# Causal Persuasion*

Anastasia Burkovskaya[1] and Egor Starkov[†2]

[1]University of Sydney, School of Economics
[2]University of Copenhagen, Department of Economics

January 2, 2026

**Abstract**

We propose a model of causal persuasion, where a receiver is aware of a subset of real-world variables and their distribution, and tries to infer causal relationships. A better-informed sender knows the full causal graph and selectively reveals additional variables to persuade the receiver about some causal relationship. We show that persuasion is only possible when the receiver's model conflicts with the true causal structure. To reveal a true causal link, the sender often needs to disclose just one or two well-chosen variables. But to dispel a perceived link—to convince the receiver there is no causal relationship—every common cause must be disclosed. Our results highlight a fundamental asymmetry in persuasion: It is much easier to prove a connection than to disprove one.

**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 the examples above 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 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 grateful to Murali Agastya, Chiara Aina, Jean-Michel Benkert, Andreas Bjerre-Nielsen, Martin Dufwenberg, Andrew Ellis, Alessandro Ispano, Klaus Kultti as well as seminar and conference audiences at Copenhagen, Sydney, NET, EWMES 2025 for the many valuable comments.

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

[1]`https://www.bloomberg.com/news/articles/2025-07-23/why-japan-s-bond-market-has-investors-worried`, retrieved Aug 08, 2025.

[2]`https://truthsocial.com/@realDonaldTrump/posts/114617666432673584`, retrieved Aug 08, 2025.

[3]`https://www.youtube.com/watch?v=OTPJYZLD6L8`, retrieved Aug 08, 2025.

[4]Kaur et al. (2025)

[5]`https://hbr.org/2021/11/leaders-stop-confusing-correlation-with-causation`, retrieved Aug 08, 2025.

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 initially has knowledge of some real-world variables and a subjective causal model that links them. The sender can disclose additional variables to 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 persuading the receiver of a certain causal connection is often "easy", while exposing a false cause can be "difficult". Finally, we show that the sender may be able to instill an incorrect model even if they are not able to lie about the true distribution.

Specifically, we assume that all agents can observe data generated by some true causal model, captured by a directed acyclic graph (DAG). The sender observes the true graph and all data, while the receiver is only aware of a subset of all variables and observes the data pertaining to them, rationalizing it with some subjective causal model. The sender can disclose additional true variables, which makes the receiver aware of them and informed of the data related to them. The sender further offers some causal model that is consistent with the data. This new model is only accepted by the receiver if their old model is refuted by the new data.

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. We present examples of such incorrect models that are consistent with the data in Section 4.

We then focus on cases where the receiver's model can be debunked. We show that in many such cases, debunking requires disclosing only one or two additional variables. If the sender aims to convince the receiver of a true causal link, it is sufficient to data mine for two independent causes of either variable. Similarly, if the sender seeks to reverse a causal link that is actually spurious—arising from an omitted variable—persuasion is possible by revealing a single independent cause of the targeted "outcome variable."

By contrast, persuading the receiver that no causal link exists requires identifying and revealing all common causes of the two variables—that is, accounting for every source of their correlation. 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 be often harder than persuading the receiver that the link runs in the opposite direction—despite no such link existing in the first place!

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. These tools made their way into economic theory through works of Spiegler (2016, 2020); Eliaz et al. (2024), and others. To our best knowledge, ours is the first paper to explicitly incorporate causal DAGs into a model of persuasion.

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 or ex ante distributions, respectively. 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.

Contemporary work by Eliaz and Rubinstein (2025) explores a related model of "Wasonian persuasion", where a sender also chooses which causal model 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.

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 withold data but cannot produce fake data (Grossman and Hart, 1980; Milgrom, 1981; see Dranove and Jin, 2010 for an overview), although in our setting the sender can offer an incorrect model to go with factually correct data. 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 survive.

We next present an illustrative example in Section 2. The formal model is set up in Section 3, and Section 4 contains some preliminary analysis. The main results are contained in Sections 5 and 6. Section 7 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 might instead think 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 $t$. The employee, however, is unaware of these factors $a, s, t$, and their subjective model of the world is as in Figure 1(b). This model is supported by the data

[6]Source: `https://poetsandquants.com/2021/05/05/lifetime-earnings-of-mbas/`, retrieved Nov 21, 2025.

[7]For 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).

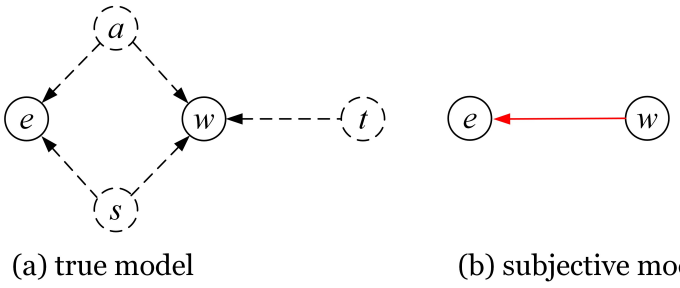



Figure 1: True DAG and employee's initial subjective model.

