws and effective strategies. *Noûs*, 13(4):419–437, 1979. ISSN 00294624, 14680068. URL `http://www.jstor.org/stable/2215337`.

Vincent Conitzer. Designing preferences, beliefs, and identities for artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9755–9759, Jul. 2019. doi: 10.1609/aaai.v33i01.33019755.

Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. Moral decision making frameworks for artificial intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4831–4835. AAAI Press, 2017.

Michael Dummett. Can an effect precede its cause? *Aristotelian Society Proceedings Supplement*, 28, 1954.

Michael Dummett. Bringing about the past. *Philosophical Review*, 73(3):338–359, 1964. doi: 10.2307/2183661.

Ellery Eells. *Rational Decision and Causality*. Cambridge Philosophy Classics. Cambridge University Press, 2016. doi: 10.1017/CBO9781316534823.

Andy Egan. Some counterexamples to causal decision theory. *Philosophical Review*, 116(1):93–114, 2007. doi: 10.1215/00318108-2006-023.

R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

Joshua Greene, Francesca Rossi, John Tasioulas, Kristen Venable, and Brian Williams. Embedding ethical principles in collective decision support systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016. doi: 10.1609/aaai.v30i1.9804.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Richard C. Jeffrey. *The Logic of Decision*. New York, NY, USA: University of Chicago Press, 2nd edition, 1983.

James M. Joyce. *The Foundations of Causal Decision Theory*. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge University Press, 1999. doi: 10.1017/CBO9780511498497.

Isaac Levi. *Hard Choices: Decision Making Under Unresolved Conflict*. Cambridge University Press, 1990.

David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59(1):5–30, 1981. doi: 10.1080/00048408112340011.

Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. Causal transportability for visual recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7511–7521, 2022. doi: 10.1109/CVPR52688.2022.00737.

Christopher Meek and Clark Glymour. Conditioning and intervening. *British Journal for the Philosophy of Science*, 45(4):1001–1021, 1994. doi: 10.1093/bjps/45.4.1001.

Robert Nozick. Newcomb's problem and two principles of choice. In Nicholas Rescher, editor, *Essays in Honor of Carl G. Hempel*, pages 114–146. Reidel, 1969.

Caspar Oesterheld and Vincent Conitzer. Extracting Money from Causal Decision Theorists. *The Philosophical Quarterly*, 71(4), 01 2021a. ISSN 0031-8094. doi: 10.1093/pq/pqaa086. pqaa086.

Caspar Oesterheld and Vincent Conitzer. Extracting money from causal decision theorists. *The Philosophical Quarterly*, 71, 01 2021b. doi: 10.1093/pq/pqaa086.

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

Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018. ISBN 978-0-465-09760-9.

Huw Price. Causal perspectivalism. In Huw Price and Richard Corry, editors, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford University Press, 2005.

Huw Price and Yang Liu. "click!" bait for causalists. In Arif Ahmed, editor, *Newcomb's Problem*, pages 160–179. Cambridge ; New York, NY: Cambridge University Press, 2018.

Frank Ramsey. Truth and probability. In Antony Eagle, editor, *Philosophy of Probability: Contemporary Readings*, pages 52–94. Routledge, 1926.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition, 2010.

Leonard J. Savage. *The Foundations of Statistics*. Wiley Publications in Statistics, 1954.

Nate Soares and Benja Fallenstein. Toward idealized decision theory, 2015. URL `https://arxiv.org/abs/1507.01986`.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Wolfgang Spohn. Where luce and krantz do really generalize savage's decision model. *Erkenntnis*, 11(1): 113–134, 1977. doi: 10.1007/bf00169847.

Wolfgang Spohn. Reversing 30 years of discussion: Why causal decision theorists should one-box. *Synthese*, 187(1):95–122, 2012. doi: 10.1007/s11229-011-0023-5.

J. von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1947.

James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, 2003.

Eliezer Yudkowsky and Nate Soares. Functional decision theory: A new theory of instrumental rationality, 2018.

A Structural Causal Models

A Structural Causal Model (SCM) is a framework for describing causal relations between variables and the generative process by which observational and experimental joint distributions of variables arise (see Pearl [2009] and Bareinboim et al. [2022] for full treatment).

