

Causal Persuasion*

Anastasia Burkovskaya¹ and Egor Starkov^{†2}

¹University of Sydney, School of Economics

²University of Copenhagen, Department of Economics

May 1, 2026

Abstract

We propose a model of causal persuasion, in which a sender selectively discloses a set of variables together with their true joint distribution and proposes a subjective causal model that binds them. A receiver is persuaded by this model only if the data conclusively identifies the causal link of interest. We characterize when such persuasion succeeds or fails, and how easily it can be achieved. We further show that if the receiver holds a pre-existing subjective model, debunking it is similar to persuading a receiver without one. To establish a true causal link, the sender often needs to disclose only one or two well-chosen variables. But to dispel a perceived link—to persuade the receiver there is no causal relationship—every common cause must be disclosed. Our results highlight a fundamental asymmetry in causal persuasion: Establishing causality is often much easier than ruling it out.

JEL-codes: D83

Keywords: causal persuasion, causal models, directed acyclic graphs, strategic communication

1 Introduction

Media offers *stories*. Politicians tell *stories*. Advertising pushes *stories*. Economists debate *stories*. “Why Investors Are Worried About Japan’s Bond Market.”¹ “Because of Tariffs, our Economy is BOOMING!”² “You’re Not You When You’re Hungry”³ “Do Financial Concerns Make Workers Less Productive?”⁴ All of these examples suggest causal relationships that rationalize the observed data. Some stories try their best to offer a good-faith explanation of the data. Some of those fail despite their best efforts.⁵ Others come up

*We are grateful to Murali Agastya, Chiara Aina, Kadir Atalay, Jean-Michel Benkert, Andreas Bjerre-Nielsen, Martin Dufwenberg, Andrew Ellis, Christian Gillitzer, Benjamin Golub, Alessandro Ispano, Anton Kolotilin, Klaus Kultti, Jay Lu, Joel Sobel, as well as seminar and conference audiences at Copenhagen, Sydney, NET, EWMES 2025, AETW 2026 for the many valuable comments.

[†]Burkovskaya: anastasia.burkovskaya@sydney.edu.au; Starkov: egor.starkov@econ.ku.dk.

¹Bloomberg, Jul 23, 2025. URL: <https://www.bloomberg.com/news/articles/2025-07-23/why-japan-s-bond-market-has-investors-worried>, retrieved Aug 08, 2025.

²Donald Trump, Truth Social, Jun 03, 2025. URL: <https://truthsocial.com/@realDonaldTrump/posts/114617666432673584>, retrieved Aug 08, 2025.

³<https://www.youtube.com/watch?v=0TPJYZLD6L8>, retrieved Aug 08, 2025.

⁴Kaur et al. (2025)

⁵“Leaders: Stop Confusing Correlation with Causation”. Harvard Business Review, Nov 05, 2021. URL: <https://hbr.org/2021/11/leaders-stop-confusing-correlation-with-causation>, retrieved Aug 08, 2025.

with a spin that is favorable for them. What makes a causal story persuasive? How can a persuader instill a causal story with a listener? What does it take to debunk a story? These are the questions that we tackle in this paper.

We are the first to develop a model of *causal* persuasion in which a sender tries to persuade a receiver about a particular causal relationship. The receiver may or may not initially have knowledge of some real-world variables and a subjective causal model that links them. The sender can disclose some variables to either persuade the receiver of some causal model, or falsify the receiver’s model and replace it with a new one. We provide conditions under which persuasion is possible and provide blueprints for effective persuasion mechanisms. We show that some properties of the data can make it “easy” to persuade the receiver of a certain causal connection and/or to debunk the receiver’s faulty model. Without these properties, the respective tasks can become “difficult” for the sender even if possible in principle. Finally, we show that the sender may be able to instill an incorrect model even if they are not able to lie about the factual data.

Specifically, we assume that there exists rich data generated by some true causal model, captured by a directed acyclic graph (DAG). The sender observes the true graph and all of the data. We consider two different cases: when the receiver has no subjective model and only needs to be persuaded, and when the receiver starts off with some subjective model already, which the sender needs to debunk before attempting persuasion. To talk about persuasion, we first assume the receiver is initially unaware of any variables. The sender can selectively disclose some of the true variables, which makes the receiver aware of them and informed of the data related to them. The sender further offers some causal model. A naïve receiver accepts this model if it is consistent with the data. A sophisticated receiver accepts the model only if the data proves it conclusively.

We show that a naïve receiver can always be persuaded that x affects y without disclosing any extra variables, as long as x and y are correlated. In contrast, persuading a sophisticated receiver is a non-trivial task. On the one hand, when the sender’s narrative aligns with the truth, revealing only one or two carefully chosen variables in addition to x and y may be effective. Disclosing these variables allows the receiver to rule out the other possibilities and conclude that x indeed causes y , as the sender suggests. However, if no such appropriate variables exist, then while persuasion may still be feasible, the sender’s model may need to be impractically large to be persuasive. On the other hand, persuading a sophisticated receiver of a narrative that is reverse to the true causality between x and y is altogether impossible (Theorem 1). Despite this, the sender *can* sometimes mislead the receiver about correlation arising from confounding variables and present it as causation, since it does not directly contradict the true causal graph.

We then consider the case when the receiver has some pre-existing subjective model. The sender needs to debunk this model before they can persuade the receiver of anything else. We show that to be falsifiable, the receiver’s model must be causally incorrect. In particular, it is not sufficient for the receiver’s model to omit some variables, but rather this omission must lead to the receiver “flipping” some causal connections relative to the true model or to creating new causal links that are not present in the true model.

In case the receiver’s model *can* be debunked, doing so is qualitatively similar to persuading a sophisticated receiver with a blank mind. In particular, it can be done by revealing at most two appropriate variables if they exist (Theorem 2) and can be arbitrarily difficult (in terms of disclosure volume) otherwise. For example, debunking some causal link (e.g., that immigration causes crime or that police funding causes crime) and persuading the receiver that no such link exists requires identifying and revealing *all* common causes of the two variables. That means accounting for every source of correlation

between them. This is only possible if the two variables are in fact not directly connected in the true causal graph. Because there may be arbitrarily many confounders, such persuasion can be very burdensome for both sender and receiver. In practice, disproving a causal link can often be harder than persuading the receiver that the link runs in the opposite direction—despite no such link existing in the first place, meaning that deception can be easier than establishing the truth. This stands in contrast to the classical Humean asymmetry for universal claims (e.g., “all swans are white”), which are difficult to verify but easy to falsify (Hume, 1748; Al-Najjar et al., 2014). The difference arises because, unlike universal claims, ruling out a causal relationship requires eliminating all competing explanations.

In our approach, we rely on directed acyclic graphs (DAGs) to represent causal models. We use the toolbox on causal discovery developed in computer science literature on Bayesian networks, see Pearl (2009) for a textbook treatment and Guo et al. (2021) or Zanga et al. (2022) for recent surveys of that literature. Causal DAGs in economic literature are primarily used to model agents with causal misperceptions; see Ambuehl et al. (2026) for an excellent overview of the relevant economic literature. To our best knowledge, ours is the first paper to explicitly incorporate causal DAGs into the strategic communication framework—demonstrating, among other things, how aforementioned causal misperceptions could arise in the first place in a strategic environment.

Our model is closest to the emerging literature on “narrative persuasion”. Schwartzstein and Sunderam (2021), Aina (2024), and Ispano (2025) explore models, in which the receiver observes the marginal distributions of different variables (such as states and signals), and the sender offers models capturing joint distributions of different variables. In Schwartzstein and Sunderam (2021) and Aina (2024), the receiver selects the model that is more plausible given the ex post variable realizations. In our model, in contrast, the receiver only adopts the sender’s proposed model if the receiver’s own model is rendered inconsistent with the data by the sender’s disclosure. Closest to ours is the work by Ispano (2025), in which the receiver adopts a model only if it is compatible with the data. However, the persuasion mechanisms in Ispano (2025) only concern the interpretation of relations between a set of known variables, whereas our model explores optimal disclosure of new variables.

Similarly, Spiegler (2020) and Eliaz et al. (2021) explore the limits of misperception in DAGs. While they do not frame their results in the context of strategic communication, they explore how large of a perceived correlation could a sender conceivably induce between the two variables by fitting a strategically chosen DAG to some objective data. In contrast, our question is qualitative rather than quantitative: “how can a sender persuade the receiver that a causal link exists (or not)?” This is as opposed to the question of how strong can this link be presented to be.

Contemporary work by Eliaz and Rubinstein (2025) explores a related model of “Wasonian persuasion”, where a sender also chooses which causal DAG to offer to the receivers. They assume that the receivers then search for the data and evaluate the proposed model based on the first-encountered relevant data point, while we assume that the receivers have access to the full population data but may not include all relevant variables in the decision frame. They characterize the distribution of DAGs that prevails in the society as a result of the sender’s strategic concerns. Eliaz and Spiegler (2020) and Eliaz et al. (2025) provide alternative characterizations of models that would be predominant in the society in an equilibrium, but do so in the absence of a strategic sender and with different selection criteria.

More broadly, our work relates to literature on strategic communication that has long

explored a question of “what makes communication persuasive?” See Little (2023) for a brief overview of the literature. Specifically, our model is connected to the literature on disclosure of verifiable information, in which the sender can withhold data but cannot produce fake data (Grossman and Hart, 1980; Milgrom, 1981; see Dranove and Jin, 2010 for an overview and Di Tillio et al., 2021 for a recent alternative model). Literature on Bayesian Persuasion considers a problem that is similar to disclosure models, but allows the sender to design information in a rich way (Kamenica and Gentzkow, 2011; see Bergemann and Morris, 2019 and Kamenica, 2019 for recent overviews). Our approach is different from both of these strands of literature in that instead of disclosing variable realizations, in our model the sender can disclose variables themselves, which has implications for which causal models remain plausible. In particular, the sender in our model can offer an incorrect model to go with factually correct data.

We next present an illustrative example in Section 2. The formal model is set up in Section 3. The main results are contained in Sections 4 and 5. Section 6 discusses some of the assumptions behind the model and outlines directions for future research.

2 Illustrative example

Consider the following illustrative example. It is established that an MBA degree is correlated with higher lifetime earnings.⁶ However, it may not be completely clear which way the causal link goes. Business schools naturally want everyone to believe that their degrees offer substantial value to prospective students. However, an employee weighing whether to leave their job for an MBA also allows for the possibility that the degree has no real value, and that the causal link runs in the opposite direction—high earners would succeed regardless and merely self-select into earning an MBA. An employer is interested in sending the employee to obtain an MBA if and only if it is valuable.

Suppose, for the sake of argument, that the true causal graph is as presented in Figure 1(a): education e (MBA degree) neither increases earnings w , nor is caused by them. Instead, the two are correlated due to two confounding variables, a person’s ability a and social skills s , that both affect earnings w .⁷ In addition, earnings w are also affected by job experience/tenure t . Can the business school persuade the employee that education affects earnings? If so, can the employer then debunk this view and persuade the employee that no actual link exists between e and w ?

Suppose the business school launches a promotional campaign around variables e , w , and t under the slogan “Experience + MBA = More Pay!” This campaign makes the employee aware of variable t and how this variable is connected to e and w in the real world. Specifically, they see that e is independent of t unconditionally, but the two are correlated conditional on w . The employee infers that this is only possible if e and t jointly determine w , as in Figure 1(b). Indeed, the conditional correlation suggests that there must be *some* connection between them, but if $e \rightarrow w \rightarrow t$ or $e \leftarrow w \leftarrow t$ or $e \leftarrow w \rightarrow t$ (or $e \rightarrow t$ or $e \leftarrow t$), then e and t would have also been unconditionally correlated. After hearing this campaign, the employee then has no choice but to conclude that it is not selection $e \leftarrow w$ that creates the correlation between e and w , but there is an actual causal

⁶“The MBA Premium: What MBAs Earn Over A Lifetime Will Shock You” Poets and Quants, May 05, 2021. URL: <https://poetsandquants.com/2021/05/05/lifetime-earnings-of-mbas/>, retrieved Nov 21, 2025.

⁷For example, Tamborini et al. (2015) show using US panel data that while the effect of education is far from zero, “accounting for key covariates reduces the estimated lifetime earnings return to college education by approximately 30%” (pp.1396–1397) and the return to graduate degrees by approximately 20% (Fig. 2).

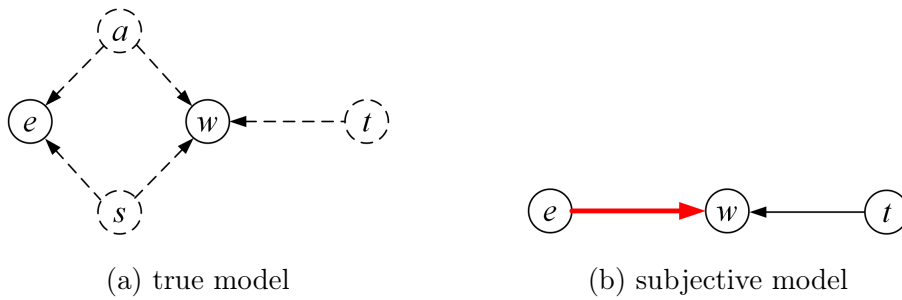


Figure 1: The true DAG and the employee’s model after the promo campaign.

NOTES: Throughout the paper, the figures use dashed variable outlines and links to represent the items that are not (made) known to the receiver. Bold red arrows represent defective links.

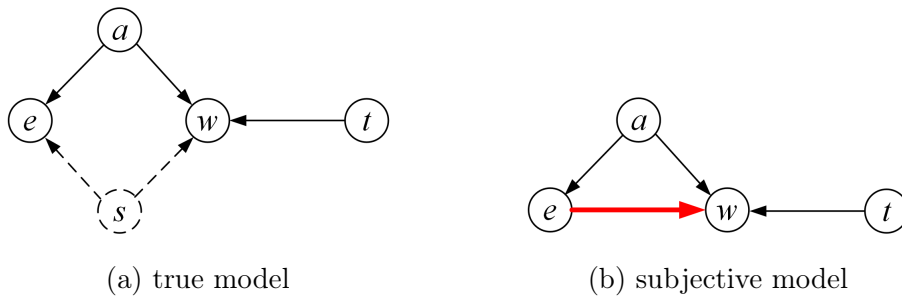


Figure 2: Employee’s model after the first chat with the employer.

link $e \rightarrow w$. The promotional campaign successfully persuades the employee by disclosing an additional variable on top of two variables of interest. Notably, persuasion is effective in spite of the suggested narrative being untruthful.

Suppose the employer, in turn, knows that the degree does not actually increase the employee’s value (and, therefore, earnings w). When an employee asks the employer to fund their education, the employer seeks to make this lack of causal connection clear—to debunk the model that the business school instilled. Suppose the employer first tells the employee that ability a is a common factor that determines both educational choices e and earnings w . The employee can see that a is indeed correlated with both e and w , and sees no contradiction in the data to the suggestion that a is the cause while e and w are the effect. But at the same time, neither can the employee see any contradiction to education e increasing earnings w , since it is clear from the data that e and w are correlated even conditional on a . The employee’s subjective model after this conversation is presented in Figure 2(b). Disclosing one common cause alone is insufficient for the employer to convince the employee that the presumed causal link does not exist.

Indeed, to convince the employee that e does not directly affect w , the employer would have to disclose all common factors influencing both e and w , which in this example include a and s , but in general there could be many more. So long as there is even a single confounding variable not known to the receiver, they see that e and w are correlated conditional on everything they know, suggesting that e and w must be causally connected. While it was easy for the business school to persuade the employee that the link runs in the opposite direction—requiring disclosure of only one variable—it is far harder for the employer to convince the employee that no direct link exists, as this demands revealing *all* relevant variables.

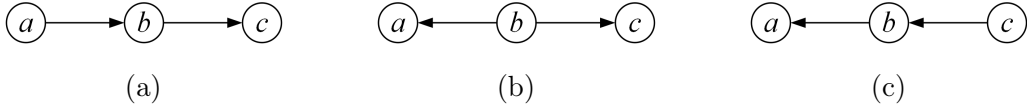


Figure 3: Multiple models consistent with the same data.

3 Model

3.1 Setup

The world. The world is described by a causal directed acyclic graph (DAG) (Ω_t, C_t) called *the true model*, where Ω_t is a set of *variables*, and $C_t : \Omega_t \rightarrow 2^{\Omega_t}$ describes directed links capturing the causal relations between them. For any variable $a \in \Omega_t$, relation $b \in C_t(a)$ is denoted equivalently as $b \rightarrow_t a$ and described as “ b has a causal effect on a ” and “ b is a parent of a ”.⁸ We further let $\bar{C}_t(a)$ denote the set of predecessors of any $a \in \Omega_t$: $b \in \bar{C}_t(a)$, denoted equivalently as $b \Rightarrow_t a$, is true if and only if there exists a path $b \rightarrow_t \dots \rightarrow_t a$ in C_t .