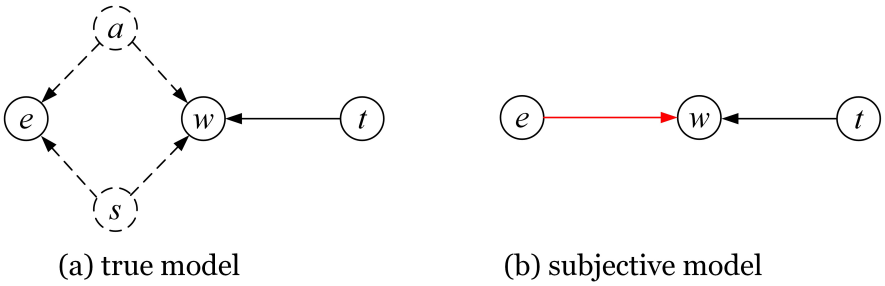
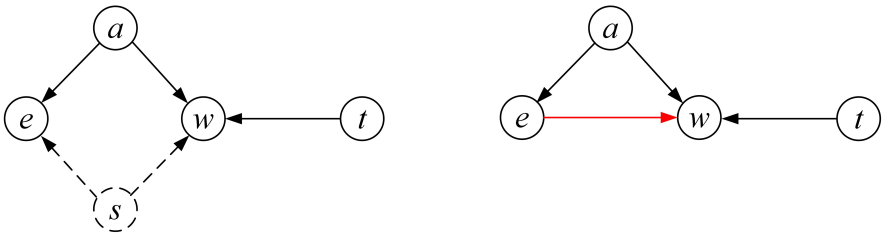



Figure 2: Employee's model after the promo campaign.

generated by the true causal model, in which $e$ and $w$ indeed end up being correlated. Can the business school persuade the employee that the link runs in the opposite direction by disclosing carefully chosen variables? If so, can the employer then persuade the employee that no actual link exists between $e$ and $w$?

Suppose the business school builds its 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 2(b). Indeed, the conditional correlation suggests that there must be *some* connection between them, but if $e \to w \to t$ or $e \leftarrow w \leftarrow t$ or $e \leftarrow w \to t$ (or $e \to t$ 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 link $e \to w$. The promotional campaign successfully persuaded skeptics like our employee by disclosing an additional variable, which rendered the employee's old subjective model inconsistent with the data. Notably, persuasion was effective even though the suggested narrative was not truthful.

The employer, on the other hand, knows that the degree does not actually increase the employee's value. When an employee asks the employer to fund their education, the employer seeks to make this lack of causal connection clear. 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




(a) true model (b) subjective model

Figure 3: Employee's model after the first chat with the employer.

conditional on $a$. The employee's subjective model after this conversation is presented in Figure 3(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 $s$. So long as there is even a single latent 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.

## 3 Model

**The World.** The world is described by a causal directed acyclic graph (DAG) $(\Omega_t, C_t)$ called *the true model*, where $\Omega_t$ is a set of *variables*, and $C_t : \Omega_t \to 2^{\Omega_t}$ 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 $b \to_t a$ and described as "$b$ *has a causal effect* on $a$" and "$b$ is a parent of $a$". 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 $b \to_t ... \to_t a$ 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 $\Omega_t$: $P(\Omega_t) \equiv \prod_{a\in\Omega_t} F_i(a|C_t(a))$. For any subset of variables $\Omega \subset \Omega_t$, let $P|\Omega$ denote the the marginal distribution of $\Omega$. In what follows, we refer to the full distribution $P$ and all marginals $P|\Omega$ as the *data*. Further, given some $\Omega \subset \Omega_t$ and $a, b \in \Omega_t \backslash \Omega$, we use the notation $(a \perp b \mid \Omega)$ if $a$ and $b$ are statistically independent conditional on $\Omega$ given $P$, and $(a \not\perp b \mid \Omega)$ if the converse is true.

**Subjective models and consistency.** We define a *(subjective) model* of the world as a DAG $(\Omega, C)$ with $\Omega \subseteq \Omega_t$. We say that $(\Omega, C)$ 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 $(a \perp b \mid S)$.

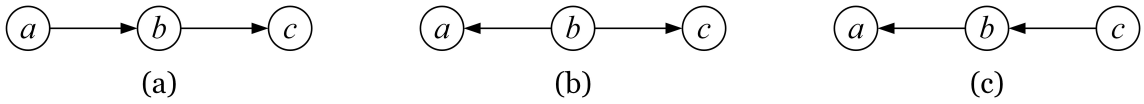



Figure 4: Example: Multiple models consistent with the same data.

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 $b$ 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 adjacent in the model. Minimality is also known as faithfulness or d-faithfulness (Nogueira et al., 2022). We assume that the true model is consistent with the data, meaning the data satisfies minimality.

Note that there may be multiple consistent models $(\Omega, C)$ for given $\Omega$ and $P|\Omega$. For example, consider different models in Figure 4: 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 $(\Omega, C)$ that is consistent with the data generated by the true model $(\Omega_t, C_t)$.[8]

Given a model $(\Omega, C)$, we say $a, b \in \Omega$ are *adjacent* if $a \to b$ or $b \to a$, are *connected* if $a \Rightarrow b$ or $b \Rightarrow a$. We say that $a, b$ are *correlated* either if they are connected, or there exists a common ancestor $c \in \bar{C}(a) \cap \bar{C}(b)$.

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

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 $\Omega_r$ but not other variables. The receiver further has some subjective model of the world $(\Omega_r, C_r)$ that is consistent with $P|\Omega_r$.

The sender knows the true model $(\Omega_t, C_t)$. The sender can *offer* a model $(\Omega_s, C_s)$ to the receiver. The sender can offer any model such that $\Omega_r \subseteq \Omega_s \subseteq \Omega_t$ (new variables can be disclosed, but the sender cannot make the receiver forget), and $(\Omega_s, C_s)$ is consistent with $P|\Omega_s$. The receiver then accepts the sender's model if and only if the receiver's model $(\Omega_r, C_r)$ is not consistent with $P|\Omega_s$. In this case we say that the sender *debunks* $(\Omega_r, C_r)$. Otherwise—if $(\Omega_s, C_s)$ does not debunk $(\Omega_r, C_r)$—we say that the receiver rejects the sender's model. In either scenario, the game ends.

We assume that the sender's objective is to *persuade* the receiver (to accept a model $(\Omega_s, C_s)$) 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). 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 assume that the sender's secondary objective (conditional on persuading the receiver) is to persuade using the least amount of additional variables, i.e., to minimize $|\Omega_s| - |\Omega_r|$.[9]

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 7. Our

[8] An example of such a situation is presented in Figure 7 and discussed further below.