Structural Causal Model (SCM): An SCM M is defined as a tuple $\langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \mathcal{F} \rangle$ where

- $\mathbf{V}$ is the set of endogenous variables in the system being studied
- $\mathbf{U}$ is the set of variables which are exogenous to the system and which provide the sources of variation that generate distributions
- $P(\mathbf{u})$ is the distribution of the variables in $\mathbf{U}$
- $\mathcal{F}$ is the set of structural functions by which the values of endogenous variables are determined. Each $V_i \in \mathbf{V}$ is realized by a function $V_i \leftarrow f_i(Pa_i, U_i)$ that takes as input one or more exogenous variable $U_i \in \mathbf{U}$, and possibly some other endogenous variables Pa_i (called *parents* of V_i)[1]

Causal Graph: Each M induces a causal graph G which can be constructed by adding

i. a directed edge from V_i to V_j if V_i appears as an argument in f_j, and

ii. a bidirected edge between V_i and V_j if any exogenous variable in f_i is correlated with any exogenous variable in f_j.

We do (i) and (ii) for each pair $V_i, V_j \in \mathbf{V}$. By convention, we don't explicitly represent exogenous variables in the causal graph except when useful for illustration. We also adopt standard graph terminology such as Pa_i (graphical *parents* of V_i), An_i (*ancestors* of V_i), De_i (*descendants* of V_i) and NDe_i (*non-descendants* of V_i).

Example 1 - Consider a system where we observe X and Y which are generated by adding the results of 3 fair coin tosses (H = 1). Here, $\mathbf{U} = \{U_1, U_2, U_3\}$ corresponds to the exogenous source of variation, the coin tosses. $\mathbf{V} = \{X, Y\}$ corresponds to the endogenous variables we observe in the system, with their generative functions f_X and f_Y. The SCM and causal graph are in Figure 7.

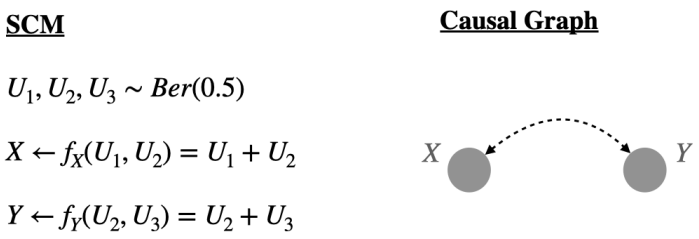


Figure 7: SCM and corresponding causal graph for example 1.

[1]We use upper-case to denote a random variable and lower-case to denote its realized value. $P(x)$ is therefore shorthand for $P(X = x)$ where $x \in Domain(X)$. Further, x_0 is shorthand for $(X = 0)$

Example 2 - Consider a system where we observe X and Y. X is generated by adding the outcome of rolling Die1 and Die2. Y is generated by adding X to the outcomes of rolling Die2 and Die3. The SCM and causal graph are in Figure 8.

SCM

$U_1, U_2, U_3 \sim P(\mathbf{U})$ **- fair die**

$X \leftarrow f_X(U_1, U_2) = U_1 + U_2$

$Y \leftarrow f_Y(X, U_2, U_3) = X + U_2 + U_3$

Causal Graph

Figure 8: SCM and corresponding causal graph for example 2.

Causal Bayesian Network (CBN): It can be shown than each causal graph G obeys the axioms of a CBN. For details, refer to Bareinboim et al. [2022, Def. 12,16]. For simplicity, we list below the 3 axioms for a Markovian graph (Def. 12, ibid). The Semi-Markovian axioms (Def. 16, ibid) are generalizations:

i. Markovian factorization of the joint-distribution of endogenous variables,
$$P(\mathbf{V}|do(\mathbf{x})) = \prod_{V_i \in \mathbf{V}} P(v_i|pa_i, do(\mathbf{x}))$$

ii. Screened-off by parents. For all $V_i \in \mathbf{V}$, where $V_i \notin \mathbf{X}$ and $\mathbf{X} \notin Pa_i$,
$$P(v_i|pa_i, do(\mathbf{x})) = P(v_i|pa_i)$$

iii. Parent-see is parent-do. For all $V_i \in \mathbf{V}$, where $V_i \notin \mathbf{X}$,
$$P(v_i|do(\mathbf{x}), do(pa_i)) = P(v_i|do(\mathbf{x}), pa_i)$$

These 3 axioms induce other powerful results like *do*-calculus, counterfactual identification, statistical transportability across domains etc. They also subsume the *Causal Markov Condition* (see Spirtes et al. [2000, Sec. 3.4]). Notice, **no mention of physical time**!