Each node $a \in \Omega_t$ represents a random variable distributed according to some c.d.f. $a \sim F_a(\cdot | C_t(a))$. We let P denote the joint distribution of all variables, so for any vector of realizations of Ω_t : $P(\Omega_t) \equiv \prod_{a \in \Omega_t} F_a(a | C_t(a))$. For any subset of variables $\Omega \subset \Omega_t$, let $P|\Omega$ denote the the marginal distribution of Ω . In what follows, we refer to the full distribution P and all marginals $P|\Omega$ as the *data*. We also assume that if a player is aware of set Ω of variables, then they observe data $P|\Omega$, meaning that they observe the whole true joint distribution of these variables. Further, given some $S \subset \Omega_t$ and $a, b \in \Omega_t \setminus S$, we use the notation $(a \perp b | S)$ if a and b are statistically independent conditional on S given P , and $(a \not\perp b | S)$ if the converse is true.

Subjective models and consistency. We define a (*subjective*) *model* of the world as a DAG (Ω, C) with $\Omega \subseteq \Omega_t$. Given a model (Ω, C) , we say that variables $a, b \in \Omega$ are *adjacent* if $a \rightarrow b$ or $a \leftarrow b$, are *connected* if $a \Rightarrow b$ or $a \Leftarrow b$. We say that a, b are *correlated* if either they are connected, or there exists a common ancestor $c \in \bar{C}(a) \cap \bar{C}(b)$. We call a triplet of variables $a - b - c$ (where b is adjacent to a and c , but a and c are not adjacent) a *collider* if $a \rightarrow b \leftarrow c$, a *chain* if $a \rightarrow b \rightarrow c$ or $a \leftarrow b \leftarrow c$, and a *fork* if $a \leftarrow b \rightarrow c$.

We say that model (Ω, C) is *consistent with data* $P|\Omega$ if the two following conditions hold:

(Markov property) There exists a collection of distribution functions $\left\{ \hat{F}_a(\cdot | C(a)) \right\}_{a \in \Omega}$ such that $P(\Omega) = \prod_{a \in \Omega} \hat{F}_a(a | C(a))$.

(Minimality) For any $a, b \in \Omega$ it holds that if $b \in C(a)$, then there is no such $S \subset \Omega$ that $(a \perp b | S)$.

The two conditions effectively require that the model and the data agree on which variables are pairwise independent. The Markov property requires that if a does not affect b according to the model, then the distribution of b in the data should not depend on the realization of a . Minimality requires the converse: that if a pair of variables are statistically independent in the data, then they must not be directly linked in the model. Minimality is also known as faithfulness or d-faithfulness (Nogueira et al., 2022).

⁸With abuse of notation, we use C_t to denote both the parental relation and the set of directed links it generates.

We assume that the true model is consistent. While the Markov property holds by definition, minimality is non-trivial and requires that any link in the true model creates a *statistical* dependency between the two variables. More generally, the following definition will prove helpful, which bridges the topology of the true model and conditional independence relations in the data.

Definition 1. *Set of variables $S \subset \Omega_t$ d-separates variables $a, b \in \Omega_t \setminus S$ if for any path between a and b in C_t , one of the following holds:*

1. *the path contains no colliders and S contains at least one variable from this path, or*
2. *the path contains at least one collider $x \rightarrow_t y \leftarrow_t z$ such that S does not contain y or any of its descendants $\bar{C}_t^{-1}(y)$.*

The Markov property implies that if S d-separates a and b then $(a \perp b \mid S)$ (Pearl, 1988, Chapter 3). Minimality implies the converse. In the end, $(a \perp b \mid S)$ in the data for some a, b, S if and only if S d-separates a and b in (Ω_t, C_t) . In the special case of $S = \emptyset$, it follows that $a \not\perp b$ if and only if there exists a path between a and b with no colliders—i.e., if and only if a and b are correlated in (Ω_t, C_t) in the sense of the definition above.

Note that there may be multiple consistent models (Ω, C) for some given data $P \mid \Omega$ on some set of variables $\Omega \subseteq \Omega_t$. For example, consider different models in Figure 3: they are all consistent with data P where variables $a, b, c \in \Omega$ are all pairwise correlated but $(a \perp c \mid b)$. Conversely, for a given variable set $\Omega \subset \Omega_t$, there may also not exist any model (Ω, C) that is consistent with the data generated by the true model (Ω_t, C_t) , see Example 2 further below. We say that link $a \rightarrow b$ (equivalently, $a \in C(b)$) in a consistent subjective model (Ω, C) is *uniquely consistent* with the data $P \mid \Omega$ if there does not exist any consistent model (Ω, C') with $b \in C'(a)$.

Players. We consider an interaction between two players: a sender and a receiver.

The receiver initially has no model in mind and is not aware of any variables. The sender knows the true model (Ω_t, C_t) . The sender can *propose* a model (Ω_s, C_s) to the receiver, which the receiver can then *accept or reject*. The sender can propose any model such that $\Omega_s \subseteq \Omega_t$ and (Ω_s, C_s) is consistent with $P \mid \Omega_s$. In other words, the sender can reveal an arbitrary subset of variables and a subjective model that describes the causal connections between them.

We assume that there are two variables of interest for the sender, $x, y \in \Omega_t$, and the sender wants to persuade the receiver that $x \rightarrow_s y$.⁹ We consider two scenarios under which the receiver accepts the proposed model:

- a *naïve* receiver accepts any model (Ω_s, C_s) that is consistent with $P \mid \Omega_s$, and
- a *sophisticated* receiver accepts model (Ω_s, C_s) if and only if the link $x \rightarrow_s y$ is uniquely consistent with $P \mid \Omega_s$. More generally, a sophisticated receiver accepts a given sender’s model if and only if every link along every path connecting x and y is uniquely consistent.

A naïve receiver is someone who is uncritical: whenever they are presented with a correlation and a story that does not contradict it, they accept such a story. A sophisticated receiver, on the other hand, is wary of reverse causality; they accept the sender’s causal narrative regarding x and y only if the data explicitly rules out all other possibilities. We assume that the receiver’s type is commonly known. Regardless of whether the receiver accepted the model, the game ends.

⁹Throughout the paper, we use subscripts to indicate which model the arrow notation refers to: e.g., “ $a \rightarrow_t b$ ” means $a \in C_t(b)$, and “ $a \Rightarrow_s b$ ” means $a \in \bar{C}_s(b)$.

In addition to the goal of persuading the receiver that $x \rightarrow_s y$, we assume that the sender’s secondary objective is to do this using the smallest number of additional variables, i.e., to minimize $|\Omega_s|$.¹⁰ We interpret this objective as aversion to model complexity. The sender may face difficulties when trying to effectively communicate complex models with a large number of variables. The receiver may find such models difficult to understand and ultimately less convincing than alternative simple models.¹¹

We note that the setup above does not describe a “game” in the strict game-theoretic sense, since the receiver is not a strategic payoff-maximizing player, but rather follows a specific choice procedure. This assumption is discussed at length in Section 6. Our problem is therefore better seen as the sender’s decision problem. This concludes the setup of the core model; the remainder of Section 3 introduces various supplementary definitions and helpful results.

3.2 Consistency

In this section, we discuss in more detail what consistency means and how to check whether a given model is consistent or not with a given dataset. There are two observations that drive the analysis. First, the minimality condition allows one to easily identify adjacent variables: if $(a \not\perp b \mid S)$ for all $S \subset \Omega$, then a and b are adjacent. Second, given a triplet of variables $a - b - c$, a collider $a \rightarrow_t b \leftarrow_t c$ looks differently in the data to chains $a \rightarrow_t b \rightarrow_t c$ and $a \leftarrow_t b \leftarrow_t c$ or a fork $a \leftarrow_t b \rightarrow_t c$. A collider in the true model generates data such that $a \perp c$ and $(a \not\perp c \mid b)$, while the other three options generate the opposite pattern: $a \not\perp c$ and $(a \perp c \mid b)$. We refer to collider patterns in the data as V-structures, defined below.

Definition 2. Variables $a, b, c \in \Omega$ constitute a V-structure in the data $P|\Omega$ if there exists $S \subset \Omega$ such that $(a \perp c \mid S)$ and $(a \not\perp c \mid S \cup \{b\})$. We further label b as the V-center, a and c as the V-parents, and variables in S as controls.

Definition 3. Variables $a, b, c \in \Omega$ constitute a direct V-structure in the data $P|\Omega$ if they constitute a V-structure and there is no $S \subset \Omega$ such that $(a \perp b \mid S)$ or $(b \perp c \mid S)$. If such S exists, we call a, b, c an indirect V-structure.

Verma and Pearl (1990, 2022) show that any pair of models are observationally equivalent (are consistent with the same data) if they have the same sets of adjacencies and colliders. Conversely, a model that has different adjacencies or different colliders relative to the true model would contradict the data. They show that as a consequence, the set of models consistent with data $P|\Omega$ (if it is nonempty) can be obtained by following the Inductive Causation (IC) algorithm described below, which relies on first identifying adjacencies, then identifying colliders $a \rightarrow b \leftarrow c$ from direct V-structures, and finally directing further links using V-structures and acyclicity. We follow the presentation of this algorithm in Pearl (2009).

Definition 4 (IC algorithm). Consider set Ω of variables and data $P|\Omega$. Construct a partially-directed graph C as follows.

1. For any pair $a, b \in \Omega$, link a and b if there is no set $S \subset \Omega$ such that $(a \perp b \mid S)$.

¹⁰One can think that the sender receives utility $|\Omega_t| - |\Omega_s|$ if the receiver accepts a model (Ω_s, C_s) such that $x \rightarrow_s y$ and utility 0 otherwise. Similarly, receiver obtains utility 1 if they behave according to the rules defined here for their type and utility 0 otherwise.

¹¹Alternatively, one could justify this objective via preference for secrecy, where the sender is not willing to reveal information unless absolutely necessary. Such a preference could arise from strategic concerns not explicitly modelled here.

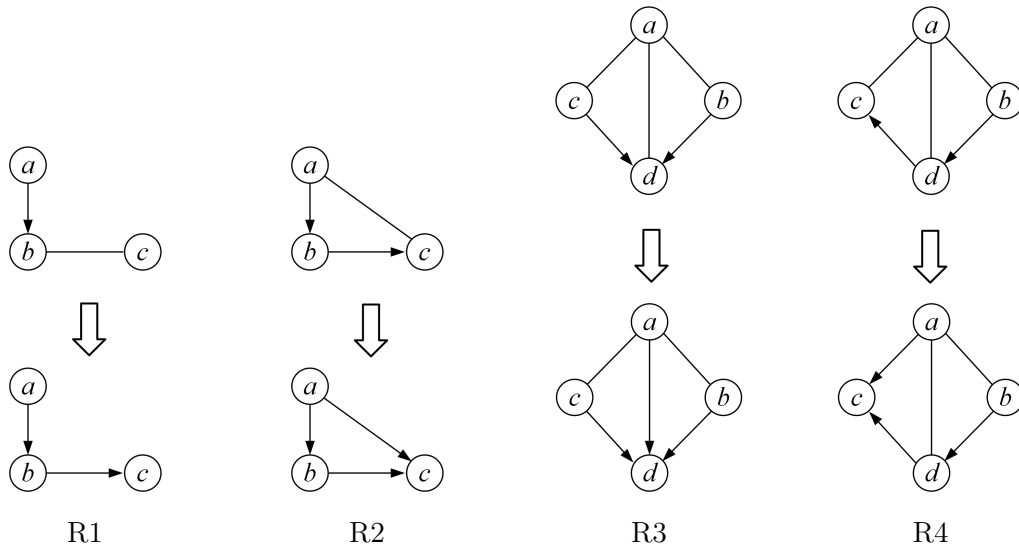


Figure 4: Meek (1995) rules.

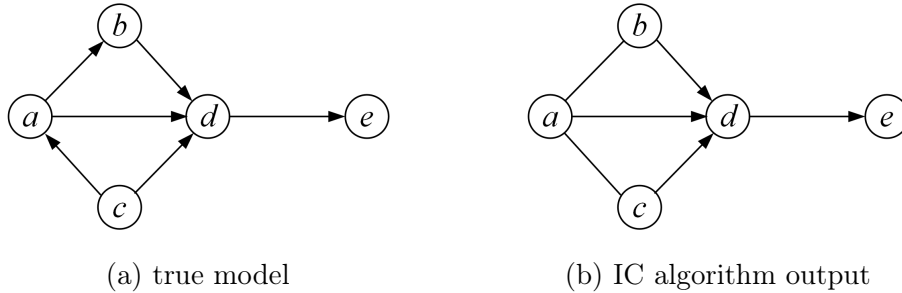


Figure 5: Applying the IC algorithm.

2. For any $a, b, c \in \Omega$ that constitute a direct V-structure, direct the links towards b .
3. For any $a, b \in \Omega$ connected by an undirected link, direct the link from a to b if:
 - (a) link $b \rightarrow a$ creates a V-structure that would have been identified in Step 2, or
 - (b) link $b \rightarrow a$ creates a cycle.

The IC algorithm outputs a partially-directed graph C that corresponds to an equivalence class of observationally-equivalent models. If any consistent model exists, then any directed acyclic selection $C' \subseteq C$ that contains the same adjacencies and no additional V-structures relative to C produces a consistent model, i.e., any such (Ω, C') is consistent with $P|\Omega$. Conversely, no model that contradicts C can be consistent. It follows that any link that was oriented in C is uniquely consistent with the data.

We further use the results of Meek (1995), who argues that Step 3 of the IC algorithm is equivalent to the iterative application of four orientation rules depicted in Figure 4. Hereinafter, we refer to to them as “Meek’s rules” and to the individual rules as R1–R4.

Example 1. Suppose the true model is as in Figure 5(a), and we use the IC algorithm on data generated by it. In Step 1, the algorithm identifies pairs of variables that are never conditionally independent. So it connects the pairs ab , ac , ad , bd , cd , and de with undirected links. In Step 2, it identifies the only V-structure $c \rightarrow d \leftarrow b$ with control a , since $(c \perp b \mid a)$ and $(c \not\perp b \mid a, d)$. In Step 3: (1) we apply R1 to $b \rightarrow d - e$ and direct $d \rightarrow e$; and (2) we apply R3 to the rectangle $\{a, b, d, c\}$ and direct the link $a \rightarrow d$. No

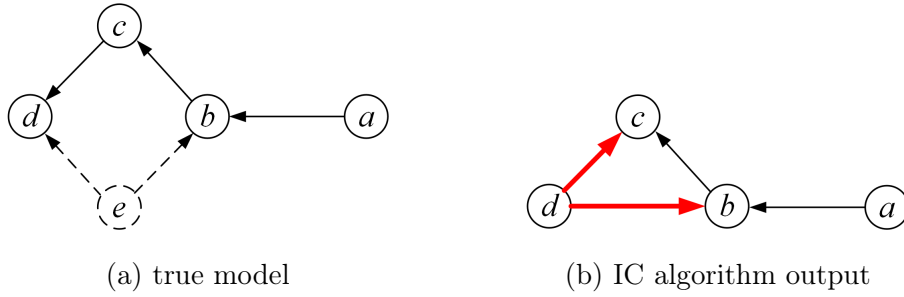


Figure 6: No models are consistent with the data on $\Omega = \{a, b, c, d\}$.

further links can be oriented. The resulting partially directed graph is shown in Figure 5(b).

If for some $\Omega \subset \Omega_t$ there does not exist a model (Ω, C') that is consistent with $P|\Omega$, then the IC algorithm will produce an inconsistent model, as the following example shows.

Example 2. Suppose the true model is as in Figure 6(a). Running the IC algorithm with variables $\Omega = \{a, b, c, d\}$ (so $e \notin \Omega$) produces the model in Figure 6(b). Specifically, in Step 1, the IC algorithm identifies which variables are linked. There is no $S \subseteq \{a, c\}$ such that $(d \perp b \mid S)$, hence d and b are connected. In Step 2, the algorithm identifies the V-structure $d \rightarrow b \leftarrow a$ with control c . In Step 3, the algorithm: (1) orients link $b \rightarrow c$ using rule R1 (which avoids creating a V-structure a, b, c that would have been identified); and (2) orients $d \rightarrow c$ using R2 (which avoids a cycle). However, the resulting model is not consistent with the data, since it violates the Markov property: the model suggests that a and d should be independent, but in the data they are not, $(a \not\perp d)$. This is happening because there does not exist any consistent model for this set of variables Ω .