[9] We interpret this objective as aversion to model complexity. Both the sender and the receiver may face difficulties when trying to process complex models with a large number of variables. Hence, we assume that the sender is looking for the simplest convincing model. Alternatively, one could justify this objective via preference for secrecy, where the sender is not willing to disclose any more variables than absolutely necessary.

problem is then better seen as the sender's decision problem. Further, 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 $(\Omega_r, C_r)$, i.e., $\Omega_r \subset \Omega_s$. We say that $(\Omega_r, C_r)$ is consistent with $P|\Omega_s$ if there exists a consistent model $(\Omega_s, C_s)$ such that for all $a, b \in \Omega_r$, if $a \to_r b$ then $a \Rightarrow_s b$.[10] For example, take our illustrative example in Section 2. The receiver's subjective model in Figure 1(b) was not consistent with the data provided in Figure 2. However, if instead the sender disclosed to the receiver only variable $a$ (instead of $t$), then the original subjective model would still be consistent with the new data.

# 4 Preliminary analysis

## 4.1 Consistency

In this section, we discuss in more details 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 $a - b - c$ (where $b$ is adjacent to $a$ and $c$, but $a$ and $c$ are not adjacent), a *collider* $a \to_t b \leftarrow_t c$ looks differently in the data to *chains* $a \to_t b \to_t c$ and $a \leftarrow_t b \leftarrow_t c$ or a *fork* $a \leftarrow_t b \to_t c$. 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 $(a \perp c \mid b)$. Verma and Pearl (1990, 2022) show that any pair of models are observationally equivalent (consistent with the same data) if they have the same sets of links and colliders. Conversely, a model that has different links or different colliders relative to the true model would contradict the data. We refer to collider patterns in the data as V-structures, defined below.

**Definition 1.** *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 $S$ as the set of* controls*.*

**Definition 2.** *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 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 $a \to b \leftarrow c$ 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 3** (IC algorithm)**.** *Consider set $\Omega$ 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 $(a \perp b \mid S)$.*
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:*

[10] When talking about multiple models $(\Omega, C_t)$, $(\Omega, C_r)$, we use subscripts to indicate which model the arrow notation refers to: "$a \to_r b$" means $a \in C_r(b)$ and "$a \to_s b$" means $a \in C_s(b)$.

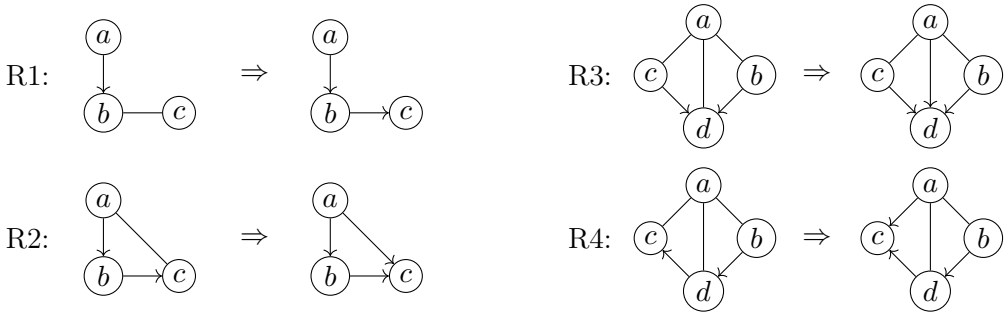



Figure 5: Meek (1995) rules.

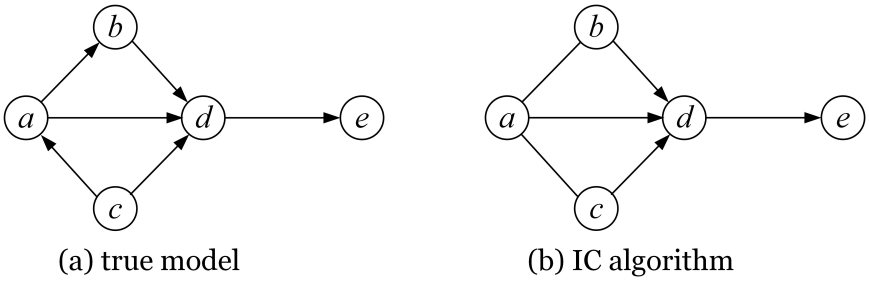



Figure 6: Example of IC algorithm.

*(a) link $b \to a$ creates a V-structure that would have been identified in Step 2, or*

*(b) link $b \to a$ 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 does not contain any additional V-structures relative to $C$ produces a consistent model, i.e., any such $(\Omega, C')$ is consistent. If a consistent model does not exist, the IC algorithm will produce an inconsistent model, as we discuss below.

We further use the results of Meek (1995), who argues that Step 3 of the IC algorithm is equivalent to iterative application of four orientation rules depicted in Figure 5. Hereinafter, we refer to to them as "Meek rules" and to the individual rules as R1–R4.

**Example 1.** *Suppose the true model is as in Figure 6(a), and we use the IC algorithm on data generated by it. In Step 1, the algorithm identifies the pairs of variables that are never conditionally independent. So it will connect with undirected links the pairs ab, ac, ad, bd, cd and de. In Step 2, it identifies the only V-structure $c \to d \leftarrow b$ 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 $b \to d - e$ and direct $d \to e$; and (2) we apply R3 to the rectangular abdc and direct the link $a \to d$. The resulting partially directed graph is shown in Figure 6(b).*

## 4.2 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 hold a model that is consistent with the data, yet incorrect. We call such models defective and offer examples below.