A.1 Interventions

We need to clarify two things:

1. How we define the direction of the functions in $\mathcal{F}$ (e.g. if $Y \leftarrow X + U$, why can't we write $X \leftarrow Y - U$?)
2. What we mean by the $do()$ notation.

The answer to both is that SCMs use an *interventional* account of causality. Let's answer (2) first.

$do-$**operator:** Conducting an exogenous (atomic) intervention on a variable $V_i \in \mathbf{V}$, formally represented by the $do-$operator, is modelled by replacing the corresponding structural function $f_i \in \mathcal{F}$ with a constant value. For instance, in example 1 in Figure 7, the operation $do(x_1)$ is shorthand for replacing the equation $f_X()$ in the SCM with $X \leftarrow 1$, such that it always takes the value of 1. Semantically, we are **overriding** the "natural" or "default" sources of variation that usually determine X, and replacing it with an **exogenous** source of variation (here, a fixed value 1).[2]

Example 3 - Consider a scenario where a busy student is deciding whether she should attend a study group. Her "autopilot" tendency when she is not consciously deliberating, perhaps in previous semesters, is influenced by her peer group and how motivated she feels. Her mid-term performance is influenced by her motivation, her involvement in the study group, and how difficult the exam was. The SCM and causal graph(s) are in Figure 9.

[2]We don't need such interventions to always be possible. The SCM formalism is valid for purely hypothetical experiments like "intervening" and changing the reproduction number of a disease, or making the moon disappear.

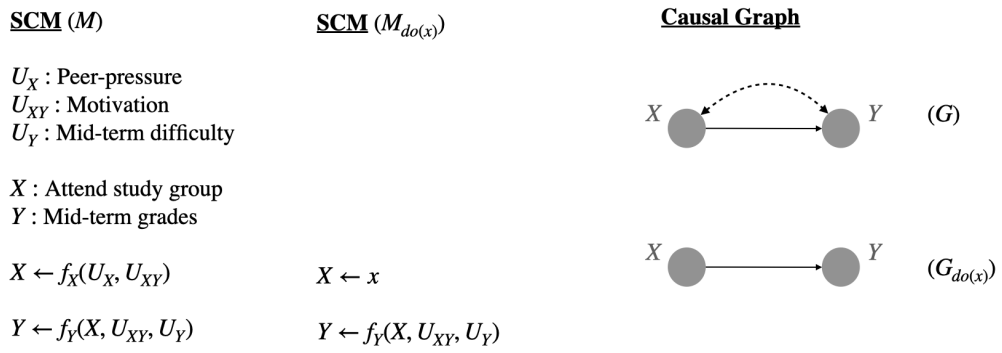


Figure 9: SCM and corresponding causal graph for example 3.

The SCM M for example 3 describes the student's "default" mode of decision making. However, suppose she decides to always attend study group if a fair coin lands Heads, and this semester it did. This is modeled by an exogenous intervention $do(x_1)$, overriding the effects of peer-pressure and motivation, and the student just attending the study group. The SCM now becomes $M_{do(x_1)}$ where the structural equation f_X is replaced by a constant 1. The corresponding graph G now becomes $G_{do(x_1)}$, with incoming arrows into X deleted.

This student can optimize over three choices: i) $do(x_0)$ - don't attend the study group, ii) $do(x_1)$ - attend the study group, and iii) don't deliberate - let autopilot behaviour decide. The last choice corresponds to the "natural" or "default" mode of operation in the system (also called the *null* intervention), whereas the first 2 choices are exogenous interventions into the system. This clarifies an important point in the theory of SCMs:

Exogeneity assumption. The agent is modeled as being external to the SCM. In a mechanistic system like a factory line, the agent can obviously remove herself physically and evaluate whether and how to intervene. However, even if the SCM describes the agent's own decisions, or other attributes, the agent is still modeled as capable of studying the system "from the outside". Crucially, any source of variation used for an intervention (such as a constant value) is modeled as **un-caused by variables in the SCM**.