3.3 Defective links and models

As argued above, the IC algorithm does not always *uniquely* identify the true model, even when all variables are observed. While the adjacencies can be pinned down correctly, not all links can be oriented from the data alone, hence the receiver may be persuaded to accept a model that is consistent with the data yet incorrect. We call such models defective and offer examples below.

Definition 5. Given a model (Ω_s, C_s) , we call $a \rightarrow_s b$ for some $a, b \in \Omega_s$ a defective link if $a \not\perp_t b$. The model (Ω_s, C_s) is defective if it has a defective link.

Example 3. Suppose the true model is $a \rightarrow b$. Running the IC algorithm on $\Omega = \{a, b\}$ yields in Step 1 that $a - b$, but since there are no V-structures in the data, the algorithm stops after Step 1 and leaves the link undirected. Hence, $a \leftarrow b$ is consistent with the data, but such a link is defective.

Example 4. Suppose the true model is as in Figure 3(a). The IC algorithm identifies in Step 1 that $a - b - c$, but since there are no V-structures in the data, the algorithm stops after Step 1 and leaves the links undirected. Then models in Figures 3(b) and 3(c) are consistent yet defective. In turn, model $a \rightarrow b \leftarrow c$ is inconsistent with the data, since it contains a collider but the data would have no corresponding V-structure.

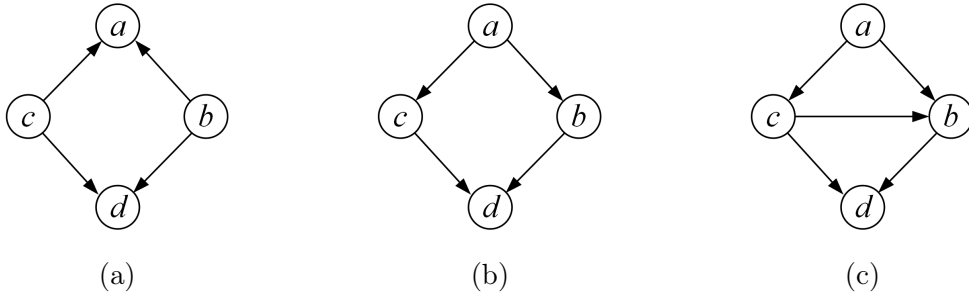


Figure 7: Simple (a and c) and non-simple (b) models.

Example 5. Suppose the true model is as shown in Figure 5(a). The IC algorithm requires the directed links presented in Figure 5(b) but imposes no restrictions on the directions of links $b - a - c$. Hence, a model with $b \rightarrow a \rightarrow c$ would still be consistent, despite both links being defective.

3.4 Simple rich worlds

For some of the results, we impose additional restrictions on the true model to ensure it is sufficiently “nice”. These conditions are stated below, and we discuss the consequences of relaxing them in Section 6.

Definition 6. The true model (Ω_t, C_t) is simple if for any $a, b, c \in \Omega_t$ that form a collider $a \rightarrow_t b \leftarrow_t c$, variables a and c are not correlated.

Definition 7. The true model (Ω_t, C_t) is rich if it is the only model consistent with P .

Simplicity requires that all direct V-structures are “obvious” in the data, in the sense that they can be identified without any controls: for any $a, b, c \in \Omega_t$ such that $(a \perp c \mid S)$ and $(a \not\perp c \mid S, b)$, we have $S = \emptyset$. We assume that even if the true model is simple, the receiver is unaware of this fact. Therefore, we allow the sender’s proposed model (Ω_s, C_s) to be non-simple, so long as it is consistent with $P \mid \Omega_s$.

Example 6. The models in Figure 7(a) and (c) are simple, but the model in (b) is not. In panel (a), $c \rightarrow a \leftarrow b$ and $c \rightarrow d \leftarrow b$ create V-structures, but because $b \perp c$, they can be identified in the data without any controls. In (b), $c \rightarrow d \leftarrow b$ creates a V-structure, but identifying it requires controlling on a , since $c \not\perp b$ and $(c \perp b \mid a)$. In (c), $c \rightarrow d \leftarrow b$ does not create a V-structure because of the link $c \rightarrow b$, which leads to $c \not\perp b$ and $(c \not\perp b \mid a)$. Similarly, the true models in Figures 5(a) and 6(a) are not simple; whereas the true model from the illustrative example in Figure 1(a) is simple.

In turn, richness requires that every link is uniquely consistent with the data, so the true model can be uniquely identified using the IC algorithm. This means that every directed link must either belong to a V-structure, or be orientable by tracing back to a V-structure using Meek’s rules. The models in Figures 1(a), 6(a) and 7(a) are rich, whereas the models in Figures 3(a,b,c), 5(a) and 7(b,c) are not.

4 Feasibility and difficulty of causal persuasion

4.1 The omitted variable power

The sender’s persuasive power in our model comes from two factors: the ability to omit variables and the ability to propose a particular model that ties the revealed variables.

We start by exploring the former. Does variable omission create the potential for causal misperceptions? I.e., what kind of causal connections *can* the receiver get wrong due to not observing all of Ω_t (omitted variable bias)? These questions are partially answered by the following lemma, which imposes a lower bound on the alignment between the true model and any consistent subjective model.

Lemma 1. *If (Ω_s, C_s) is consistent with $P|\Omega_s$, then the following hold for any $a, b \in \Omega_s$:*

1. *a and b are correlated in C_s if and only if they are correlated in C_t ;*
2. *if a and b are adjacent in C_t , then they are adjacent in C_s .*

Proof. The first statement follows from the observation that a, b are correlated in a consistent model if and only if $a \not\perp b$ in the data, and that the true model is consistent.

The second statement: since C_s is consistent, it must align with the output of the IC algorithm. Suppose by contradiction that a, b are not adjacent in C_s . From Step 1 of the IC algorithm, there then exists $S \subset \Omega_s$ such that $(a \perp b \mid S)$. Minimality then implies that a and b cannot be adjacent in the true model. \square

In words, the lemma above says that if two variables are correlated in the true model (i.e., they are connected by a path of chains and forks but no colliders), then they must be correlated as well in any subjective model that only includes a subset of variables but is consistent with the data on these variables. The converse, however, also holds: if x and y are *not* correlated in the true model—implying that $x \perp y$ in the data—then they cannot be correlated in any subjective model consistent with the data. Therefore, the sender cannot persuade the receiver that $x \rightarrow_s y$ in this case, regardless of whether the receiver is sophisticated or naïve.

Example 7. *Suppose the true model is $a \rightarrow_t b \leftarrow_t c$. If the sender discloses all three variables, they constitute a V-structure, hence the true model is uniquely identified by the IC algorithm. If the sender only discloses $\Omega_s = \{a, c\}$, then their proposed model must be such that a and c are not adjacent, since $a \perp c$.*

One possible misperception not ruled out by the lemma is that a consistent subjective model can draw a link between two variables that is nonexistent in the true model, due to omitting other relevant variables. This means that if x and y are correlated in the true model, then Lemma 1 does not preclude the existence of a subjective model (Ω_s, C_s) that includes $x \rightarrow_s y$ and is consistent with the data $P|\Omega_s$. This may be the case even if $x \not\rightarrow_t y$ but, for example, $x \leftarrow_t c \rightarrow_t y$ for some $c \in \Omega_t$, meaning that link $x \rightarrow_s y$ is defective. This is explored in more detail in further sections.

Example 8. *Suppose the true model is $a \rightarrow_t b \rightarrow_t c$, as in Figure 3(a), and the sender only discloses $\Omega_s = \{a, c\}$. Then both $a \rightarrow_s c$ and $c \rightarrow_s a$ are consistent with the data, since $a \not\perp c$. The latter model is defective. The former model, $a \rightarrow_s c$, draws a non-existent link from a to c , but is not defective according to our definition, since $a \Rightarrow_t c$, so the implication that a affects c is correct.*

4.2 Persuading a naïve receiver

We proceed by characterizing the benchmark case of persuading a naïve receiver. Lemma 1 above implies that if x and y are not correlated in the true model (or the data), then no consistent model can include the link $x \rightarrow_s y$. Persuasion is therefore impossible in this case. If, however, x and y are correlated in the true model, then persuading a naïve receiver is trivial. This is captured by the following proposition, which is obvious and left without proof.

Proposition 1. *If $x, y \in \Omega_t$ are correlated in the true model, then a naïve receiver is persuaded by (Ω_s, C_s) such that $\Omega_s = \{x, y\}$, $x \rightarrow_s y$. Otherwise a naïve receiver cannot be persuaded that $x \rightarrow_s y$.*

A naïve receiver is someone who cannot tell correlation from causation. Hence, presenting a pair of correlated variables and claiming that one of them affects the other is sufficient for the sender to persuade such a receiver. No additional variables are required. As we see shortly, this is not the case for a sophisticated receiver.

Example 9. *Suppose vaccination rates for some disease are positively correlated with the number of confirmed cases of that disease, due to more vaccinations in response to (or in anticipation of) an outbreak. Presenting the data on these two correlated variables to a naïve receiver with a claim that vaccines cause the disease would be sufficient to persuade them in this narrative.*

4.3 Persuading a sophisticated receiver

Moving on to a more interesting question: when can the sender persuade a *sophisticated* receiver that $x \rightarrow_s y$, and how? Again, Lemma 1 implies that a necessary condition for persuasion to be possible is that x and y are correlated in the true model. However, this is no longer sufficient. In the example considered above, a sophisticated receiver would be left unconvinced that vaccines cause the disease, since this is not *uniquely consistent* with the data. They would recognize that correlation does not imply causation and would therefore consider the possibility that high case numbers may lead to higher vaccination rates.

Our first result for sophisticated receivers is negative in that it identifies a further case (in addition to the one implied by Lemma 1) in which persuasion is impossible.

Theorem 1. *In a simple world, if $x \leftarrow_t y$ then a sophisticated receiver cannot be persuaded that $x \rightarrow_s y$.*

Proof. See Appendix. □

Theorem 1 argues that the sender can never persuade a sophisticated receiver of reverse causation, i.e., that the causal link goes in the direction opposite to the truth. In Example 9 above, the true model is such that higher case numbers cause higher vaccination rates, meaning there is no genuine data that the sender could pull up to persuade a sophisticated receiver that the vaccine is the source of the disease.

Proof of Theorem 1 in the Appendix proceeds by contradiction, assuming that $x \leftarrow_t y$ and attempting to find a true model (Ω_t, C_t) and a subset of revealed variables Ω_s such that $x \rightarrow_s y$ would be uniquely consistent with $P|\Omega_s$. A link being uniquely consistent with the data means that it must be oriented by the IC algorithm when it is run on $P|\Omega_s$, where the link can be oriented in Step 2 of the algorithm from a direct V-structure in the data or in Step 3 via one of Meek’s rules. We show that in every case, “mis-orienting” a link—orienting a link in direct contradiction of the true model—requires either that the true model is not simple, or that another link had already been mis-oriented by the algorithm. As a consequence, when the true model is simple, there can never be any “first mistake” in the IC algorithm, i.e., no link can be the first one to be mis-oriented.

Moving on from Theorem 1, it is reasonable to ask whether it is possible to persuade a sophisticated receiver if $x \not\leftarrow_t y$. As we show in what follows, under some additional conditions, persuasion is not only possible in this case, but can be done by disclosing at

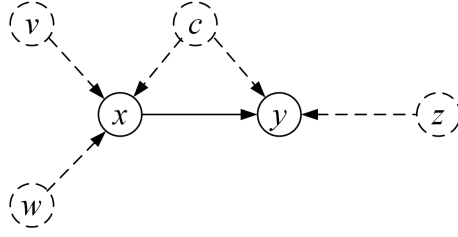


Figure 8: (Ω_t, C_t) with obvious and non-obvious causes.

most two variables in addition to x, y . If such variables exist, we say that persuasion is “easy”. We show later that in case these conditions do not hold, then even if persuasion is possible, it may be “difficult” in that it may require disclosing arbitrarily many variables.

Definition 8. Given $x, y \in \Omega_t$, we call $z \in \Omega_t$ an obvious cause of y given x if $z \in \bar{C}_t(y)$ and z is not correlated with x .

Definition 9. Given $x, y \in \Omega_t$ such that $x \Rightarrow_t y$, we call $w \in \Omega_t$ a non-obvious cause of y given x if $w \in \bar{C}_t(x)$ and x d -separates w and y .

Definition 10. Given $x, y \in \Omega_t$, we call $c \in \Omega_t$ a confounding variable (or a confounder) for (x, y) if C_t contains a path $c \Rightarrow_t x$ that does not pass through y and a path $c \Rightarrow_t y$ that does not pass through x . We denote the set of such confounders as $\tilde{C}_t(x, y) \subset \Omega_t$.

Proposition 2. Suppose $x \not\Rightarrow_t y$.

1. If there exists an obvious cause z of y given x , then a sophisticated receiver can be persuaded that $x \rightarrow_s y$ by a model (Ω_s, C_s) with $\Omega_s = \{x, y, z\}$.
2. If $x \Rightarrow_t y$ and there exist two non-obvious causes v, w of y given x such that $v \perp w$, then a sophisticated receiver can be persuaded that $x \rightarrow_s y$ by a model (Ω_s, C_s) with $\Omega_s = \{v, w, x, y\}$.
3. If $x \Rightarrow_t y$ and there exist a non-obvious cause w of y given x and a confounder $c \in \tilde{C}_t(x, y)$ such that $c \perp w$, then a sophisticated receiver can be persuaded that $x \rightarrow_s y$ by a model (Ω_s, C_s) with $\Omega_s = \{c, w, x, y\}$.

Proof. Figure 8 depicts all three cases.

1. Revealing variables $\Omega_s = \{x, y, z\}$ when z is an obvious cause creates a V-structure $x \rightarrow y \leftarrow z$, hence $x \rightarrow_s y$ is uniquely consistent with $P|\Omega_s$.
2. Revealing variables $\Omega_s = \{v, w, x, y\}$ when v and w are non-obvious causes creates a V-structure $v \rightarrow x \leftarrow w$ since $v \perp w$, and $x \rightarrow_s y$ is then uniquely consistent with $P|\Omega_s$ by Meek’s rule R1.
3. Revealing variables $\Omega_s = \{c, w, x, y\}$ when w is a non-obvious cause and c is a confounder creates a V-structure $w \rightarrow x \leftarrow c$ since $c \perp w$, and $x \rightarrow_s y$ is then uniquely consistent with $P|\Omega_s$ by Meek’s rule R1 (from link $w \rightarrow_s x$).

□

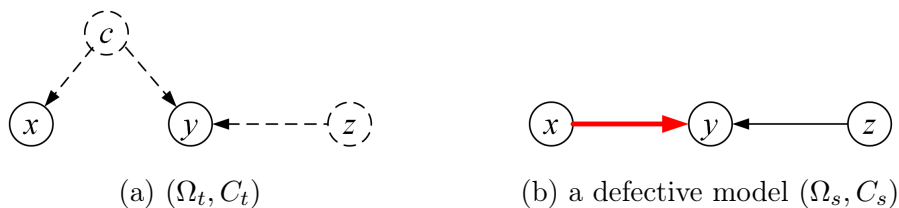


Figure 9: Misleading persuasion using an obvious cause.

Obvious and non-obvious causes describe variables that suffice to persuade a sophisticated receiver. In case an obvious cause exists, revealing it forms a V-structure together with the two variables of interest, which would be identifiable from the data. In case two non-obvious causes exist, they would form a V-structure with x and make $x \rightarrow_s y$ orientable from Meek’s rule R1. The same happens if we have a non-obvious cause and a confounder. We now present some examples of how this result can be applied.

Example 10. Consider screen time and child development, where a child’s higher screen time x has a negative effect on child development y (refer to Figure 8 for the full causal graph, with other variables exemplified further). A public health board (sender) would like to educate parents (receivers) and persuade them that the link is causal. To do so, the board can reveal an obvious cause z of child development outcomes given screen time, such as parental smoking, air pollution, or genetics. This obvious cause would not be correlated with screen time, and would hence form a V-structure with the two variables of interest in the joint data on the three variables. This would prove a causal link between screen time and child development.

Alternatively, the board could reveal a non-obvious cause w , such as school technology adoption (whether tablets or laptops are used in the classroom), together with a confounder c like parental engagement. If these two variables are independent (which may be the case if parents are limited in their choice of schools), they would form a V-structure with screen time in the data, but neither would form a V-structure with screen time and development outcomes. This would prove the link between screen time and development is causal.