**Definition 4.** *Given a model $(\Omega_r, C_r)$, we call $a \to_r b$ for some $a, b \in \Omega_r$ a* defective link *if $a \not\Rightarrow_t b$. The receiver's model $(\Omega_r, C_r)$ is* defective *if it has a defective link.*

**Example 2.** *Suppose the true model is $a \to b$, and the receiver is aware of both variables. The IC algorithm identifies 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, the receiver may believe that $a \leftarrow b$, which is consistent with the data, but such a link is defective.*

**Example 3.** *Suppose the true model is as in Figure 4(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 4(b) and 4(c) are consistent yet defective. However, note that model $a \to b \leftarrow c$ is inconsistent with the data, since $a \not\perp c$, so there is no V-structure in the data.*

**Example 4.** *Suppose the true model is as shown in Figure 6(a). The IC algorithm requires the directed links presented in Figure 6(b) but imposes no restrictions on the directions of links $b - a - c$. Hence, a model with $b \to a \to c$ would still be consistent, despite both links being defective.*

In addition to the inconclusiveness of the IC algorithm, we also assume that the receiver is, initially, only aware of a subset $\Omega_r \subset \Omega_t$ of all relevant variables. Does this variable omission create more potential for causal misperceptions? I.e., what kind of causal connections *can* the receiver get wrong due to not observing all of $\Omega_t$ (omitted variable bias)? These questions are partially answered by the following lemma.

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

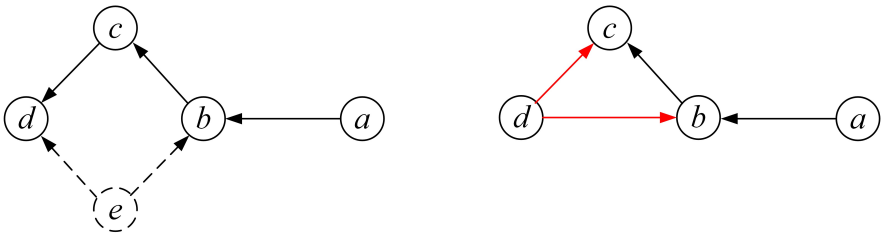
1. *$a$ and $b$ are correlated in $C_r$ if and only if they are correlated in $C_t$;*
2. *if $a$ and $b$ are not adjacent in $C_r$, then they are not adjacent in $C_t$.*

*Proof.* The first statement: suppose there exist $a, b \in \Omega_r$ that are correlated in $C_t$ but not in $C_r$. Then $(a \not\perp b \mid S)$ for any $S \subset \Omega_t$ in the data generated by $C_t$. But according to the IC algorithm (which we can use to check that $C_r$ is consistent with $P|\Omega_r$), it must be that $(a \perp b \mid S)$ for some $S \subset \Omega_r$, which is a contradiction. The other case is analogous.

The second statement: since $C_r$ is consistent, it must align with the output of the IC algorithm. Step 1 of the algorithm implies that $a$ and $b$ are not connected if and only if there exists $S \subset \Omega_r$ such that $(a \perp b \mid S)$. Minimality then implies that $a$ and $b$ cannot be adjacent in the true model. □

In words, the lemma above says that if two variables *that the receiver is aware of* 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 in the receiver's model as well. Conversely, if the two variables are not directly adjacent in the true model and receiver is aware of enough controls ($S$) to disentangle them, then the receiver must not connect the two either. So the receiver must make accurate judgments regarding which variables are adjacent and which are not. One possible misperception not ruled out by the lemma is that the receiver can draw a link that does not exist in the true model, due to lacking awareness of some other variables.

**Example 5.** *Suppose the true model is $a \to b \leftarrow c$. If the receiver is aware of all three variables, they constitute a V-structure, hence the true model is uniquely identified by the IC algorithm. If the receiver is only aware of $\Omega_r = \{a, c\}$, then they must think that $a$ and $c$ are not connected, since $a \perp c$.*




(a) true model (b) IC algorithm

Figure 7: A true model without consistent models for observable variables $a, b, c, d$.

**Example 6.** *Suppose the true model is $a \to b \to c$, and the receiver is only aware of $\Omega_r = \{a, c\}$. Then both $a \to c$ and $c \to a$ are consistent with the data, since $a \not\perp c$. The latter model is defective. The former model, $a \to c$, draws a non-existent link from $a$ to $c$, but is not defective according to our definition, since $a \Rightarrow c$ in the true model. The receiver's belief that $a$ affects $c$ is correct in this case, and we assume that their unawareness that this effect goes through $b$ is immaterial.*

Note that a consistent model might not exist for some subsets $\Omega \subset \Omega_t$ of variables. In this case, the IC algorithm would produce an inconsistent model if used; an example follows below. In our analysis, however, we require the receiver to start off with a consistent model in the first place and the sender to supply the data together with a consistent model. Hence, whenever we apply the IC algorithm, the underlying assumption is that a consistent model exists for a given $\Omega$.

**Example 7.** *Suppose the true model is as in Figure 7(a), and the receiver is only aware of variables $\Omega_r = \{a, b, c, d\}$ (so $e \notin \Omega_r$). Running the IC algorithm with variables $\Omega_r$ produces the model in Figure 7(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 $d \to b \leftarrow a$ with control $c$. In Step 3, the algorithm: (1) orients link $b \to c$ using rule R1 (which avoids creating a V-structure $a, b, c$ that would have been identified); and (2) orients $d \to c$ 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 $\Omega_r$.*

### 4.3 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 7.

**Definition 5.** *The true model $(\Omega_t, C_t)$ is* simple *if for any $a, b, c \in \Omega_t$ that form a collider $a \to_t b \leftarrow_t c$, variables $a$ and $c$ are not correlated.*

**Definition 6.** *The true model $(\Omega_t, C_t)$ 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. In other words, it requires that 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 $S = \emptyset$.

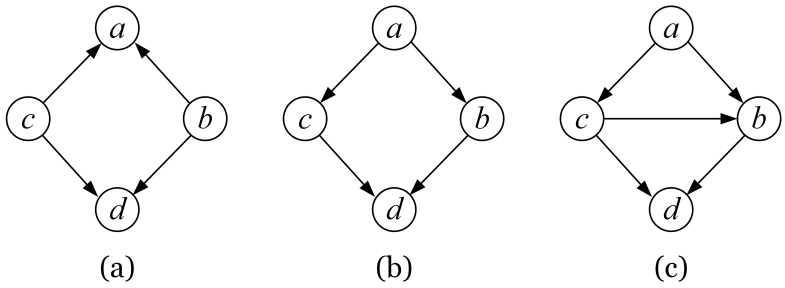



Figure 8: Examples of simple (a and c) and non-simple (b) models.