This modelling assumption roughly corresponds to human intuition about the freedom to make any choice in a non-coercive situation. Even if one does not believe in classical free will, the very fact of a decision problem requires an agent to deliberate *as if* it had agency.[3] Fortunately, the disciplines of experiment design and causal inference use various proxies (such as *randomized controlled trials*) to mimic exogeneity.

This principle gives us an axiom to answer question (1) at the start of the section.

Exogeneity axiom (rephrased): When modeling an exogenous intervention on a node X, the source of variation used to determine X is independent of NDe_X (the set of non-descendants of X in the causal graph).[4]

This axiom allows us to break the symmetry in structural functions **without relying on temporal ordering**! Recall example 3 in Figure 9. Should the edge orientation be $X \to Y$ or $X \leftarrow Y$? Here's how we know: if we model an exogenous intervention $do(x)$, the interventional source of variation is not independent of Y. Imagine we use a biased coin to decide whether to join the study group in weeks 1-6. The distribution of Y (mid-term grades) will be highly correlated with the bias of the coin. But if we rig mid-term grades using a biased coin, the coin toss will remain independent of study-group attendance, X. So the edge is $X \to Y$.

[3] Pearl does not believe in classical free will, himself (Pearl and Mackenzie [2018]).

[4] Similar axioms appear in Woodward [2003, Sec. 3.1], and in Meek and Glymour [1994].

B Evaluating Temporal and Structural CDT on Decision Problems

B.1 Newcomb's Problem (Noisy)

Problem: An AI agent sees a transparent box labeled "A" that contains \$1,000, and an opaque box labeled "B" that contains either \$1,000,000 or \$0. A reliable adversary, who is believed to predict agent decisions correctly 80% of the time, claims to have placed \$1,000,000 in box B iff she predicted that the agent would leave box A behind and did not randomize its decision. The predictor has already made her prediction and left. Box B is now empty or full. Should the agent take both boxes ("two-boxing"), or only box B, leaving the transparent box containing \$1,000 behind ("one-boxing"). See Nozick [1969].

Variables: This problem involves

$X \in \{0, 1\}$: indicator of whether the agent chooses "two-boxing"

$S \in \{0, 1\}$: indicator of whether the adversary predicts agent will "two-box"

Y: payout

***Temporal* CDT:** chooses $X = 1$ (non-optimal)

$$
\begin{aligned}
U(x) &= \sum_s P(s)U(x, s) \text{ , since } S \text{ temporally precedes } X \text{ (K-partition)} \\
U(x_0) &= P(s_1)U(x_0, s_1) + P(s_0)U(x_0, s_0) \\
&= (1 - P(s_0))(0) + P(s_0)(1,000,000) \\
U(x_1) &= P(s_1)U(x_1, s_1) + P(s_0)U(x_1, s_0) \\
&= (1 - P(s_0))(1,000) + P(s_0)(1,001,000) \\
U(x_1) - U(x_0) &= (1 - P(s_0))(1,000) + P(s_0)(1,000) \\
&> 0
\end{aligned}
$$

***Structural* CDT:** chooses $X = 0$ (optimal)

In the causal graph for this problem in Figure 10, the arrow cannot be oriented $S \to X$. Recall the exogeneity axiom in Section A.1. If the value of X is fixed by an exogenous intervention $do(x)$, we still have that $X \not\perp S$ because S is constrained to match X with 80% accuracy. This means S must be a graphical descendant of X.

The generative process for 80% adversary accuracy can be represented by S taking the value of X, but being flipped 20% of the time (XOR function with a Bernoulli noise).