Example 11. Consider a politician (sender) running on an anti-immigration platform who seeks to persuade voters (receivers) that higher immigration x causes an increase in crime rates y .¹² In reality, suppose there is no direct causal link between immigration and crime (Marie and Pinotti, 2024); instead, they are correlated due to an unobserved confounder, such as urban density c . The politician may then highlight an obvious cause z of crime—such as weather, bar closing hours, or technological changes (e.g., the adoption of digital wallets reducing cash theft). Because this factor affects crime but is independent on immigration, it creates the appearance of a V-structure in the data with crime as V-center, see Figure 9. This selective disclosure (misleadingly) suggests a causal link between immigration and crime.

What if the premise of Proposition 2 does not hold? Is it easy to persuade a sophisticated receiver if there are no suitable obvious or non-obvious causes? The following example shows that persuasion can be arbitrarily difficult in this case, in the sense of requiring the sender to reveal arbitrarily many variables in addition to x, y .

¹²Donald Trump stated in his presidential campaign announcement speech (June 16, 2015) that immigrants are “bringing drugs, they’re bringing crime, they’re rapists.” Full transcript: <https://time.com/3923128/donald-trump-announcement-speech/>, retrieved Apr 03, 2026.

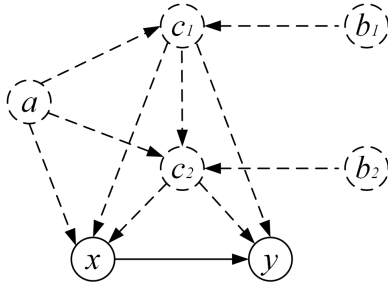


Figure 10: Persuasion without a cause.

Example 12. Suppose the true model is such that $\Omega_t = \{x, y, a, b_1, \dots, b_n, c_1, \dots, c_n\}$, and the causal graph is given by $\{a, b_i\} \rightarrow_t c_i \rightarrow_t \{x, y\}$ for all $i \in \{1, \dots, n\}$, $a \rightarrow_t x$, $x \rightarrow_t y$, and $c_i \rightarrow_t c_j$ for any $i > j$. Such a model for $n = 2$ is depicted in Figure 10. This model contains no obvious or non-obvious causes of y given x (all other variables are confounding for x, y). Further, the model is rich, so disclosing the true model would persuade a sophisticated receiver. To see this, run the IC algorithm on the full dataset P . Then all undirected links are identified in Step 1. Step 2 identifies V -structures $a \rightarrow c_i \leftarrow b_i$ for all i , and $c_i \rightarrow c_j \leftarrow b_j$ for all i, j with $i > j$.¹³ Step 3 of the IC algorithm then orients, for all i :

$$\begin{aligned} b_i &\rightarrow c_i \xrightarrow{R1} c_i \rightarrow \{x, y\}, \\ a &\rightarrow c_i \rightarrow x \xrightarrow{R2} a \rightarrow x, \\ a &\rightarrow x \xrightarrow{R1} x \rightarrow y. \end{aligned}$$

Persuading a sophisticated receiver that $x \rightarrow_s y$ in this case requires revealing all variables in $\{c_1, \dots, c_n\}$. This is because orienting $x \rightarrow_s y$ can only be done using R1 as above, which requires that a is not adjacent to y in (Ω_s, C_s) , meaning that $\Omega_s \supset \{c_1, \dots, c_n\}$.¹⁴ Since n is arbitrary, this means that arbitrarily many variables may be required to persuade a sophisticated receiver in this case.

This case reflects the core challenge faced by a large portion of empirical economics, which seeks to establish causal relationships between variables of interest. The sheer size of this literature underscores the difficulty of the task, as these variables are typically jointly determined by a large number of confounding factors. Therefore, credible identification (and persuasion) requires conditioning on a rich set of covariates or exploiting exogenous variation.¹⁵ Consequently, persuading a sophisticated receiver in such a setting would require the sender to provide information at the level of an advanced empirical economist.

To conclude, the feasibility and the difficulty of persuasion depend on both the sophistication of the receiver and the topology of the causal graph. Persuading a naïve receiver is trivial, as any correlation can be presented as causal, so long as this correlation exists in the data. In contrast, persuading a sophisticated receiver is feasible in a smaller range

¹³In all of the respective colliders in the true model, pairs (a, b_i) and (c_i, b_j) are not correlated, meaning that the model is also simple.

¹⁴Orienting $x \rightarrow_s y$ further requires that $a \rightarrow_s x$, which can also only be oriented using R2 as in the true model and thus requires that $a \rightarrow_s c_i \rightarrow_s x$ for some i were oriented. The latter requires that $b_i \in \Omega_s$ for some i .

¹⁵A canonical example is the estimation of the returns to education, where ability, family background, and other unobserved characteristics confound the relationship between schooling and earnings; see, for instance, Card (1999).

of scenarios (only when the true model does not contradict the proposed narrative), and always involves revealing additional variables. In the best case, only one or two suitable (obvious or non-obvious) causes need to be revealed in addition to the variables of interest x and y to create a persuasive narrative. But in the absence of such variables, persuasion may become arbitrarily difficult. Moreover, if the world is not (sufficiently) rich, even full disclosure of the true model may fail to persuade a sophisticated receiver.

5 Persuasion and debunking of pre-existing models

The previous section investigated how to persuade a receiver with a blank mind. However, what if the receiver already has a subjective model in mind by the time the sender enters the picture? Can the sender change the receiver’s mind? As we show in this section, debunking a receiver’s model is qualitatively similar to persuading a sophisticated receiver, even when the receiver is actually naïve. We begin by introducing the modified problem.

5.1 Modified setup: Receiver with a pre-existing model

Suppose the receiver is initially aware of a subset $\Omega_r \subset \Omega_t$. This means that they only observe the data $P|\Omega_r$ on variables in Ω_r but not other variables. The receiver further has some subjective model of the world (Ω_r, C_r) that is consistent with $P|\Omega_r$.

As before, the sender knows the true model (Ω_t, C_t) and can propose a model (Ω_s, C_s) to the receiver. The sender can propose any model (Ω_s, C_s) that is consistent with $P|\Omega_s$ and is such that $\Omega_r \subseteq \Omega_s \subseteq \Omega_t$ (the sender knows the set of variables known by the receiver and cannot make the receiver forget any variables). If (Ω_r, C_r) is not consistent with $P|\Omega_s$, we say that the sender *debunks* the receiver’s model (Ω_r, C_r) , and conditional on that:

- a *naïve* receiver accepts any model (Ω_s, C_s) that is consistent with $P|\Omega_s$;
- a *sophisticated* receiver accepts model (Ω_s, C_s) if and only if every link along every path between x and y in (Ω_s, C_s) is uniquely consistent with $P|\Omega_s$.

Otherwise—if (Ω_s, C_s) does not debunk (Ω_r, C_r) —the receiver rejects the sender’s model for sure.

We again assume that the sender’s objective is to *persuade* the receiver—to accept a model (Ω_s, C_s) —such that either $x \Rightarrow_s y$ for some $x, y \in \Omega$, or that x and y are not adjacent (both cases are considered).¹⁶ We focus on the interesting case where the receiver is initially aware of both variables, $x, y \in \Omega_r$, but their existing model does not satisfy the sender’s objective. Further, we maintain the sender’s secondary objective to persuade using the smallest number of new variables, i.e., to minimize $|\Omega_s| - |\Omega_r|$.

For the receiver’s choice procedure above to be well-defined, we need to define consistency for situations when the dataset $P|\Omega_s$ covers more variables than are included in the model (Ω_r, C_r) , i.e., $\Omega_r \subset \Omega_s$. We say that (Ω_r, C_r) is consistent with $P|\Omega_s$ when $\Omega_r \subset \Omega_s$ if there exists a consistent model (Ω_s, C_s) such that for all $a, b \in \Omega_r$, if $a \rightarrow_r b$ then $a \Rightarrow_s b$. For instance, take our illustrative example in Section 2. The employee’s subjective model after hearing the business school advertisement is as in Figure 1(b). If the employer reveals variable a , the employee’s subjective model does not contradict the new data, and they simply extend the subjective model to the one in Figure 2(b).

¹⁶Our results regarding when and how the sender can successfully rule out a link between x and y also apply directly to the persuasion problem considered in Section 4. This case was not covered previously because it is difficult to think of a real-world setting in which the sender would want to reveal two variables that the receiver was not aware of, and try to prove to the receiver that the two are not causally related.

The sender’s problem in this modified model is two-fold. First, they choose which new variables $\Omega_s \setminus \Omega_r$ to present to the receiver to render their initial model (Ω_r, C_r) inconsistent with the new data. This incompatibility would be contained in one or more causal links from the receiver’s model that cannot exist in any model consistent with the new evidence. We then say that model (Ω_s, C_s) *debunks a link* $a \rightarrow_r b$ for some $a, b \in \Omega_r$ if there is no model consistent with $P|\Omega_s$ such that $a \Rightarrow b$. The receiver’s model (Ω_r, C_r) is debunked if any of its links is debunked. Results in Section 5.2 discuss when and how the receiver’s model can be debunked. The second part of the sender’s problem is proposing a new model C_s that is consistent with the data and is then accepted by the receiver of a given type. Sections 5.3 and 5.4 discuss when and how the sender can successfully persuade the receiver.

5.2 Debunking the receiver’s model

Debunking a receiver’s pre-existing model requires debunking at least one causal link. This needs to be done regardless of the receiver’s type, and the task in either case is essentially equivalent to persuading a sophisticated receiver about that link, as characterized in Section 4.3. This is because debunking a link $x \leftarrow_r y$ requires revealing such data that this link is not consistent with it—which means either rendering $x \rightarrow_s y$ (or $x \Rightarrow_s y$) uniquely consistent with the data or showing that x and y are correlated purely due to confounders but $x \not\Leftarrow_s y$ and $x \not\Rightarrow_s y$ (this option is explored in Section 5.4). Consequently, our results regarding what a sophisticated receiver can and cannot be persuaded of have direct implications for which models can and cannot be debunked. We start by stating a corollary of Theorem 1, showing that our notion of a defective link (and a defective model) provide a necessary condition for what can be debunked.

Corollary 1. *In a simple world, the sender can never debunk a non-defective link.*

Corollary 1 argues that the sender can debunk only defective links in the receiver’s model, which is the causal relations that the receiver believes in, $x \leftarrow_r y$, but which do not exist in reality, $x \not\Leftarrow_t y$. In particular, if the receiver’s model $x \leftarrow_r y$ is incorrect only in that it omits some proxy variable a (meaning that the true model is such that $x \leftarrow_t a \leftarrow_t y$), then revealing a would only make the receiver expand their model, but not abandon it. This result immediately implies the following.

Corollary 2. *In a simple world, the sender can never debunk a non-defective model.*

The latter corollary says that if the receiver’s model is a simplified version of the true model in the sense that it may omit some variables, but does not contain any defective links, then such a model cannot be debunked by the sender. If the true model is rich, the converse is also true: any defective receiver’s model can be debunked by presenting the true model (Ω_t, C_t) . In this case, the question becomes: how difficult is it to debunk a defective model? Is it necessary to reveal the whole true model, or would revealing a subset of variables suffice?

We next show that when the true model is simple and rich, Proposition 2 can be extended to debunking models. Specifically, we argue that obvious or non-obvious causes can be used to “easily” debunk a defective model. Case 1 of Theorem 2 below corresponds to Case 1 of Proposition 2, whereas Case 2 of the theorem combines the other two cases from the proposition.

Theorem 2. *Suppose the true model is simple and rich, and the receiver’s model has defective link $x \leftarrow_r y$.*

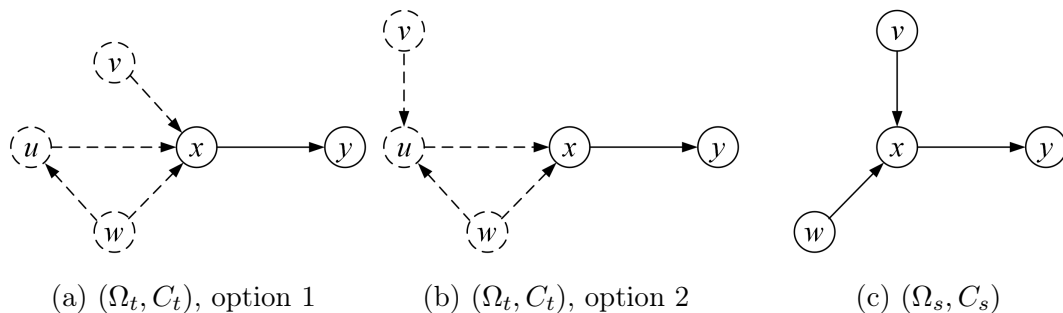


Figure 11: Debunking with a non-obvious cause.

1. If there exists an obvious cause of y given x , then the receiver's model can be debunked by revealing one new variable (the obvious cause).
2. If $x \Rightarrow_t y$ and there exists a non-obvious cause of y given x , then the receiver's model can be debunked by revealing at most two new variables.

Proof. See Appendix. □

Theorem 2 generalizes Proposition 2 to the case of the receiver having an arbitrary pre-existing model. In Proposition 2, the receiver did not have any model in mind, so it was safe to assume the causes would be adjacent to the variables of interest. Here, by contrast, the receiver already holds a subjective model involving additional variables that might stand between the candidate causes and the variables of interest. Nevertheless, the result continues to hold.

The proof of Theorem 2, much like that of Proposition 2, proceeds by constructing a V-structure either upstream of y (using an obvious cause) or upstream of x (using a non-obvious cause). The presence of such a V-structure then permits the sequential application of Meek's rules to orient the desired link $x \rightarrow_s y$. The existence of the required V-structure and the identifiability of the target link are ensured by the simplicity and richness of the true model. In particular, the richness of the true model is only used in the second part of the theorem, to guarantee that either a second non-obvious cause, or a confounding variable exist. Alternatively, the result can be formulated without relying on richness, in a way that is equivalent to Proposition 2. We now present a few examples demonstrating how Theorem 2 can be applied.

Example 13. To illustrate the second part of Theorem 2, suppose the true model is as shown in Figure 11(a), and the receiver's model is $\Omega_r = \{x, y\}$ with $x \leftarrow_r y$. In this case, no obvious cause of y given x exists. However, because the world is rich, we must be able to trace the link $x \rightarrow y$ back to some V-structure upstream from x . In this case, revealing either v and w , or v and u achieves the purpose. However, note that the necessary V-structure does not have to contain only direct parents. If the true model is as shown in Figure 11(b), then by revealing v and w , the sender will be able to debunk the defective link as well, which will result in a subjective model as in Figure 11(c).

We now further present a few more applied examples of how causal misperceptions can be debunked. For the examples that follow, we assume that the true model (Ω_t, C_t) is as in Figure 8, the receiver's initial model is given by $x \leftarrow_r y$, and all the respective variables are as in different rows of Table 1. A disclaimer is needed that the problems discussed in these examples are complex and multi-faceted. The authors do not claim to have any easy answers and do not claim that the "true models" in these examples capture reality

v, w	x	y	z
economy, foreign policy	voters preferences	poll results	poll-day weather
party lines, demographics, culture, voting system	polarization	SoMe algorithms	platform design
supply shocks (wars, nat. disasters) demand shocks (mark-up, foreign demand)	inflation	interest rates	β -shock, demographics

Table 1: Examples of obvious and non-obvious causes.

accurately. Instead, we assume for sake of the argument that the true causal links in every model go in a certain way, and demonstrate how these links could be proven by causal narratives.