**Example 8.** *The models in Figure 8(a) and (c) are simple, but the model in (b) is not. In panel (a), $c \to a \leftarrow b$ and $c \to d \leftarrow b$ form V-structures, but because $b \perp c$ they do not require any controls. In (b), $c \to d \leftarrow b$ form a V-structure, but identifying it requires $a$ as control, since $c \not\perp b$ and $(c \perp b \mid a)$. In (c), $c \to d \leftarrow b$ is not a V-structure anymore because of the link $c \to b$. Similarly, the true models in Figures 6(a) and 7(a) are not simple; whereas the true model from the illustrative example in Figure 1(a) is simple.*

We assume that even if the true model is simple, the receiver is unaware of this fact. Therefore, we allow their original model $(\Omega_r, C_r)$ and the sender's proposed model $(\Omega_s, C_s)$ to be non-simple, so long as they are consistent with $P|\Omega_r$ and $P|\Omega_s$, respectively.

In turn, richness requires that the true model debunks any other model. This is equivalent to requiring that the true model can be uniquely identified using the IC algorithm. This means that every directed link in the true model 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), 7(a) and 8(a) are rich, whereas the models in Figures 4(a,b,c), 6(a) and 8(b,c) are not.

# 5 When can the receiver's model be debunked?

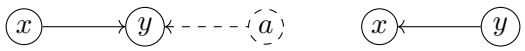
The sender's problem in our model is two-fold. First, they choose which new variables $\Omega_s \backslash \Omega_r$ to present to the receiver to render their initial model $(\Omega_r, C_r)$ inconsistent with the new data. This incompatibility would be contained in one or more causal links from the receiver's model that cannot exist in any model consistent with the new evidence. We then say that model $(\Omega_s, C_s)$ *debunks a link* $x \leftarrow_r y$ for some $x, y \in \Omega_r$ if $x \to y$ in any model consistent with $P|\Omega_s$. The receiver's model $(\Omega_r, C_r)$ is *debunked* if any of its links is debunked. Our main results in this section 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. This is effectively a tie-breaking rule prescribing which (consistent) model the receiver should adopt in case their old model is debunked. Section 6 relies on this assumption to discuss when the sender can successfully persuade the receiver, and we discuss alternative modelling approaches in Section 7.

We start by showing that our notion of a defective link (and a defective model) provides a necessary condition for what can be debunked.

**Theorem 1** (No debunking non-defective links.)**.** *In a simple world, the sender can never debunk a non-defective link.*

*Proof.* See Appendix. □




(a) (b)

Figure 9: Obvious cause.

Theorem 1 argues that the sender can only debunk defective links in the receiver's model, which is those causal relations that the receiver believes, $a \to_r b$, but which do not exist in reality, $a \not\Rightarrow_t b$. In particular, if the receiver's belief $a \to_r b$ omits some proxy variable $c$—meaning that the true model is such that $a \to_t c \to_t b$—then revealing variable $c$ would only make the receiver expand their model, but not abandon it. This result immediately implies the following corollary.

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

The 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 model can be debunked by presenting the true model $(\Omega_t, C_t)$; this is the definition of a rich model.

We next show that if the true model is simple and rich, then under some conditions, any defective model can be debunked by disclosing at most two new variables, so debunking a defective model is "easy". As we show later, in case these conditions do not hold, debunking a model is "difficult" in that it may require disclosing arbitrarily many variables even in a simple rich world.

**Definition 7.** *Given $x, y \in \Omega_t$, we call $l \in \Omega_t$ a* latent variable for $(x, y)$ *if $C_t$ contains a path $l \Rightarrow_t x$ that does not pass through $y$ and a path $l \Rightarrow_t y$ that does not pass through $x$. We denote the set of such latent variables as $\Omega_L(x, y) \subset \Omega_t$.*

**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 $z \in \Omega_t$ a* non-obvious cause of $y$ given $x$ *if $z \in C_t(x) \backslash \Omega_L(x, y)$.*

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

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. □

**Example 9.** *To illustrate the first part of Theorem 2, suppose the true model is as in Figure 9(a), and the receiver's model is as in Figure 9(b). In this case, $a$ is the obvious cause of $y$ given $x$, so by revealing it the sender is able to debunk and flip the defective link.*

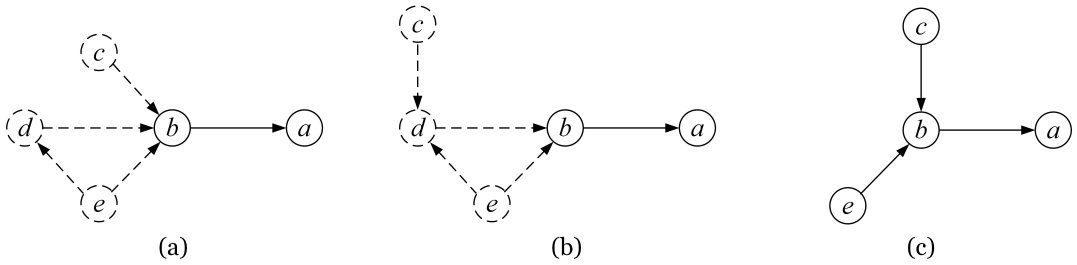



Figure 10: Example application of Theorem 2 (part 2).

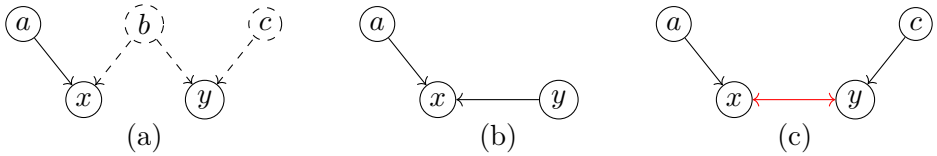



Figure 11: Debunking may not lead to persuasion.