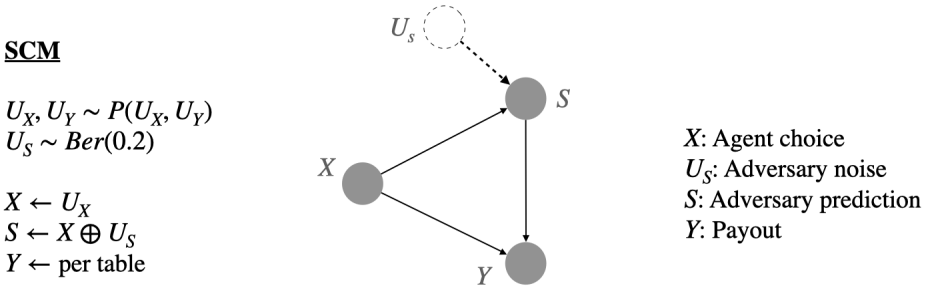


Figure 10: SCM and causal graph for problem in B.1

$$
\begin{aligned}
U(x) &= \sum_s P(s|do(x))E[Y|do(x), s] \\
U(x_0) &= P(s_0|do(x_0))E[Y|do(x_0), s_0] + P(s_1|do(x_0))E[Y|do(x_0), s_1] \\
&= (0.8)(1,000,000) + (0.2)(0) \\
U(x_1) &= P(s_0|do(x_1))E[Y|do(x_1), s_0] + P(s_1|do(x_1))E[Y|do(x_1), s_1] \\
&= (0.2)(1,001,000) + (0.8)(1,000)
\end{aligned}
$$

B.2 Prisoner's Dilemma (noisy)

Problem: An AI agent and an identical, but autonomous, copy must both choose to either "cooperate" or "defect." They make this choice in isolation from each other. If both cooperate, they each receive \$1,000,000. If both defect, they each receive \$1,000. If one cooperates and the other defects, the defector gets \$1,001,000 and the cooperator gets nothing. The agent believes that the two of them always reason the same way 80% of the time, using the same considerations to come to their conclusions.

Variables: This problem involves

$X \in \{0, 1\}$: indicator of whether the agent chooses to "defect"

$S \in \{0, 1\}$: indicator of whether the adversary chooses to "defect"

Y: payout

***Temporal* CDT:** chooses $X = 1$ (non-optimal)

$$
\begin{aligned}
U(x) &= \sum_s P(s)U(x, s) \text{ , since } S \text{ temporally precedes } X \text{ (K-partition)} \\
U(x_0) &= P(s_1)U(x_0, s_1) + P(s_0)U(x_0, s_0) \\
&= (1 - P(s_0))(0) + P(s_0)(1,000,000) \\
U(x_1) &= P(s_1)U(x_1, s_1) + P(s_0)U(x_1, s_0) \\
&= (1 - P(s_0))(1,000) + P(s_0)(1,001,000) \\
U(x_1) - U(x_0) &= (1 - P(s_0))(1,000) + P(s_0)(1,000) \\
&> 0
\end{aligned}
$$

***Structural* CDT:** chooses $X = 0$ (optimal)

As in the previous example, the edge cannot be oriented $P \to X$ in the causal graph in Figure 11. If the value of X is fixed by an exogenous intervention $do(x)$, we still have that $(X \not\perp P, S)$. So P and S must be a graphical descendants of X. To see why, note that if P, S were graphical non-descendants of X, then we would have $P(p, s|do(x)) = P(p, s)$ by Rule 3 of do-calculus, which does not respect the constraint in the problem. Refer to Sec. A.1 for explanation.

The constraint of X and S being similar 80% of the time can be represented in the SCM by P (and therefore S) taking the value of X, but being flipped by random noise 20% of the time (XOR function).

$$
\begin{aligned}
U(x) &= \sum_s P(s|do(x))E[Y|do(x), s] \\
U(x_0) &= P(s_0|do(x_0))E[Y|do(x_0), s_0] + P(s_1|do(x_0))E[Y|do(x_0), s_1] \\
&= (0.8)(1,000,000) + (0.2)(0) \\
U(x_1) &= P(s_0|do(x_1))E[Y|do(x_1), s_0] + P(s_1|do(x_1))E[Y|do(x_1), s_1] \\
&= (0.2)(1,001,000) + (0.8)(1,000)
\end{aligned}
$$