Example 14. *Suppose an autocrat attempts to suppress the results of the pre-election public opinion polls under the pretense that poll results y affect voters’ preferences x and thereby bias the election outcomes. The opposition (sender) wants to debunk this narrative to persuade the voters (receivers) to mobilize and protect electoral transparency and free speech. Suppose further the autocrat’s narrative is indeed false (defective) in this example, and voters’ preferences are correlated with poll results simply because the latter accurately represent the former ($x \rightarrow_t y$). The opposition can then debunk it by revealing that weather on the day of the poll biases the poll results without affecting the voters’ actual preferences.¹⁷ This makes weather an obvious cause z in this example. Alternatively, the opposition could reveal two non-obvious causes v, w , which is any pair of independent variables that affect public opinion but have no effect on poll results, except through changes in public opinion. Such variables can include state of the economy (growth, inflation, or unemployment) and foreign policy events (wars and conflicts or trade agreements and expressions of support).*

Example 15. *Consider now a politician (receiver) who believes that the recommendation algorithms of social media platforms y lead to social polarization x which is undesirable and warrants regulating platforms. Suppose the truth is the converse: that polarized individuals simply prefer polarized media sources, and by engaging with such content “teach” the recommendation algorithm (Robertson et al., 2023). The platform (sender) would like to debunk the receiver’s view and to show the truth. One way to do so is to reveal an obvious cause z , such as some aspect of platform design that affects the composition of the users’ feed (together with the data on an internal experiment varying this aspect and relating it to which items users engage with). An alternative approach would be to reveal two non-obvious causes v, w that do not affect social media recommendations except through their effect on social polarization. These could be the sentiment of politicians’ statements and cross-sectional variation from demographics or culture.*

Example 16. *Suppose a voter (receiver) believes that the interest rate y set by the central bank drives inflation x , seeing how the two are correlated. The conservative central bank (sender) would like to explain that its monetary policy actually responds to inflation in an*

¹⁷Damsbo-Svendsen and Hansen (2023) show in a meta-analysis that weather on the voting day affects voting behavior. By a similar logic, it is reasonable to argue that weather on a polling day affects the poll results.

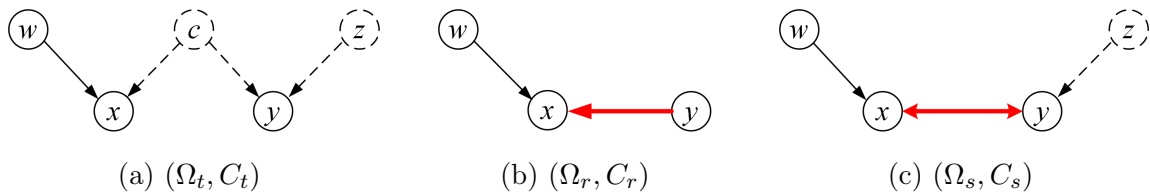


Figure 12: Debunking may not lead to persuasion.

attempt to curb the business cycles. To do so with data, the sender can show another factor z that affects interest rates but is not correlated with inflation, which would be the obvious cause in this case. Any variable affecting the savings rate, such as ageing population¹⁸ or a shock to the intertemporal discount factor β , would be the natural candidate. Alternatively, the central bank could disclose independent, non-obvious determinants v, w of inflation, such as aggregate supply and demand shocks, including geopolitical disruptions, pandemics, natural disasters, as well as mark-up shocks and foreign demand shocks. These affect inflation without entering the policy rule directly, and therefore influence interest rates only through the central bank's endogenous response.

Note that Theorem 2 focuses purely on debunking and does not incorporate the constraint that the sender must offer a new consistent model, which may require disclosing more variables even in the cases covered by Theorem 2. An example is presented below.

Example 17. Suppose the true model is as in Figure 12(a), and the receiver's model is $w \rightarrow_r x \leftarrow_r y$, depicted in Figure 12(b). This model can be debunked by revealing just one variable, z , which is an obvious cause of y given x . However, then both w, x, y and x, y, z constitute V -structures, with the former implying $x \leftarrow y$ and the latter implying $x \rightarrow y$, as depicted in Figure 12(c). These contradicting implications mean that there is no consistent model for $\Omega = \{w, x, y, z\}$. If the sender is required to propose a consistent model, then to debunk the receiver's model they must also disclose confounder c in addition to the obvious cause z and propose the true model from Figure 12(a).

Similarly to persuasion, if the premise of Theorem 2 does not hold, it may be difficult to debunk the receiver's defective model. This is in the sense of requiring the sender to reveal arbitrarily many new variables. Suppose that the receiver's model is $x \leftarrow_r y$ and is defective, meaning $x \not\leftarrow_t y$. This opens up two cases: either $x \Rightarrow_t y$, or $x \not\Rightarrow_t y$. In the former case, Example 12 applies, which demonstrates that the sender may need to reveal arbitrarily many variables to debunk the receiver's model $x \leftarrow_r y$. The latter case is discussed in Section 5.4.

5.3 Persuading a receiver with a pre-existing model

The previous subsection asked the question of "when can a defective model be (easily) debunked?" We now proceed to answer the main question: "when can the sender persuade the receiver?" The sender wants to persuade the receiver that $x \Rightarrow_s y$. If the receiver already believes that $x \Rightarrow_r y$ then nothing needs to be done. Hence we look at the interesting cases, which are when the receiver's initial model is such that $x \not\Rightarrow_r y$. We consider different cases corresponding to whether the respective link is present in the true model or not.

¹⁸"The impact of population ageing on monetary policy." Vox EU, Mar 05, 2019. URL: <https://cepr.org/voxeu/columns/impact-population-ageing-monetary-policy>

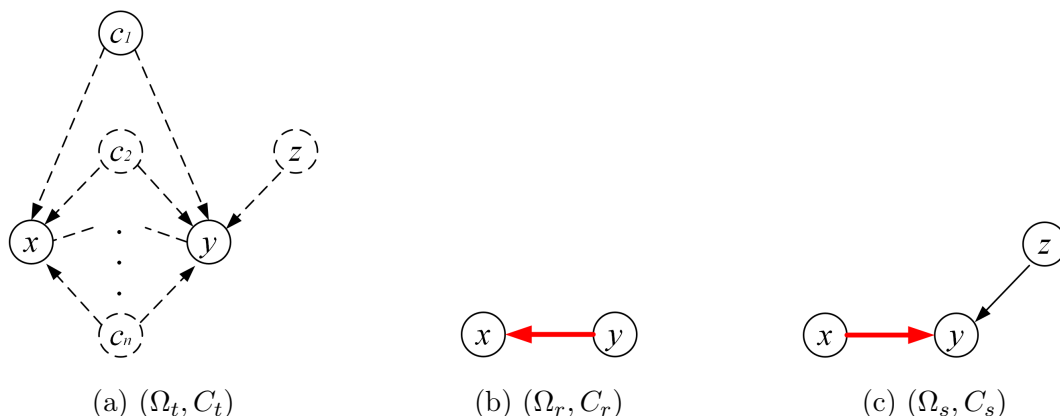


Figure 13: Replacing a defective model with a defective model.

Truth is on the sender’s side. In a rich world, the sender can always persuade the receiver that $x \Rightarrow_s y$ when truth is on their side—i.e., the true model (Ω_t, C_t) is such that $x \Rightarrow_t y$. By definition of richness, this can be done by revealing the true model. In other words, while persuasion may be difficult (due to the true model containing many variables and links), it is possible in principle. Further, it follows immediately from Theorem 2 that if an appropriate cause of y exists, then such persuasion is, in fact, easy, in the sense of only requiring the sender to reveal one or two additional variables. This applies regardless of whether the receiver is naïve or sophisticated. The result is summarized by the following corollary.

Corollary 3. *If the true model is simple and rich and such that $x \Rightarrow_t y$ and there exists an (obvious or non-obvious) cause of y given x , the sender can persuade the receiver that $x \Rightarrow_s y$ with at most two variables.¹⁹*

Spurious correlation. If the true model is such that x and y are not adjacent and $x \not\Rightarrow_t y$, $x \not\leftarrow_t y$, but rather there exist some confounders $c \in \tilde{C}_t(x, y)$ such that $c \Rightarrow_t x, y$, then two cases are possible. If x and y are not adjacent in the receiver’s model, the sender can do nothing. This is because the receiver is aware of sufficiently many other variables to realize there is no direct link between x and y . The sender cannot change that regardless of the receiver’s type, since revealing additional variables can only destroy some links in the receiver’s model but cannot create new spurious correlations and links.

If, on the other hand, $x \leftarrow_r y$ (due to the receiver being unaware of some of the confounders $c \in \tilde{C}_t(x, y)$), then the sender can flip this link and persuade the receiver that $x \Rightarrow_s y$, as is illustrated by the following example. This is possible even though there is no actual link between x and y in the true model, meaning the sender debunks the receiver’s defective model and replaces it with another defective model. This is also possible regardless of whether the receiver is naïve or sophisticated.

Example 18. *Suppose the true model is as depicted in Figure 13(a), and the receiver’s model is $x \leftarrow_r y$, as in Figure 13(b). Then the sender can persuade the receiver that $x \rightarrow_s y$ by revealing variable z (for the proposed model see Figure 13(c)), in which case x, y, z form a V-structure in the data.*

¹⁹Note specifically that the assumption $x \Rightarrow_t y$ rules out the situation described in Figure 12 and Example 17, where revealing an obvious cause produced an inconsistent model.

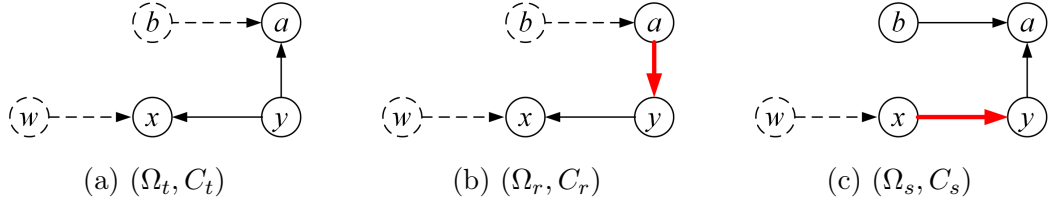


Figure 14: Persuasion by nitpicking.

However, such deception is only possible if the true model allows it and if the receiver is sufficiently unaware. In particular, the receiver must not be aware of enough variables to realize that x and y are not adjacent, as discussed above. But on top of this, the receiver’s (un)awareness must accommodate the sender’s defective model, as illustrated by the following example.

Example 19. *Suppose the true model is as depicted in Figure 12(a), and the receiver’s model is as in Figure 12(b). As argued in Example 17, while variable z is an obvious cause of y given x , revealing it does not allow the sender to persuade the receiver that $x \rightarrow y$, since there is no consistent model for $\Omega_s = \Omega_r \cup \{z\} = \{w, x, y, z\}$. If the sender must present a consistent model, their only choice is between leaving in place the receiver’s model from Figure 12(b) with $x \leftarrow_r y$ and revealing the true model from Figure 12(a) with x not adjacent to y .*

Truth is against the sender. If the true model is such that $x \leftarrow_t y$, then Corollary 1 shows that the sender can never debunk *this* link and present such data that only $x \Rightarrow_s y$ is consistent with it. This implies that a sophisticated receiver cannot be persuaded in this case.

However, this does not mean that persuasion is impossible with a naïve receiver, since the sender may still be able to target some other defective link in the receiver’s model. If such a link exists, the sender can “nitpick” the naïve receiver’s model: debunk some defective link that is not directly related to the targeted link, and then use this opportunity to deceive the receiver. The following example demonstrates how such “persuasion by nitpicking” can work.

Example 20. *Suppose the true world is as shown in Figure 14(a). The naïve receiver observes variables a , x and y and thinks the model is as shown in Figure 14(b). Specifically, the receiver correctly believes that $x \leftarrow_r y$, but also mistakenly believes that $a \rightarrow_r y$. The sender would like to deceive the receiver and persuade them that $x \rightarrow_s y$. Because $x \leftarrow_t y$ is the truth, the sender cannot debunk this link directly. Instead, the sender can debunk the defective link $a \rightarrow_r y$. In this example, the sender is able to reveal variable b , which is the obvious cause for a given y . The sender then will offer the model shown in Figure 14(c), which will be accepted by the naïve receiver.*

However, as before, successful deception along the lines of the example above requires that the link of interest must not be uniquely consistent with $P|\Omega_r$. Instead, models with both $x \Rightarrow y$ and $x \Leftarrow y$ must be consistent with the data available to the receiver. Otherwise, if the naïve receiver can unambiguously infer from the data they observe that $x \Leftarrow y$, then no nitpicking can help the sender persuade them that $x \Rightarrow y$. Further, the scope for such deception also depends on the relationship between the link of interest and the defective link. If both links trace back to the same V-structure, deception may not be possible, as demonstrated by the following example.

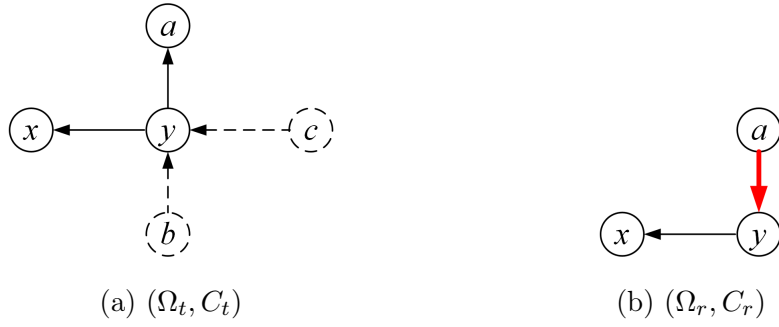


Figure 15: Example: persuasion by nitpicking not possible.

Example 21. Suppose the true model is as shown in Figure 15(a), and the naïve receiver’s subjective model is as in Figure 15(b). The receiver correctly believes $x \leftarrow_r y$ but incorrectly believes $a \rightarrow_r y$. The sender would like to persuade the receiver that $x \rightarrow_s y$. However, in this case, the targeted link $a \leftarrow_t y$ and the link of interest $x \leftarrow_t y$ both trace back to the same V -structure c, y, b . Thus, to debunk $a \rightarrow_r y$, the sender has to reveal b and c . But revealing b and c also makes $x \rightarrow_s y$ inconsistent with the data, so deception is not possible.

5.4 Dissuading a receiver with a pre-existing model

What if the sender wants to rule out a link between x and y ? In other words, the receiver believes that $x \leftarrow_r y$, and the sender wants to persuade them that there is no link between x and y (i.e., propose a model such that $x \not\leftarrow_s y$ and $x \not\rightarrow_s y$). This could, for example, be the case when the sender tries to debunk a defective model fuelled by spurious correlation, which has previously been imprinted on the receiver.

Example 22. In the beginning of the 20th century, a major outbreak of localized paralytic polio epidemics led to over 27,000 cases and more than 6,000 deaths in the US. Due to recurring summer epidemics, by the 1940s-1950s polio was widely feared as a devastating threat (Mehndiratta et al., 2014). During that period, Dr. Benjamin Sadler was a strong proponent of the theory that refined sugars cause polio in children. His advocacy resulted in a substantial drop in sales of ice cream, cookies and Coca Cola in Asheville, NC, where he practiced (Williams, 2013). His theory was based on the observation that polio epidemics occur in summers, when children increase their sugar intake (ice cream in particular). In reality, as it was later discovered, polio spread primarily through fecal-oral transmission facilitated by close contact and poor hygiene.²⁰ Warmer weather was the confounding variable behind the correlation observed by Sadler: it led to increased social interactions, particularly in swimming pools and rivers (which fostered polio transmission), as well as increased ice cream consumption.

Ruling out a direct link between two correlated variables x and y requires the sender to demonstrate the proxy and/or confounding variables that capture all sources of correlation between x and y . Without such complete disclosure, x and y would still be correlated from the receiver’s perspective conditional on any set of other variables known to them, and hence the receiver would still draw a direct causal link between x and y . In other words, the sender needs to reveal a set of variables that d-separates x and y (Definition 1).

²⁰“Clinical Overview of Poliomyelitis” CDC, May 9, 2024. URL: <https://www.cdc.gov/polio/hcp/clinical-overview/index.html>, retrieved April 3, 2026.

The statement is formalized in the following proposition, which is trivial and is presented without proof.

Proposition 3. *Suppose the receiver’s model contains a link $x \leftarrow_r y$. This link is debunked by a sender’s model (Ω_s, C_s) where x is not adjacent to y only if the latter contains a set of variables $S \subset \Omega_s$ that d-separates x and y . Further,*

1. *if $x \leftarrow_t y$ or $x \rightarrow_t y$, then no such set $S \subset \Omega_t$ exists;*
2. *if x and y are not adjacent in (Ω_t, C_t) , then such a set $S \subset \Omega_t$ necessarily exists.*

In some situations, it is trivial that no such d-separating set exists. Specifically, if $x \leftarrow_t y$ or $x \rightarrow_t y$, then Lemma 1 implies that x and y must be adjacent in any consistent subjective model as well. It follows that it would not be possible to dissuade the receiver of a link between x and y in this case, regardless of the receiver’s type.

If x and y are not adjacent (but still correlated) in the true model, then revealing a d-separating set is necessary to rule out a direct link between the two variables in the receiver’s model. Even though it is not sufficient, as we argue below, this necessary task may already be difficult. The d-separating set can be arbitrarily large, which is demonstrated by the following simple example.