**Example 10.** *For another example, consider the illustrative example in Figure 1(a): tenure $t$ is an obvious cause of earnings $w$ given education $e$. The business school uses the obvious cause $t$ to debunk the employee's original subjective model. This latter example also illustrates that Theorem 2 can be applied to deceive the receiver: the sender there is able to convince the receiver that $e \to_s w$ by showing an obvious cause, but this is a defective link as well, produced by the existence of omitted latent variables.*

**Example 11.** *To illustrate the second part of Theorem 2, suppose the true model is as shown in Figure 10(a), and the receiver's model is $\Omega_r = \{a, b\}$ with $b \leftarrow_r a$. In this case, no obvious cause of $a$ given $b$ exists. However, because the world is rich, we must be able to trace the link $b \to a$ back to some V-structure. In this case, revealing either $c$ and $e$, or $c$ and $d$ 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 10(b), then by revealing $c$ and $e$, the sender will be able to debunk the defective link as well, which will result in a subjective model as in Figure 10(c).*

We should 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 more variables even in the cases covered by Theorem 2. An example is presented below.

**Example 12.** *Suppose the true model is as in Figure 11(a), and the receiver's model is $a \to_r x \leftarrow_r y$, depicted in Figure 11(b). This model can be debunked by revealing just one variable, $c$, which is an obvious cause of $y$ given $x$. However, then both $a, x, y$ and $x, y, c$ constitute V-structures, with the former implying $x \leftarrow y$ and the latter implying $x \to y$, as depicted in Figure 11(c). These contradicting implications mean that there is no consistent model for $\Omega = \{a, x, y, c\}$. If the sender is required to propose a consistent model, they must also disclose latent variable $b$ in addition to the obvious cause $c$ and propose the true model from Figure 11(a).*

What if the premise of Theorem 2 does not hold? Is it easy to debunk a defective model if there are no obvious or non-obvious causes? In what follows, we consider two scenarios that suggest that debunking can be arbitrarily difficult in this case, in the sense of requiring the sender to reveal arbitrarily many new variables.

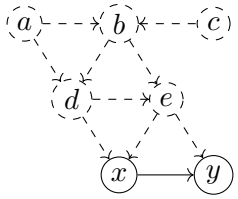



Figure 12: No obvious or non-obvious causes exist for $y$ given $x$.

Consider first the case when $x \to_t y$, $x \leftarrow_r y$, and no obvious or a non-obvious cause for $y$ given $x$ exist in a simple rich world. Simplicity restricts heavily how the true causal graph can look in this case. Specifically, all other variables $z \in \Omega$ must be either proxies ($x \to_t z \to_t y$), or latent variables ($x \Leftarrow_t z \Rightarrow_t y$). It is immediate that revealing proxy variables alone can not help debunk the defective link $x \leftarrow_r y$. We can further construct an example, where revealing latent variables is not helpful either—unless the sender reveals all of them.

**Example 13.** *Suppose the true model is as depicted in Figure 12. This model is indeed simple (the only direct V-structure is $a \to_t b \leftarrow_t c$) and rich (the true model is fully identified), and there exist no obvious or non-obvious causes of $y$ given $x$ (all other variables are latent for $x, y$). Suppose further that the receiver's subjective model is given by $\Omega_r = \{x, y\}$ and $x \leftarrow_r y$. One can see, by bruteforcing all possible combinations, that debunking the defective link $x \leftarrow_r y$ requires the sender to reveal all other variables $a, b, c, d, e$. One can expand this model by adding more variables between $c$ and $x, y$ to create an example where arbitrarily many variables must be revealed to debunk $x \leftarrow_r y$.*

Next, we look at the case when the receiver's model has defective link $x \leftarrow_r y$ (while $x \not\Leftarrow_t y$), but $x \not\Rightarrow_t y$, and rather the two variables have one or more common causes. In this case, we show in the proposition below that debunking the receiver's model can be done by revealing *all* irreducible latent variables. Since there can be arbitrarily many such variables, the sender may need to present a very elaborate and complicated model to the receiver, who may have a difficult time following this complexity. We therefore label this persuasion method as "difficult", and argue that if one of the methods outlined in Theorem 2 are available to the sender, they are preferred.

**Definition 10.** *A latent variable $z \in \Omega_L(x, y)$ is* irreducible *if $\bar{C}_t(z) \cap \Omega_L(x, y)$ is empty.*

**Proposition 1.** *Suppose the true model is simple, and the receiver's model has defective link $x \leftarrow_r y$ when $x \not\Leftarrow_t y$ and $x \not\Rightarrow_t y$. Then the receiver's model can be debunked by revealing all irreducible latent variables in $\Omega_L(x, y)$ (which is nonempty).*

*Proof.* Conversely, suppose that $\Omega_s$ includes the set $\bar{\Omega}_L(x, y) \subseteq \Omega_L(x, y)$ of all irreducible latent variables. Then $x \perp y \mid \bar{\Omega}_L(x, y)$, meaning the receiver identifies the absence of a link between $x$ and $y$ in Step 1 of the IC algorithm. This debunks $(\Omega_r, C_r)$. □

**Example 14.** *To illustrate Proposition 1, consider Figure 13. In both panels, $c$ and $d$ are latent variables, however, only $d$ is irreducible. Proposition 1 states that to disprove the connection between $a$ and $b$, it is enough to reveal all irreducible variables, i.e., it is enough to show $d$. Note, however, that this result is sufficient, but not necessary: revealing $c$ (but not $d$) would also be sufficient in Figure 13(b), but would not suffice in Figure 13(a).*

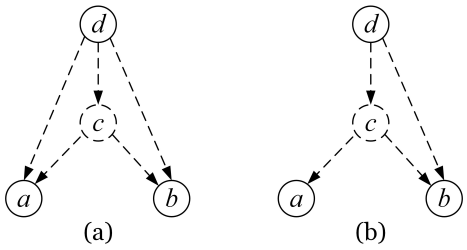



Figure 13: Example application of Proposition 1.