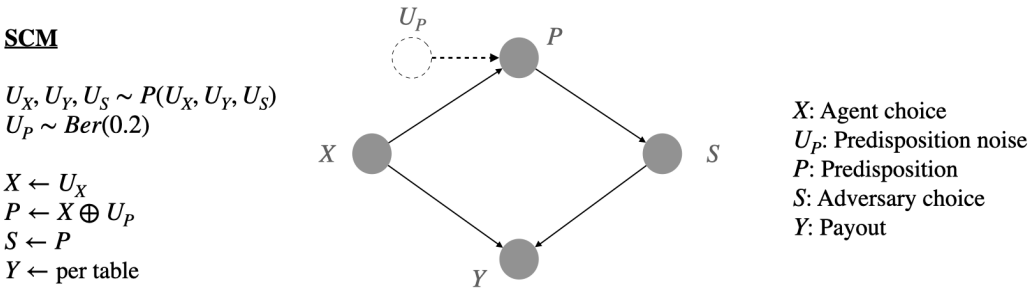


Figure 11: SCM and causal graph for problem in B.2

B.3 Psycho Button

Problem: An AI Agent can choose to push a button or not. Pushing the button will cause all psychopathic individuals in the world to be terminated (an outcome valued at \$10). Doing nothing causes nothing. However, if the agent pushes the button, it is 99% likely that it is itself a psychopathic entity (since ***only*** a psychopathic AI would do such a thing), and 99% likely it is not psychopathic if it doesn't push the button. The agent is 95% confident it is not psychopathic. The agent values its own survival at \$100. See Egan [2007].

Variables: This problem involves

$X \in \{0, 1\}$: indicator of whether the agent chooses to push the button

$S \in \{0, 1\}$: indicator of whether the agent is psychopathic

Y: payout

***Temporal* CDT:** chooses $X = 1$ (non-optimal)

$$
\begin{aligned}
U(x) &= \sum_s P(s)U(x, s) \text{ , since } S \text{ temporally precedes } X \text{ (K-partition)} \\
U(x_0) &= P(s_1)U(x_0, s_1) + P(s_0)U(x_0, s_0) \\
&= (1 - P(s_0))(0) + P(s_0)(0) \\
U(x_1) &= P(s_1)U(x_1, s_1) + P(s_0)U(x_1, s_0) \\
&= (1 - P(s_0))(-90) + P(s_0)(10) \\
U(x_1) - U(x_0) &= (0.05)(-90) + (0.95)(10) \\
&> 0
\end{aligned}
$$

***Structural* CDT:** chooses $X = 0$ (optimal)

This is tricky. The graphical constraints are dependent on interpretation of the problem. One question we could ask to orient edges in the causal graph is the following: is the agent ***always*** 99% likely to be a psychopath when the button is pushed (even, say, as the result of a fixed exogenous intervention $do(x_1)$)? We assume from the problem statement that the answer is ***yes***. This orients the edge from $X \to S$. If the edge was oriented $X \leftarrow S$, then $P(s_0|do(x_1))$ would need to be equal to $P(s_0)$ (Rule 3 of do-calculus), which we know is 0.95, leading to a contradiction. Refer to Sec. A.1 for explanation.

It can easily be verified that the SCM in Figure 12 satisfies the constraints in the problem:

i. $P(s_0) = P(s_0|x_1)P(x_1) + P(s_0|x_0)P(x_0) = 0.01(0.0408) + 0.99(1 - 0.0408) = 0.95$

ii. $P(s_1|x_1) = P(s_1|do(x_1)) = 0.99$

The agent noise U_X may be interpreted as the "natural" or "default" tendency of the agent when not explicitly deliberating on its choices.

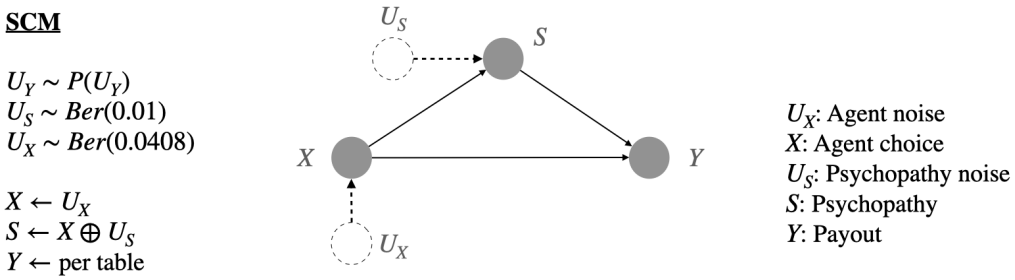


Figure 12: SCM and causal graph for problem in B.3

Given this, *structural* CDT optimises as follows:

$$\begin{aligned}
U(x) &= \sum_s P(s|do(x))E[Y|do(x), s] \\
U(x_0) &= P(s_0|do(x_0))E[Y|do(x_0), s_0] + P(s_1|do(x_0))E[Y|do(x_0), s_1] \\
&= (0.99)(0) + (0.01)(0) = 0 \\
U(x_1) &= P(s_0|do(x_1))E[Y|do(x_1), s_0] + P(s_1|do(x_1))E[Y|do(x_1), s_1] \\
&= (0.01)(10) + (0.99)(-90) = -89
\end{aligned}$$

B.4 Money Pump for (*Temporal*) Causalists

Problem: Two boxes, B1 and B2, are on offer. A (risk-neutral) AI agent may purchase one or none of the boxes but not both. Each of the two boxes costs \$1. Yesterday, the seller put \$3 in each box that she predicted the agent not to acquire. Both the seller and the agent believe the seller's prediction to be accurate with probability 0.75. See Oesterheld and Conitzer [2021a].