Example 23. *Suppose the true model is as represented in Figure 13(a). Then the only set that d-separates x and y is $S \equiv \{c_1, \dots, c_n\}$. If we assume that the receiver’s model is given by $\Omega_r = \{x, y\}$ and $x \leftarrow_r y$, as in Figure 13(b), then revealing the true (lack of) causal relation between x and y requires revealing all variables $\{c_1, \dots, c_n\}$. By increasing n , we can make the number of such variables arbitrarily large.*

Moreover, finding a *minimal* set S that d-separates given x, y is, in general, a difficult problem. It is well known as the Minimum Set Cover Problem, which seeks the smallest collection of subsets whose union covers a given universe, and it was shown to be NP-complete (Karp, 2009). Exact solutions typically rely on integer programming and branch-and-bound methods (Nemhauser and Wolsey, 1988). For large instances, approximation algorithms (Johnson, 1974) and linear programming relaxations with rounding techniques (Lovász, 1975) are often used.

As alluded to previously, revealing a d-separating set is by itself not necessarily sufficient to debunk the receiver’s model: revealing S breaks the direct link between x and y , but may not help identify the directions of the links between x, y and all the d-separating variables. For instance, if the true model is $x \leftarrow_t c \rightarrow_t y$ and the receiver starts with a model $x \leftarrow_r y$, then revealing c does not debunk the receiver’s model. In particular, the receiver would regard the data on x, y, c as consistent with an extended model $x \leftarrow_r c \leftarrow_r y$ and reject the sender’s model claiming $x \leftarrow_s c \rightarrow_s y$.

However, if truth is on the sender’s side in the sense that $x \not\leftarrow_t y$ and $x \not\rightarrow_t y$, and the d-separating set S reveals sufficiently many V-structures, then persuasion may be possible. For instance, in Example 23, set S contains multiple independent confounders, which induce the V-structures that allow the receiver to infer that $x \leftarrow_s c_i \rightarrow_s y$ without a doubt for all $i \in \{1, \dots, n\}$. This rules out the possibility that $x \leftarrow_r y$ and debunks the receiver’s model. Further, since this orients all links on all paths between x and y , even a sophisticated receiver would then accept a model with $\Omega_s = \{x, y, c_1, \dots, c_n\}$ and $x \leftarrow_s c_i \rightarrow_s y$ for all c_i .

At the same time, note that if the sender’s goal is merely to debunk the receiver’s original model $x \leftarrow_r y$ (without regard as to which alternative model is put in place), then truthful persuasion may not be the best course of action when $x \not\leftarrow_t y$ and $x \not\rightarrow_t y$. As discussed in Section 5.3, the strategy of replacing one defective model with another may

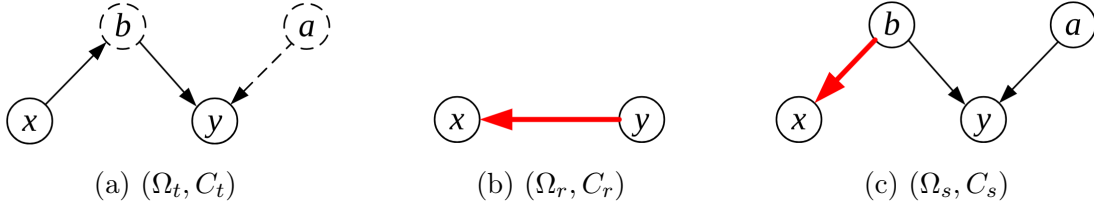


Figure 16: Ruling out a defective link: $x \leftarrow_r y$, $x \Rightarrow_t y$.

indeed be more feasible than persuading the receiver that x and y do not have a causal relation. Indeed, the same receiver’s model is debunked in Example 18 by a (defective) model that reveals one new variable, while debunking it in Example 23 by a non-defective model requires revealing arbitrarily many new variables. Persuading with a defective model may thus be easier than proving the truth, regardless of the receiver’s type.

Finally, we move on to the case when the sender wants to persuade the receiver that $x \not\leftarrow_s y$ and $x \not\Rightarrow_s y$ but the truth is against the sender in that either $x \leftarrow_t y$, or $x \Rightarrow_t y$. In this case, Theorem 1 suggests that sender can never prove conclusively that x and y are correlated purely due to confounders. This then implies that a sophisticated receiver can not be persuaded. However, if the receiver is naïve, then just like in Example 20 above, the sender may still be able to persuade them that x and y do not affect one another. If the receiver believes $x \leftarrow_r y$ and this link is defective, meaning $x \Rightarrow_t y$, then the sender may be able to debunk this link and offer a fitting consistent model with confounders. Otherwise, if $x \leftarrow_t y$ and the receiver’s model is correct, then “persuasion by nitpicking” may still be possible, with the sender debunking some irrelevant defective link that the receiver’s model may have. The following examples demonstrate these possibilities in the two respective cases (when $x \Rightarrow_t y$ and $x \leftarrow_t y$).

Example 24. To see how the sender can rule out a link between x and y when $x \leftarrow_r y$ and $x \Rightarrow_t y$, see Figure 16. The true model is presented in panel (a). The naïve receiver’s model in panel (b) has defective link $x \leftarrow_r a$, whereas in the true model $x \Rightarrow_t y$. The sender can propose the model in Figure 16(c), debunking the naïve receiver’s defective model: V -structure b, y, a immediately implies $b \rightarrow_s y \leftarrow_s a$, which is incompatible with $x \leftarrow_s y$ and b being only a proxy for this channel. The sender can then instill the belief that $x \leftarrow_s b \rightarrow_s y$, contrary to the true model, since this model is consistent with $P|\Omega_s$ despite being defective.

Example 25. To see how the sender can rule out a link between x and y when $x \leftarrow_r y$ and $x \leftarrow_t y$, see Figure 17. The true model is presented in panel (a) and is simple and rich. The naïve receiver’s model in panel (b) has defective link $x \rightarrow_r a$, whereas in the true model $x \leftarrow_t a$. The sender can exploit this and reveal variables b, d , proposing the model in panel (c), which is also defective but compatible with the data $P|\Omega_s$. It debunks the naïve receiver’s defective model and instills the belief that $x \leftarrow_s d \rightarrow_s y$, contrary to the true model.

6 Discussion

In this section, we discuss various assumptions behind the model and the potential consequences of relaxing these assumptions.

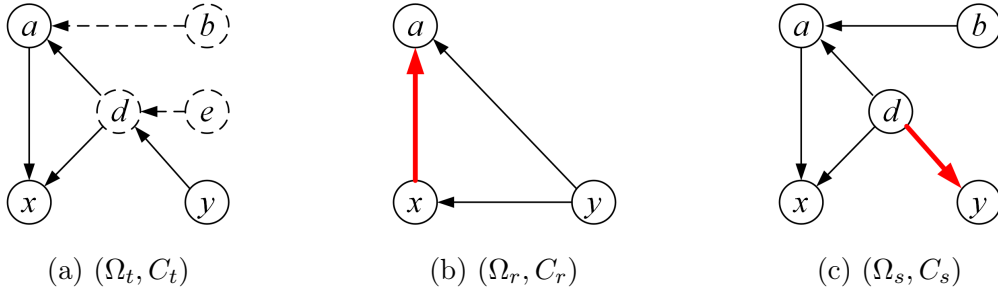


Figure 17: Ruling out a correct link by nitpicking: $x \leftarrow_r y$, $x \leftarrow_t y$.

6.1 Assumptions regarding the receiver.

Receiver is unaware of their unawareness. We assume that our receiver is unaware of their own unawareness and does not even consider existence of variables they do not observe. One could relax this assumption and make the receiver aware of the fact that there may exist other variables they do not observe. This approach encounters a conceptual issue that allowing for confounding variables can explain *any* data. Specifically, any dataset can be explained by assuming the existence of some confounding variable c that affects all other observed variables: $c \rightarrow_r x$ for all $x \in \Omega_r \setminus \{c\}$ (“God’s will” is one example of such c ; various conspiracy theories also rely on the existence of such unobserved—and unobservable—confounding factors). It is unclear then how the receiver should choose between multiple consistent models that rely on confounders.

However, some conclusions can be reached by assuming the receiver allows for the existence of unobserved variables in principle, but applies the Occam’s razor by not relying on them unless absolutely necessary.²¹ Using this approach, the receiver can still recover the set of causal models that are consistent with the data by using the IC^* algorithm (see Pearl, 2009, ch.2) instead of the IC algorithm this paper applies. Using the IC^* algorithm, the receiver is able to find a consistent model for each set of observable variables, as well as discover confounders on their own in some cases. For example, in the situation in Figure 6(b), the receiver would realize that the model produced by the IC algorithm is not consistent with the data, so there must be a confounding variable affecting b and d . The receiver would not observe its distribution, but would become aware of its existence and would even infer the correct causal structure. Further assumptions would be required on how the receiver accepts or rejects models in this scenario. Moreover, the IC^* algorithm is only one of many approaches in the literature on causal discovery, with there being no clear consensus regarding the best way to work with situations in which some relevant variables may be unobservable (Peters et al., 2017, ch. 9.4.1).

Receiver is certain of their model. We assume that the receiver always holds at most one subjective model in mind. If their initial model becomes irreconcilable with the data, the receiver replaces it with another (singular) model. Given that the receiver’s model is very likely to be incorrect—in its incompleteness at the very least—this approach violates the rational expectations assumption that is standard in economic theory. A more standard approach would be to assume the receiver is uncertain regarding which model of the world is correct, assigning some subjective probabilities to different models being true

²¹Gottesman (2025) argues, in a different but somewhat related setting, that it is optimal for the receiver to not include unrelated variables in their model if a sender does not suggest them. This is true even if the receiver makes strategic inferences from the sender’s messages given the sender’s strategic concerns.

and updating these probabilities via Bayes’ rule after observing new information from the sender. “Information” in this case may include both direct evidence shown by the sender and any kind of indirect inference the receiver can make from the sender’s strategic motives and observed behavior. Indeed, this is the approach adopted by the “narrative persuasion” literature (c.f. Schwartzstein and Sunderam, 2021; Aina, 2024; Spano, 2025).

Our approach is closer instead to a model known in epistemology as the AGM paradigm (Peppas and Williams, 1995). The AGM paradigm prescribes that belief changes should follow the principle of minimal change. In line with this principle, *expansion* of a subjective model of the world is the simplest change and should be employed when new data is compatible with the current set of beliefs. In our model, once the receiver has a subjective model, they first try to expand their pre-existing model (Ω_r, C_r) to incorporate new data. However, when new data is incompatible with the pre-existing model, the model is *revised* in some way based on the sophistication of the receiver.

Experimental evidence suggests that our approach is closer than Bayesian updating to the thought process people employ in the real world. In particular, Aina and Schneider (2024) run a lab experiment, in which participants must assign subjective probabilities to different models of the world after observing some evidence that can be interpreted through the lens of these models. They find that most participants assign subjective probability of one to the model with the best fit of the data, while only a small share of the participants’ guesses are consistent with Bayesian updating.

Receiver discards the debunked model. The receiver’s response in our model to evidence that is incompatible with their subjective model depends on their type. A naïve receiver discards their own model and simply accepts the sender’s proposed model (assuming it is consistent with the new data). While a sophisticated receiver is more critical of the sender’s model, they also discard the debunked model completely. An alternative assumption, which would be in line with the principle of minimal revision mentioned above, could be that the receiver only discards the debunked links. They could then attempt to construct a new model that is consistent with the new evidence while remaining as close as possible to their original model. This approach would make persuasion more difficult for the sender. Our results could then be understood as the upper bound on the feasibility of persuasion in the best case for the sender.

On a separate note, our model allows a sophisticated receiver to discard their pre-existing model after it is debunked but reject the sender’s proposed model if it is not supported well enough by the evidence. We do not specify which model the sophisticated receiver adopts in such a situation, as this is not required for our goals. If necessary for other results, it may be plausible to assume that the receiver follows the approach outlined above, retaining a part of their debunked model and attempting to revise it to make it consistent with the data.

6.2 Assumptions regarding the sender and the world.

Perfect data. Our model assumes that both players can perfectly observe the joint distribution of all variables (that they are aware of), and that this distribution is sufficiently nice in the sense of any link in the true model being identifiable in the data. This abstracts from decades of econometric literature and centuries of statistics literature, which together deal with an immense variety of issues that arise in statistical estimation of such joint distributions. This abstraction is intentional, since we would like to explore the issues that are specific to persuasion and selective disclosure. However, all of the issues with statistical estimation still remain in the real world. Incorporating them in our model would lead to

agents having less confidence in subjective models. On the one hand, this may make persuasion more difficult, especially with sophisticated receivers, since the sender would have a harder time debunking pre-existing models with limited and imperfect data, and may struggle to find enough arguments in favor of their proposed model. But on the other hand, the receivers’ pre-existing models would also be based on limited and imperfect data, and as a consequence, they may be easier to debunk in this case than in our benchmark.

Agents cannot do interventions. We consider learning from observational data, where the receiver tries to recover the true causal graph by merely looking at existing data, and the sender can only provide more of the existing data (other variables). Specifically, we assume that neither the sender, nor the receiver can run experiments, directly manipulating some of the variables and thereby generating additional data points. Allowing for such interventions could allow the sender to more easily demonstrate some truthful links, as well as potentially ease the receiver’s causal discovery problem. Allowing the receiver to run experiments could also mitigate the risk of being deceived by the sender. The flip side is that if the whole dataset is formed by the receiver’s choices and the realized consequences, underexperimentation could lead to poor data (see Gratton et al., 2025 for one such example). Overall, we find the issues related to optimal learning and persuasion with interventions to be an appealing avenue for future research.²²

Rich worlds. Some of our results (specifically, Theorem 2) assumed the world is rich—namely that the true graph can be uniquely identified. As mentioned in the discussion following the theorem, this assumption is not necessary. In general, if the world is not rich, then persuading a sophisticated receiver and debunking pre-existing models may be more difficult, since the link of interest is no longer guaranteed to be identifiable. At the same time, conditional on debunking the receiver’s model, deceiving the naïve receiver may be simpler because the receiver may have less evidence in favor of the true model.

Simple worlds. Most of our results are formulated for simple worlds—those where all V-structures are obvious in the sense of not requiring any additional control variables. This assumption is relevant for our results in two ways. One is that more possibility results would be available. In particular, Proposition 2 (and the corresponding version of statements in Theorem 2) would continue to hold as is. However, we could adjust the definitions of obvious and non-obvious causes to take into account the control variables required to identify the relevant V-structures. Allowing for such “non-simple” causes would open up more persuasion avenues, with the caveat that controls would have to be included in the set of the revealed variables, making persuasion more difficult than with “simple” causes.

In turn, the proofs of the impossibility results (Theorem 1 and Corollaries 1 and 2 that follow from it) show that when we run the IC algorithm, it can only mis-orient a link (identify $a \leftarrow b$ when in the true model, $a \Rightarrow_t b$) if either another link has been mis-oriented or the true model is not simple. However, we are not aware of examples in which links are mis-oriented when the true model is non-simple. This raises the possibility that the results may extend to non-simple environments, but a formal treatment is left for future research.

²²For a brief overview of the literature on causal discovery with interventions see Zanga et al. (2022, Section 5).

7 Conclusion

In this paper, we propose the first model of causal persuasion. Our main contribution is setting up a novel tractable model of strategic communication of causal models that explicitly deals with how causation can be established and debunked. We show that unless the receiver is naïve, persuading them of a causal link requires disclosing additional variables. The latter task is also closely related to debunking a pre-existing model that a (naïve or sophisticated) receiver may have.

Our model emphasizes the asymmetry in communicating causal models. We show that it is easier for the sender to persuade the receiver that a particular causal link exists than to persuade that a link does not exist. Persuasion difficulty in this case is measured by the number of causal variables that need to be revealed by the sender to debunk the receiver’s subjective causal model. The reasoning behind this is that larger models with more variables are more difficult for the sender to communicate and for the receiver to understand.

Our model is set up in a way to make persuasion as simple as possible for the sender, while restricting them to disclosure of only truthful data. We show that the scope for deception nonetheless exists in this case, with the sender being able to spin the correlation generated by confounding variables as causation, even if the receiver is sophisticated. Further, the sender may be able to debunk the receiver’s defective (incorrect) model and replace it with another defective model. Conversely, if a sender wants to persuade a sophisticated receiver that no causal link exists between two variables and any correlation between them is spurious, then this is only possible if this is so in the true model. Even then, the absence of a link may be difficult to prove—more difficult than flipping a link.