# 6 When can the sender persuade the receiver?

The previous section asked the question of "when can a defective model be easily debunked?" In this section, we proceed to answer the main question: "when can the sender persuade the receiver?" In doing so, we consider different cases corresponding to whether the sender wants to persuade the receiver that a link exists or does not exists, as well as whether the respective link is present in the true model or not.

## 6.1 Sender wants to persuade of a link

We start with a case when the sender wants to persuade the receiver that $x \Rightarrow_s y$ for some $x, y$. 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 $x \not\Rightarrow_r y$.

**Truth is on the sender's side.** In a rich world, the sender can always the sender wants to persuade the receiver that $x \Rightarrow_s y$ when truth is on their side—i.e., the true model $(\Omega_t, C_t)$ 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 the sender to reveal one or two additional variables. This is summarized by the following corollary.

**Corollary 2.** *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 with at most two variables.*

**Truth is against the sender.** If the true model is such that $x \Leftarrow_t y$, then Theorem 1 shows that the sender can never debunk *this* link (i.e., present such data that only $x \Rightarrow_s y$ is consistent with it). However, this does not mean that persuasion is impossible, 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 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 15.** *Suppose the true world is as shown in Figure 14(a). The receiver observes variables $a$, $b$ and $c$ and thinks the model is as shown in Figure 14(b). Specifically, the receiver correctly believes that $a \to_r b$, but also mistakenly believes that $c \to_r a$. The sender would like to deceive the receiver and persuade them that $a \leftarrow_s b$. Because $a \to_t b$ is*

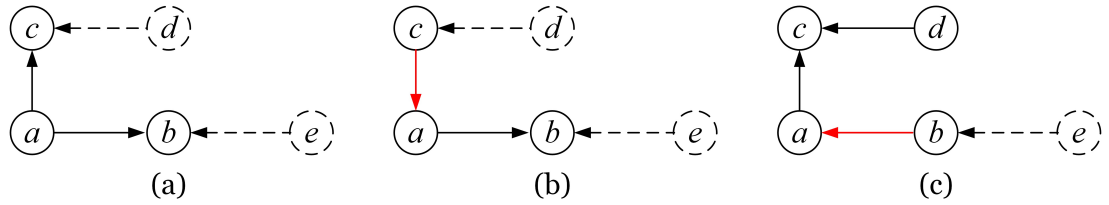



Figure 14: Example: persuasion by nitpicking.

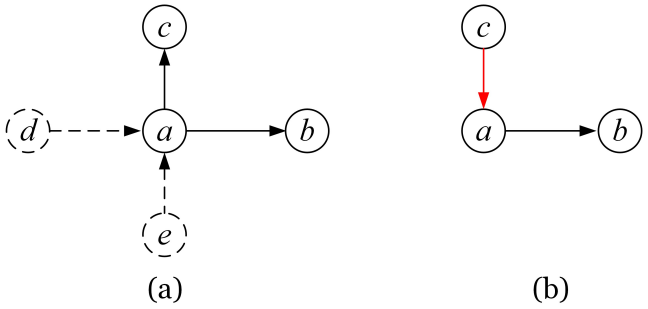



Figure 15: Example: persuasion by nitpicking not possible.

*the truth, the sender cannot debunk this link directly. Instead, the sender can debunk the defective link $c \to_r a$. In this example, the sender is able to reveal variable $d$, which is the obvious cause for $c$ given $a$. The sender then will offer model shown in Figure 14(c), which will be accepted by the receiver.*

Note, however, that successful deception along the lines of the example above requires that the link of interest must not be oriented by the IC algorithm given $\Omega_r$. In other words, models with both $x \Rightarrow y$ and $x \Leftarrow y$ must be consistent with the data available to the receiver. Otherwise, if the 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.

**Example 16.** *Suppose the true model is as shown in Figure 15(a), and the receiver's subjective model is as in Figure 15(b). The receiver correctly thinks $a \to_r b$ but incorrectly thinks $c \to_r a$. The sender would like to persuade the receiver that $a \leftarrow_s b$. However, in this case, the defective link $a \to_t c$ and $a \to_t b$ both trace back to the same V-structure $d, a, e$. Thus, to debunk $c \to_r a$, the sender has to reveal $d$ and $e$. But revealing $d$ and $e$ also makes $a \leftarrow_s b$ inconsistent with the data, so deception is not possible.*

**Truth is more complicated.** If the true model is such that $x$ and $y$ are not adjacent, but rather there exist some latent variables $\Omega_L(x, y)$ (such that $z \Rightarrow_t x, y$), 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, since revealing additional variables can only destroy some links in the receiver's model but cannot create new spurious correlations and links.

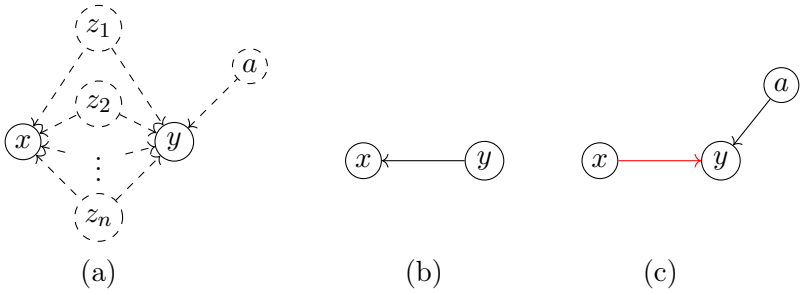



Figure 16: Replacing a defective model with a defective model.