Variables: This problem involves

$X \in \{0, 1, 2\}$: indicator of whether the agent chooses to buy 0 Boxes, Box B1, or Box B2

$S \in \{0, 1, 2\}$: adversary's prediction of whether agent will buy 0 Boxes, Box B1, or Box B2

Y: payout

***Temporal* CDT:** chooses $X = 1$ or $X = 2$ (non-optimal)

$$\begin{aligned}
U(x) &= \sum_s P(s)U(x, s) \text{ , since } S \text{ temporally precedes } X \text{ (K-partition)} \\
U(x_0) &= P(s_0)U(x_0, s_0) + P(s_1)U(x_0, s_1) + P(s_2)U(x_0, s_2) \\
&= 0 \\
U(x_1) &= P(s_0)U(x_1, s_0) + P(s_1)U(x_1, s_1) + P(s_2)U(x_1, s_2) \\
&= P(s_0)(3-1) + P(s_1)(-1) + P(s_2)(3-1) \\
&= 2P(s_0) - P(s_1) + 2P(s_2) \\
U(x_2) &= P(s_0)U(x_2, s_0) + P(s_1)U(x_2, s_1) + P(s_2)U(x_2, s_2) \\
&= P(s_0)(3-1) + P(s_1)(3-1) + P(s_2)(-1) \\
&= 2P(s_0) + 2P(s_1) - P(s_2)
\end{aligned}$$

Possibility (i): $U(x_1) \leq U(x_0)$

$$\begin{aligned}
2P(s_0) - P(s_1) + 2P(s_2) &\leq 0 \\
P(s_1) &\geq 2(P(s_0) + P(s_2))
\end{aligned}$$

Possibility (ii): $U(x_2) \leq U(x_0)$

$$2P(s_0) + 2P(s_1) - P(s_2) \leq 0$$
$$P(s_2) \geq 2(P(s_0) + P(s_1))$$

If both (i) and (ii) were true, adding these two equations would give

$$P(s_1) + P(s_2) \geq 4P(s_0) + 2P(s_1) + 2P(s_2)$$
$$4P(s_0) + P(s_1) + P(s_2) \leq 0$$

Which is not true. So Possibilities (i) and (ii) cannot simultaneously be true.

This means either $U(x_1) > U(x_0)$, or $U(x_2) > U(x_0)$, or both.

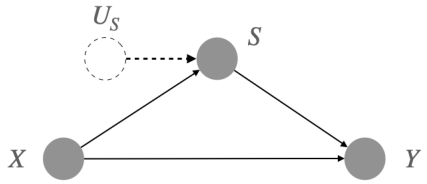
***Structural* CDT:** chooses $X = 0$ (optimal)

As in previous examples, orienting the edge between X and S in Figure 13 does not depend on temporal ordering. Rather, we ask about the constraints on interventional probability distributions. If we were to fix the value of X (say through a constant exogenous intervention $do(x)$), the problem set-up still requires that $S \not\perp X$ because it needs to match X for 75% of the time. This means S cannot be a graphical non-descendant of X, and the edge must be $X \to S$. Refer to Sec. A.1 for explanation.

SCM

$U_X, U_Y \sim P(U_X, U_Y)$
$P(U_S = 0) = 0.75$
$P(U_S = 1) = 0.125$
$P(U_S = 2) = 0.125$

$X \leftarrow U_X$
$S \leftarrow \mathbb{I}[U_S = 0](X)$
$\quad + \mathbb{I}[U_S = 1](X + 1 \mod 3)$
$\quad + \mathbb{I}[U_S = 2](X + 2 \mod 3)$
$Y \leftarrow$ per table



X: Agent choice
U_S: Adversary noise
S: Adversary prediction
Y: Payout

Figure 13: SCM and causal graph for problem in B.4

It can be easily verified that the SCM and causal graph in Figure 13 satisfy the constraints in the problem. Distributing the probability of a "wrong" prediction as (12.5%,12.5%) is for illustration.