This paper makes only the first step towards a theory of causal persuasion. The discussion in Section 6 outlines many assumptions that could be relaxed or reinterpreted in future work. These include making the receiver privy to the sender’s strategic motives and making indirect inferences from the sender’s messages, allowing the receiver’s prior belief over models to be non-degenerate, or allowing the receiver to retain a part of their subjective model in the face of conflicting evidence. Additionally, competition between senders with different motives appears to be an interesting direction for future research.

A Appendix

A.1 Proof of Theorem 1

This proof shows that if $x \leftarrow_t y$, then $x \rightarrow_s y$ cannot be uniquely consistent with $P|\Omega_s$, which is required to persuade a sophisticated receiver. Since a link is uniquely consistent with $P|\Omega_s$ if and only if it is oriented by the IC algorithm run on Ω_s , we show that the IC algorithm run on any $\Omega_s \subseteq \Omega_t$ cannot orient $x \rightarrow_s y$. We say that a link between $a, b \in \Omega_s$ is *mis-oriented* if the IC algorithm outputs $a \rightarrow_s b$ when in the true model, $a \leftarrow_t b$.

Direct case. We start with the case $x \leftarrow_t y$ and show that for any $\Omega_s \subseteq \Omega_t$, the IC algorithm cannot orient $x \rightarrow_s y$. The link can be mis-oriented in Step 2 or Step 3 of the IC algorithm. We show below that any such mis-orientation requires one of the following:

1. that at least one link has already been mis-oriented, or
2. that the true model is not simple.

We conclude then that if the true model is simple, there cannot be any “first mistake”,

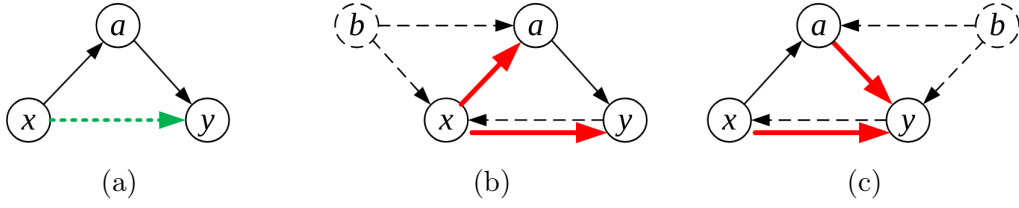


Figure 18: R2 and non-simple models required to orient a link incorrectly.

NOTES: Green dotted arrows in this and following figures represent the link identified by the respective Meek’s rule. As before, bold red arrows represent the identified defective links, and the dashed black links and variables represent the items that are not (made) known to the receiver.

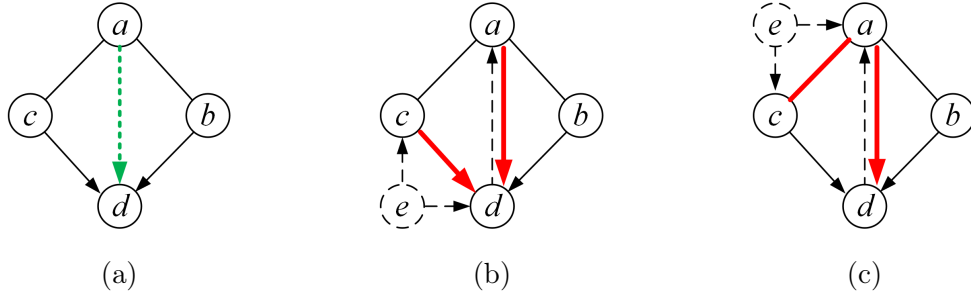


Figure 19: R3 and non-simple models required to orient a link incorrectly.

the first link to be mis-oriented by the IC algorithm, since any such mistake requires that another mistake had already been made.

To mis-orient $x \rightarrow_s y$ in Step 2, the sender’s model needs to contain a direct V-structure $x \rightarrow_s y \leftarrow_s a$ for some a . This means that $a \perp x$ and $(x \not\perp a \mid y)$ in the data, and such that $(a \not\perp y \mid S)$ for any $S \subseteq \Omega_s \setminus \{y, a\}$. The latter fact for $S = \emptyset$ is equivalent to a and y being correlated in the true model. Since $x \leftarrow_t y$, it follows that a and x are then also correlated in the true model, meaning $x \not\perp a$, yielding a contradiction.

Mis-orienting $x \rightarrow_s y$ in Step 3 of the IC algorithm must occur due to application of one of Meek’s rules. Rule R1 cannot produce the “first mistake” (be the first to mis-orient a link). R1 orients $x \rightarrow_s y$ only if $a \rightarrow_s x$ for some a not adjacent to y . But if $a \rightarrow_t x$ then $a \rightarrow_t x \leftarrow_t y$ form a direct V-structure in $P|\Omega_s$, which would have been identified in Step 2. Hence mis-orienting $x \rightarrow_s y$ via R1 requires that $a \not\rightarrow_t x$, meaning $a \rightarrow_s x$ is also defective. The following cases are possible.

1. If $a \leftarrow_t x$ (or $a \leftarrow_t x$) then $a \rightarrow_s x$ must have been mis-oriented.
2. If a, x are not adjacent (or correlated) in the true model, then $a \rightarrow_s x$ being identified by the IC algorithm contradicts Lemma 1.
3. If a, x are correlated but not connected, and there is a confounder $b \in \tilde{C}_t(a, x)$, then b, x, y still form a V-structure in $P|\Omega_s$ which would have been identified in Step 2.

Rule R2 (“orient $x \rightarrow_s y$ if $x \rightarrow_s a$ and $a \rightarrow_s y$ for some a ”, see Figure 18(a)) can obviously not mis-orient a link unless one of the other links is defective, since $x \rightarrow_t a \rightarrow_t y$ together with $y \rightarrow_t x$ violates acyclicity. If either $x \rightarrow_s a$, or $a \rightarrow_s y$ have been mis-oriented, then we are done. If one of these links is due to a hidden confounding variable, then the world cannot be simple. Specifically, if $x \rightarrow_s a$ is due to a hidden confounder b such that $x \leftarrow_t b \rightarrow_t a$, then b, x, y form a V-structure in P conditional on a , see Figure 18(b). Similarly, if $a \rightarrow_s y$ is due to a hidden confounder b such that $a \leftarrow_t b \rightarrow_t y$, then

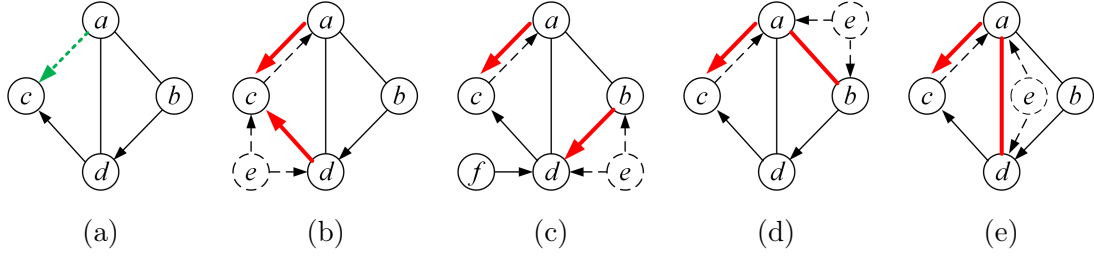


Figure 20: R4 and non-simple models required to orient a link incorrectly.

x, a, b form a V-structure in P conditional on y , see Figure 18(c).

For either of rules R3 and R4 to mis-orient a link, one of the following cases must apply.

1. At least one of two other required link directions have been mis-oriented. In this case, we have traced back to an already existing mis-oriented link.
2. At least one of two other required link directions come from a hidden confounding variable.

(a) Consider R3 in Figure 19(a). If link $c \rightarrow_s d$ is defective due to omitted variable e , then it must have come from the underlying true model shown by black arrows in Figure 19(b). If $a \rightarrow_t c$, then we have a non-simple V-structure a, c, e with control d , hence the world is not simple. Alternatively, if $c \rightarrow_t a$, then note that we must have $b \rightarrow_t a$ (by acyclicity from $b \rightarrow_t d \rightarrow_t a$), so we have a V-structure b, a, c that would have been identified in Step 2.

(b) Consider R4 in Figure 20(a). Then we have two directed links to examine:

- i. If link $d \rightarrow_s c$ is defective due to omitted variable e , then it must have come from the underlying true model shown in Figure 20(b). Yet, if this was the true model, the link should have been identified in the opposite direction (as $c \rightarrow_s d$) since c, d, b form a V-structure conditional on a in the data. So we have a contradiction.
- ii. If link $b \rightarrow_s d$ was incorrectly identified due to omitted variable e , then it must have come from the underlying true model shown in Figure 20(c). However, the true model in Figure 20(c) cannot be simple: either c, a, b form a V-structure conditional on d, e (if $a \leftarrow_t b$) or a, b, e form a V-structure conditional on c, d (if $a \rightarrow_t b$).

3. At least one of two other required blank (unoriented) links in (Ω_s, C_s) are defective due to omitted confounding variables.

(a) Consider R3 in Figure 19(a). If the link $a - c$ is defective due to omitted variable e , then it must have come from the underlying true model shown by black arrows in Figure 19(c). Then the true model is not simple, since e, a, d form a V-structure conditional on c .

(b) Consider R4 in Figure 20(a). Then we have two blank links to consider.

- i. If the blank link $a - b$ is defective due to omitted variable e , then it must have come from the underlying true model shown by black arrows in Figure 20(d). However, then c, a, e form a V-structure conditional on b, d , hence the true model is not simple.

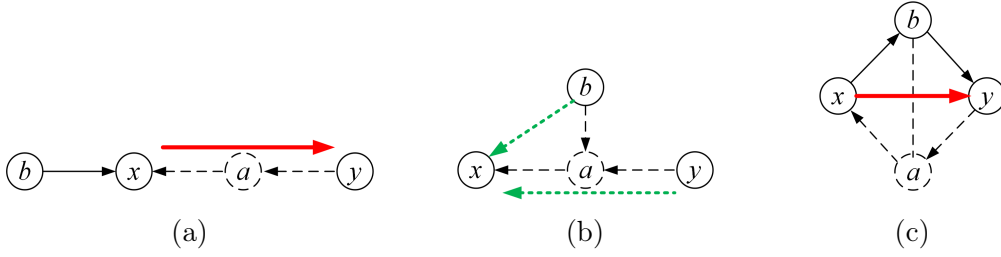


Figure 21: R1, R2, and models required to orient a link incorrectly, indirect case.

- ii. If the blank edge $a - d$ is defective due to omitted variable e , then it must have come from the underlying true model shown by black arrows in Figure 20(e). However, then c, a, e form a V-structure conditional on d , hence the true model is not simple.

We conclude that if $x \leftarrow_t y$, then mis-orienting this link as $x \rightarrow_s y$ requires a non-simple true model. If the true model is simple, then mis-orienting a link is only possible if another link has already been mis-oriented, hence there can be no “first mistake”, and hence no mis-orientation can occur.

Indirect case. Consider then the case when $x \leftarrow_t y$, and x and y are not adjacent in the true model. The argument above implies that if the receiver is made aware of all variables that are (along all paths) between x and y in the true model, then a link can never be mis-oriented, hence $x \Rightarrow_s y$ cannot be uniquely consistent with $P|\Omega_s$. Suppose then that there exists a latent variable a such that $x \leftarrow_t a \leftarrow_t y$ and $a \notin \Omega_s$. We want to show that the IC algorithm can not identify $x \rightarrow_s y$ in this situation. (The case with more than one latent variable in between x and y then follows by induction.)

Note that if x is the only child of a and a is the only parent of x , then we can treat x and a as a single variable for all means and purposes, and so no new situations arise. We ignore this case in what follows. We can show that the link $x \rightarrow_s y$ cannot be identified from a direct V-structure in Step 2 of the IC algorithm using the same argument as in the direct case above. We now show that the link $x \rightarrow_s y$ cannot be identified in Step 3 of the IC algorithm. In doing so, we only need to consider the cases not covered in the direct case above.

Suppose rule R1 orients $x \rightarrow_s y$ from some $b \rightarrow_s x$ where b is not adjacent to y . The logic from the direct case above shows that this cannot happen when b is adjacent to x in the true model (regardless of whether it is adjacent to a), see Figure 21(a) for illustration. Suppose then instead that $x \leftarrow_t a \leftarrow_t b, y$. But then b, x, y form a direct V-structure in the data $P|\Omega_s$, as shown in Figure 21(b), which would identify $x \leftarrow_s y$, yielding a contradiction. Hence, again it must be that the link $b \rightarrow_s x$ is defective.

Suppose rule R2 orients $x \rightarrow_s y$ from some oriented links $x \rightarrow_s b \rightarrow_s y$. However, it cannot be that $x \rightarrow_t b \rightarrow_t y$, since together with $x \leftarrow_t y$ this violates acyclicity, see Figure 21(c). It follows that either $x \rightarrow_s b$, or $b \rightarrow_s y$ is defective.

Suppose rule R3 orients $x \rightarrow_s y$ from some oriented links $z \rightarrow_s y \leftarrow_s b$ for some z and b adjacent to x and not to each other in the sender’s model, see Figure 22(a) for illustration. If $z \rightarrow_t y \leftarrow_t b$, then acyclicity implies $z \rightarrow_t x$ and $b \rightarrow_t x$, see Figure 22(b). Then z, x, b is a V-structure that should have been identified in Step 2 of the IC algorithm – a contradiction, since $x - z$ and $z - b$ must not form a V-structure for R3 to apply. Therefore, either $z \rightarrow_s y$ or $y \leftarrow_s b$ is defective.

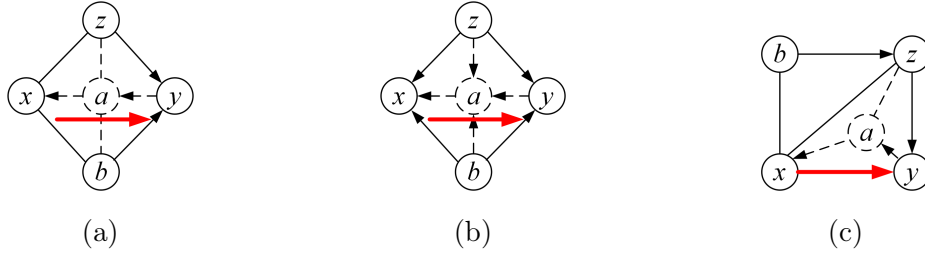


Figure 22: R3, R4, and models required to orient a link incorrectly, indirect case.

Suppose rule R4 orients $x \rightarrow_s y$ from some oriented $b \rightarrow_s z \rightarrow_s y$ for some b and z adjacent to x , see Figure 22(c) for illustration. If $b \rightarrow_t z \rightarrow_t y$, then acyclicity implies $b \rightarrow_t x$. However, then a, x, b form a V-structure conditional on z , contradicting simplicity. Therefore, either $b \rightarrow_s z$, or $z \rightarrow_s y$ is defective.

In all of the cases above, the respective defective link either has been mis-oriented, or has been mis-identified due to a hidden confounder. In the latter case, we can apply the same logic as in the direct case to either arrive at a contradiction (with the model being simple or with the link being unoriented after Step 2), or to find a mis-oriented link. In either case, we again conclude that if the true model is simple, then mis-orienting a link is only possible if another link has already been mis-oriented, hence there can be no “first mistake”, and hence no mis-orientation can occur. This completes the proof.

A.2 Proof of Theorem 2

We begin by establishing a supplementary Lemma.

Lemma 2. *Suppose the true model (Ω_t, C_t) is simple and rich, there exist $x, y \in \Omega_t$ such that $x \Rightarrow_t y$, there exists no obvious cause of y given x , and $\tilde{C}_t(x, y)$ is empty. Then there must exist a direct V-structure a, b, c “upstream from x ”, meaning $a, c \in \bar{C}_t(x)$ and $b \in \bar{C}_t(x) \cup x$.*

Proof. If (Ω_t, C_t) is rich and $x \Rightarrow_t y$, then all links along the path from x to y must be oriented in either Step 2 or Step 3 of the IC algorithm when it is run on the set of all variables, Ω_t . Note that R2 does not establish new connections (meaning if a link $d \rightarrow_t e$ is oriented by R2, then some other link $d \rightarrow_t f$ along the path $d \Rightarrow_t e$ must have already been oriented by some other criterion), so we can ignore it. Further, R3 never applies in simple models (since the structure in R3 contains a V-structure with controls). Hence, any link in a simple rich model must be oriented in either Step 2 from a direct V-structure, or in Step 3 by either R1 or R4.