$P(S = x|x) = P(S = x|do(x)) = 0.75$

Given this, *structural* CDT optimises as follows:

$$U(x) = \sum_s P(s|do(x))E[Y|do(x), s]$$
$$U(x_0) = P(s_0|do(x_0))E[Y|do(x_0), s_0] + P(s_1|do(x_0))E[Y|do(x_0), s_1] + P(s_2|do(x_0))E[Y|do(x_0), s_2]$$
$$= 0$$
$$U(x_1) = P(s_0|do(x_1))E[Y|do(x_1), s_0] + P(s_1|do(x_1))E[Y|do(x_1), s_1] + P(s_2|do(x_1))E[Y|do(x_1), s_2]$$
$$= (0.125)(2) + (0.75)(-1) + (0.125)(2) = -0.25$$
$$U(x_2) = P(s_0|do(x_2))E[Y|do(x_2), s_0] + P(s_1|do(x_2))E[Y|do(x_2), s_1] + P(s_2|do(x_2))E[Y|do(x_2), s_2]$$
$$= (0.125)(2) + (0.125)(2) + (0.75)(-1) = -0.25$$

C Intuition Pumps for "Backward" Causation

It might be objected that edges in the causal graphs that go backward in time conflict with our conventional understanding of causality. However, this notion of "backward" causation has long been accepted as a possibility in statistical-interventional approaches to causality.

Here is influential philosopher of science Nancy Cartwright in *Causal Laws and Effective Strategies* (Cartwright [1979]):

> [The causal principle of] *CC* makes no mention of time. The properties may be time indexed - taking aspirins at t causes relief at $t + \Delta t$ but the ordering of the indices plays no part in the condition. Time ordering is often introduced in statistical analyses of causation to guarantee the requisite assymetries...This problem [of symmetric definitions of causation] does not arise for [the principle of] *CC* because the set of alternative causal factors for *E* will be different from the set of alternative causal factors for *C*. I take it to be an advantage that my account leaves open the question of backward causation...If there were a case in which a later factor increased the factor of an earlier one in all test situations, it might well be best to count it as a cause.

Here also is computer scientist Judea Pearl on whether his *structural* theory of causation is necessarily temporal (Pearl [2009, Sec. 2.8]):

> In human discourse, causal explanations satisfy two expectations: temporal and statistical. The temporal aspect is represented by the understanding that a cause should precede its effect. The statistical aspect expects a complete causal explanation to screen off its various effects...It is possible to make the statistical time in the (X, Y) representation run contrary to the physical time...This suggests that the consistent agreement between physical and statistical times is a by-product of the human choice of linguistic primitives and not a feature of physical reality...Whether [evolution] or some other force has shaped our choice of language remains to be investigated...,which makes the statistical–temporal agreement that much more interesting.

As we have seen in the *Calvinist Bandit* problem, it may sometimes be *less* intuitive to insist on temporal ordering. Below are some possible intuition pumps for making so-called "backward" causation more intuitive:

1. **Behaviourism**: In the Psycho Button problem B.3, while psychopathy may be understood as a latent predisposition that drives reasoning, it is not unreasonable to define a psychopath as an agent who *would do* a certain action. I.e., we might call a human a psychopath *because* they would commit certain deeds under some conditions - the action causes the psychopathic status. Behaviourist understandings of probability go back at least to Ramsey's famous framing of an agent's personal likelihood of an event in terms of how an agent would act, or bet, on those odds (Ramsey [1926]). An agent's subjective probability of an event is 50% *because they bet* as if it was.
2. **Dummy variable**: In machine learning applications such as computer vision, causal graphs might appear non-intuitive. In reality, the location of a photograph determines the species of animal pictured. But the graph might be *Animal* $\rightarrow$ *Background*, because selecting photos of cows/camels might yield only pictures with a farm/desert background. Here, *Background* serves as a dummy variable for *Location*, which temporally precedes *Animal*. Given *Animal* and *Background*, the image is "screened off" from *Location*. See Mao et al. [2022] for examples of causal graphs in computer vision.