Consider a path from x to y and take some variable $d \in \Omega_t$ such that $x \rightarrow_t d \Rightarrow_t y$ (possibly with $d = y$ if x and y are adjacent). Focus on the link $x \rightarrow_t d$. Proceed according to cases outlined above.

1. Suppose link $x \rightarrow_t d$ is oriented in Step 2 of the IC algorithm. Then it must be a part of a direct V-structure x, d, e for some $e \in \Omega_t$. But then $x \perp e$ and $x, e \Rightarrow_t y$, meaning x, y, e constitute a (possibly indirect) V-structure. Therefore, e is an obvious cause of y given x , which contradicts the premise of the lemma.
2. Suppose link $x \rightarrow_t d$ is oriented in Step 3 by Meek’s rule R1. Rule R1 goes up the causal tree: it can only identify the link $x \rightarrow_t d$ if x is a child of another node that is not correlated with d conditional on x . Hence, there must exist $g \in \Omega_t$ not adjacent

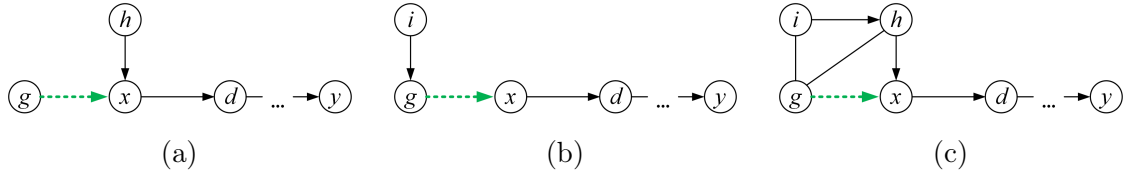


Figure 23: Options for orienting link $g \rightarrow_t x$.

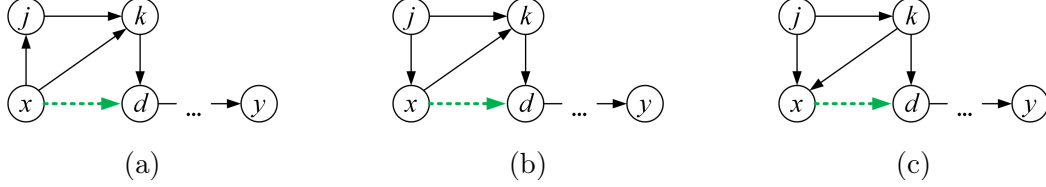


Figure 24: Options for orienting link $x \rightarrow_t d$ via R4.

to d s.t. $g \rightarrow_t x$, and this link was oriented earlier in the algorithm. Consider again cases regarding how $g \rightarrow_t x$ could have been oriented:

- (a) If $g \rightarrow_t x$ was oriented in Step 2 of the IC algorithm as a part of some direct V-structure g, x, h then this is exactly the V-structure required by the lemma, so we are done. This case is depicted in Figure 23(a).
- (b) If $g \rightarrow_t x$ was oriented in Step 3 by rule R1, then there must exist another link $i \rightarrow_t g$ that was oriented earlier in the algorithm. This case is depicted in Figure 23(b).
- (c) If $g \rightarrow_t x$ was oriented in Step 3 by rule R4, then there must exist another link $h \rightarrow_t x$ that was oriented earlier in the algorithm. This case is depicted in Figure 23(c).

The cases above imply that orienting a link upstream from x requires either a direct V-structure upstream from x , or another link that has already been oriented. In the latter case, we can apply the logic again to the other link that was already oriented and obtain the same conclusion. Proceeding by induction, we will inevitably (since the graph is finite) arrive at some V-structure upstream from x , fulfilling the requirement of the lemma.

3. Suppose link $x \rightarrow_t d$ is oriented in Step 3 by Meek's rule R4. Then there exist $j, k \in \Omega_t$ adjacent to x such that $j \rightarrow_t k \rightarrow_t d$, with these latter links having already been oriented by the IC algorithm. Focus on the link $j \rightarrow_t k$ and consider different cases for how it could have been oriented and how x, j, k are connected (see Figure 24):

- (a) If $x \leftarrow_t k$, as in Figure 24(c), then k is a confounder for x, y , which violates the premise of the lemma requiring that $\tilde{C}_t(x, y)$ is empty.
- (b) Link $j \rightarrow_t k$ was oriented in Step 2 of the IC algorithm as a part of a direct V-structure j, k, l for some $l \in \Omega_t$. If $x \rightarrow_t k$ (we are in cases (a) or (b) of Figure 24), then $x \perp l$, so l is an obvious cause of y given x —a contradiction to the premise of the lemma.
- (c) Link $j \rightarrow_t k$ was oriented in Step 3 by Meek's rule R1. Then there exists $m \in \Omega_t$ s.t. link $m \rightarrow_t j$ was oriented earlier in the algorithm. If $x \rightarrow_t j$, then such

m is an obvious cause of y given x , which is a contradiction. If $x \leftarrow_t j$, then $m \rightarrow_t j$ is a link upstream from x , so from case 2 above we know there must exist a direct V-structure upstream from x .

- (d) Link $j \rightarrow_t k$ was oriented in Step 3 by Meek's rule R4. Then there exist $n, o \in \Omega_t$ adjacent to j such that $n \rightarrow_t o \rightarrow_t k$, with these latter links having already been oriented by the IC algorithm. We can then repeat the analysis in case 3, focusing on link $n \rightarrow_t o$. Proceeding by induction, we will either find a V-structure upstream from x , or arrive at a contradiction.

This concludes the proof of Lemma 2. \square

We now proceed to the proof of Theorem 2.

The receiver's model has defective link $x, y \in \Omega_r$ such that $x \leftarrow_r y$ but $x \not\leftarrow_t y$. By Lemma 1, x is correlated with y in C_r if and only if they are correlated in C_t . Proceed according to the cases from the theorem.

Case 1: there exists an obvious cause z of y given x : $z \Rightarrow_t y$ and $z \perp x$. Then revealing z creates a V-structure because $x \Rightarrow_s y \Leftarrow_s z$ (so $z \not\perp x \mid y$), hence the receiver must conclude that $x \Rightarrow_s y$, debunking (Ω_r, C_r) . In what follows, we present this argument in detail.

By assumption, x and y are adjacent in (Ω_r, C_r) . Consider $\Omega_s \equiv \Omega_r \cup z$. Then x and y are still adjacent in any consistent (Ω_s, C_s) , since $(x \not\perp y \mid S \cup z)$ for any $S \subset \Omega_r$. If y and z are adjacent in (Ω_s, C_s) , then x, y, z constitute a direct V-structure in $P|\Omega_s$ because $x \Rightarrow_t y \Leftarrow_t z$ (so $z \not\perp x \mid y$). The receiver must identify this V-structure in Step 2 of the IC algorithm and conclude that $x \rightarrow_s y$, which debunks (Ω_r, C_r) . If y and z are not adjacent in (Ω_s, C_s) , but there exists a path $z \Rightarrow_s y$ s.t. all variables $w \in \Omega_s$ along this path (i.e., such that $z \Rightarrow_s w \Rightarrow_s y$) are also obvious causes of y given x , then we arrive at a contradiction: for some such $w \in \Omega_r$, variables x, y, w form a direct V-structure in $P|\Omega_r$, hence the receiver must have concluded that $x \rightarrow_r y$ in Step 2 of the IC algorithm.

Therefore, it remains to consider the case when y and z are not adjacent in (Ω_s, C_s) , and every path $z \Rightarrow_s y$ contains some $w \in \Omega_s$ such that $z \Rightarrow_s w \Rightarrow_s y$ and $w \not\perp x$. Fix one such path and take such w which is closest to z in (Ω_s, C_s) along that path. It is without loss to assume that z and w are adjacent in (Ω_s, C_s) . Then w must be adjacent to x in (Ω_r, C_r) and (Ω_s, C_s) , since otherwise—if there existed $S \subset \Omega_r$ such that $(x \perp w \mid S)$ — x, y, w would form a V-structure in $P|\Omega_r$ conditional on S , hence the receiver would have concluded that $x \rightarrow_r y$ in Step 2 of the IC algorithm, which contradicts the assumption that $x \leftarrow_r y$.

Further, it must be that $w \not\Rightarrow_t x$, since otherwise $z \Rightarrow_t x$, so $z \not\perp x$, meaning z would not be an obvious cause. Then $x \perp z$ by assumption and $(x \not\perp z \mid w)$ by the above, hence x, w, z is a direct V-structure in $P|\Omega_s$. The receiver must then orient the links $x \rightarrow_s w \leftarrow_s z$ in Step 2 of the IC algorithm. Since $w \Rightarrow_t y$, Meek's rule R1 then orients all links along at least one path $w \Rightarrow_s y$. Finally, from $x \rightarrow_s w \Rightarrow_s y$ R2 orients $x \rightarrow_s y$, debunking link $x \leftarrow_r y$ and the receiver's model (Ω_r, C_r) .

Case 2: $x \Rightarrow_t y$ and there is a non-obvious cause w of y given x . If there also exists an obvious cause z of y given x , then case 1 above applies and proves the statement, hence for the remainder of this proof assume there is no such obvious cause z . If $\tilde{C}_t(x, y)$ is non-empty, the sender can disclose w and any $c \in \tilde{C}_t(x, y) \cap C_t(x)$. Note that w must be not correlated with c , since otherwise x would not be d-separating w and y , which contradicts

the definition of w . Therefore, disclosing w and c creates a direct V-structure w, x, c that is identified by the receiver in Step 2 of the IC algorithm. Link $x \rightarrow_s y$ is then oriented in Step 3 of the IC algorithm by Meek’s rule R1, debunking link $x \leftarrow_r y$ and the receiver’s model (Ω_r, C_r) .

If $\tilde{C}_t(x, y) = \emptyset$ and the true model is rich, then by Lemma 2, there exists a direct V-structure $a \rightarrow_t b \leftarrow_t c$ with $b \in \tilde{C}_t(x) \cup x$. Suppose then that the sender reveals its V-parents: $\Omega_s \equiv \Omega_r \cup \{a, c\}$. If x is adjacent to a, c in (Ω_s, C_s) , then the receiver must identify a direct V-structure $a \rightarrow_s x \leftarrow_s c$ in Step 2 of the IC algorithm because: (i) $a \perp c$, but (ii) $(a \not\perp c \mid b)$ implies $(a \not\perp c \mid x)$. Then Meek’s rule R1 orients link $x \rightarrow_s y$ because $(a \perp y \mid x)$, debunking (Ω_r, C_r) . Alternatively, if x is not adjacent to a, c in (Ω_s, C_s) , then there exists $d \in \Omega_r$ that is adjacent to a, c and such that $a, c \rightarrow_t d \Rightarrow_t x$. Then a, d, c form a direct V-structure in (Ω_s, C_s) , which must be identified by the receiver in Step 2 of the IC algorithm. Meek’s rule R1 then orients all links along at least one path $d \Rightarrow_s x$ and, eventually, the link $x \rightarrow_s y$, debunking (Ω_r, C_r) . This concludes the proof of Theorem 2.

References

- Aina, C. (2024). Tailored Stories. Working paper.
- Aina, C. and Schneider, F. H. (2024). Weighting Competing Models. Working paper.
- Al-Najjar, N., Pomatto, L., and Sandroni, A. (2014). Claim Validation. *American Economic Review*, 104(11):3725–36.
- Ambuehl, S., Bhui, R., and Thysen, H. C. (2026). Mental models of causal structure in economics and psychology.
- Bergemann, D. and Morris, S. (2019). Information Design: A unified Perspective. *Journal of Economic Literature*, 57(1):44–95.
- Card, D. (1999). The Causal Effect of Education on Earnings. *Handbook of Labor Economics*, 3:1801–1863.
- Damsbo-Svendsen, S. and Hansen, K. M. (2023). When the election rains out and how bad weather excludes marginal voters from turning out. *Electoral Studies*, 81:102573.
- Di Tillio, A., Ottaviani, M., and Sørensen, P. N. (2021). Strategic Sample Selection. *Econometrica*, 89(2):911–953.
- Dranove, D. and Jin, G. Z. (2010). Quality Disclosure and Certification: Theory and Practice. *Journal of Economic Literature*, 48(4):935–963.
- Eliaz, K., Galperti, S., and Spiegler, R. (2025). False narratives and political mobilization. *Journal of the European Economic Association*, 23(3):983–1027. Publisher: Oxford University Press.
- Eliaz, K. and Rubinstein, A. (2025). Wasonian Persuasion. Working paper.
- Eliaz, K. and Spiegler, R. (2020). A Model of Competing Narratives. *American Economic Review*, 110(12):3786–3816.
- Eliaz, K., Spiegler, R., and Weiss, Y. (2021). Cheating with Models. *American Economic Review: Insights*, 3(4):417–434.

- Gottesman, A. (2025). A limitation of Evidence-Backed Communication. Working paper.
- Gratton, G., Lee, B. E., and Yousaf, H. (2025). Bad Democracy Traps. *The Economic Journal*, page ueaf111.
- Grossman, S. J. and Hart, O. D. (1980). Disclosure laws and takeover bids. *The Journal of Finance*, 35(2):323–334.
- Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. (2021). A survey of Learning Causality with Data: Problems and Methods. *ACM Computing Surveys*, 53(4):1–37.
- Hume, D. (1748). *An Enquiry Concerning Human Understanding*. London: Millar.
- Ispano, A. (2025). The perils of a coherent narrative. *Economic Theory*.
- Johnson, D. S. (1974). Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9(3):256–278.
- Kamenica, E. (2019). Bayesian Persuasion and Information Design. *Annual Review of Economics*, 11(1):249–272.
- Kamenica, E. and Gentzkow, M. (2011). Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615.
- Karp, R. M. (2009). Reducibility among combinatorial problems. In *50 Years of Integer Programming 1958-2008: from the Early Years to the State-of-the-Art*, pages 219–241. Springer Berlin Heidelberg.
- Kaur, S., Mullainathan, S., Oh, S., and Schilbach, F. (2025). Do Financial Concerns Make Workers Less Productive? *The Quarterly Journal of Economics*, 140(1):635–689.
- Little, A. T. (2023). Bayesian Explanations for Persuasion. *Journal of Theoretical Politics*, forthcoming.
- Lovász, L. (1975). On the ratio of optimal integral and fractional covers. *Discrete Mathematics*, 13(4):383–390.
- Marie, O. and Pinotti, P. (2024). Immigration and Crime: An International Perspective. *Journal of Economic Perspectives*, 38(1).
- Meek, C. (1995). Causal Inference and Causal Explanation with Background Knowledge. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–410.
- Mehndiratta, M. M., Mehndiratta, P., and Pande, R. (2014). Poliomyelitis: Historical Facts, Epidemiology, and Current Challenges in Eradication. *Neurohospitalist*, 4(4):223–229.
- Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, pages 380–391.
- Nemhauser, G. L. and Wolsey, L. A. (1988). *Integer and Combinatorial Optimization*. Wiley, New York.
- Nogueira, A. R., Pugnana, A., Ruggieri, S., Pedreschi, D., and Gama, J. (2022). Methods and tools for causal discovery and causal inference. *WIREs Data Mining and Knowledge Discovery*, 12(2):e1449.

- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.
- Pearl, J. (2009). *Causality*. Cambridge university press.
- Peppas, P. and Williams, M.-A. (1995). Constructive Modelings for Theory Change. *Notre Dame Journal of Formal Logic*, 36(1).
- Peters, J., Janzing, D., and Scholkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT press.
- Robertson, R. E., Green, J., Ruck, D. J., Ognyanova, K., Wilson, C., and Lazer, D. (2023). Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature*, 618(7964):342–348.
- Schwartzstein, J. and Sunderam, A. (2021). Using Models to Persuade. *American Economic Review*, 111(1):276–323.
- Spiegler, R. (2020). Can Agents with Causal Misperceptions be Systematically Fooled? *Journal of the European Economic Association*, 18(2):583–617.
- Tamborini, C. R., Kim, C., and Sakamoto, A. (2015). Education and Lifetime Earnings in the United States. *Demography*, 52(4):1383–1407.
- Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–270.
- Verma, T. and Pearl, J. (2022). Equivalence and Synthesis of Causal Models. In Geffner, H., Dechter, R., and Halpern, J. Y., editors, *Probabilistic and Causal Inference*, pages 221–236. ACM, New York, NY, USA, 1 edition.
- Williams, G. (2013). *Paralysed with Fear: The Story of Polio*. Palgrave Macmillan UK, London.
- Zanga, A., Ozkirimli, E., and Stella, F. (2022). A survey on Causal Discovery: Theory and Practice. *International Journal of Approximate Reasoning*, 151:101–129. arXiv:2305.10032 [cs].