If, on the other hand, $x \Leftarrow_r y$ (due to the receiver being unaware of some of the latent variables $z \in \Omega_L(x, y)$), then the sender can flip this link and persuade the receiver that $x \Rightarrow_s y$. Note that this is possible even though there is no actual link between $x$ and $y$ in the true model, so the sender can debunk the receiver's defective model and replace it with another defective model. An example follows.

**Example 17.** *Suppose the true model is as depicted in Figure 16(a), and the receiver's model is $x \leftarrow_r y$, as in Figure 16(b). Then the sender can persuade the receiver that $x \rightarrow_s y$ by revealing variable $a$, in which case $x, y, a$ form a V-structure in the data.*

However, such deception is, again, 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 18.** *Suppose the true model is as depicted in Figure 11(a), and the receiver's model is as in Figure 11(b). As argued in Example 12, while variable $c$ 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 $\Omega_s = \Omega_r \cup \{c\} = \{a, c, x, y\}$. If the sender must present a consistent model, their only choice is between leaving in place the receiver's model from Figure 11(b) with $x \leftarrow_r y$ and revealing the true model from Figure 11(a) with $x$ not adjacent to $y$.*

## 6.2 Sender wants to break a link

What if the sender wants to break 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 $x \not\Rightarrow_s y$). By minimality, this requires showing the receiver a set $S$ of variables such that $(x \perp y \mid S)$, which only exists if truth is on the sender's side in the sense that $x$ and $y$ are not adjacent in the true model, as argued by the following result.

**Proposition 2.** *Suppose the true model is simple, and the receiver's model has defective link $x \leftarrow_r y$ when $x \not\Leftarrow_t y$. Then persuading the receiver that $x$ and $y$ are not connected requires revealing some set $S$ of variables such that $(x \perp y \mid S)$.*

*If $x \not\Rightarrow_t y$, then the set $\bar{\Omega}_L(x, y)$ of all irreducible latent variables is one such set $S$.*

*Proof.* If there is no such set $S \subset \Omega_s$, then $(x \not\perp y \mid S')$ for all $S' \subset \Omega_s$, so the receiver identifies a link between $x$ and $y$ in Step 1 of the IC algorithm. By Proposition 1, the set $\bar{\Omega}_L(x, y)$ of all irreducible latent variables is one such set $S$, since $\big(x \perp y \mid \bar{\Omega}_L(x, y)\big)$. □

The result above implies that even if such persuasion is possible, it may be "difficult" in the sense of requiring the sender to disclose arbitrarily many variables. This is illustrated by the following example.

**Example 19.** *Suppose the true model is as depicted in Figure 16(a). The set of (irreducible) latent variables for $x, y$ in this case is $\bar{\Omega}_L(x,y) = \Omega_L(x,y) = \{z_1, z_2, ..., z_n\}$. Only conditioning on all variables from this set allows one to see that $x$ and $y$ are uncorrelated: $(x \perp y \mid S)$ only if $S = \Omega_L(x,y)$. Consequently, if the receiver is unaware of at least one latent variable from this set, $x$ and $y$ will be adjacent in their model. If the sender wants to persuade the receiver that there is no link between $x$ and $y$, they must reveal all variables in $\Omega_L(x,y)$ that the receiver is not already aware of. By increasing $n$, we can make the number of such variables arbitrarily large.*

It follows also that if the sender's objective is to convince the receiver that $x \not\Leftarrow_s y$ when the correlation between $x$ and $y$ is actually driven by a large number of latent variables $\{z_1, ..., z_n\}$, then it might be easier for the sender to convince the receiver that $x \Rightarrow_s y$ than to prove the truth. One can see this by comparing Examples 17 and 19.

# 7 Discussion

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

## 7.1 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 might exist other variables that they do not observe. This approach encounters a conceptual issue that latent variables can explain *any* data. Specifically, any dataset can be explained by assuming the existence of some latent variable $d$ that affects all other observed variables: $d \to x$ for all $x \in \Omega$ ("God's will" is one example of such $d$; various conspiracy theories also rely on the existence of such unobserved—and unobservable—latent factors). It is unclear then how the receiver should choose between multiple consistent models that rely on latent variables.

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.[11] 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 latent variables on their own in some cases. For example, in the situation in Figure 7(b), the receiver will realize that the model produced by the IC algorithm is not consistent with the data, so there must be a latent variable affecting $b$ and $d$. The receiver will not observe its distribution, but will become aware of its existence and will even infer the correct causal structure.

[11] Gottesman (2025) argues, in a slightly different 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.

**Receiver is certain of their model.** We assume that the receiver starts off with a singular model in mind, and only when this model becomes irreconcilable with the data does the receiver open up to other models. Given that the receiver's initial model is 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 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; Ispano, 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, the receiver first tries to expand their initial model $(\Omega_r, C_r)$ to incorporate new data. However, when new data is incompatible with the existing model, the model is *revised*. In our case, this revision entails abandoning the receiver's initial model and accepting the sender's model. A more minimal revision would have the receiver keep parts of the initial model that do not contradict the data; this approach is discussed below.

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 throws out the whole debunked model.** We assume that when facing new evidence incompatible with their initial model, the receiver throws away the entire model and simply accepts the sender's proposed model (assuming it is consistent with the new data). A more cautious receiver might instead opt for a smaller revision and only discard the debunked link, attempting to then find a new consistent model that is as close as possible to their initial model. Our results continue to hold in this setting if the sender is trying to persuade the receiver of something that is true. Deception, however, is harder in this scenario. For example, Figure 14(c) presented an example of a "wrong link persuasion": debunking the receiver's model by exposing one wrong link, and then offering a model where another link is also flipped (which was correct in the receiver's initial model). This trick would no longer work with a "minimally-revising" receiver. Since Theorem 1 implies that non-defective links can never be falsified, the only remaining scope for deception in this case would be due to latent variables, as in the illustrative example in Section 2.

### 7.2 Assumptions regarding the sender and the world.

**Sender 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

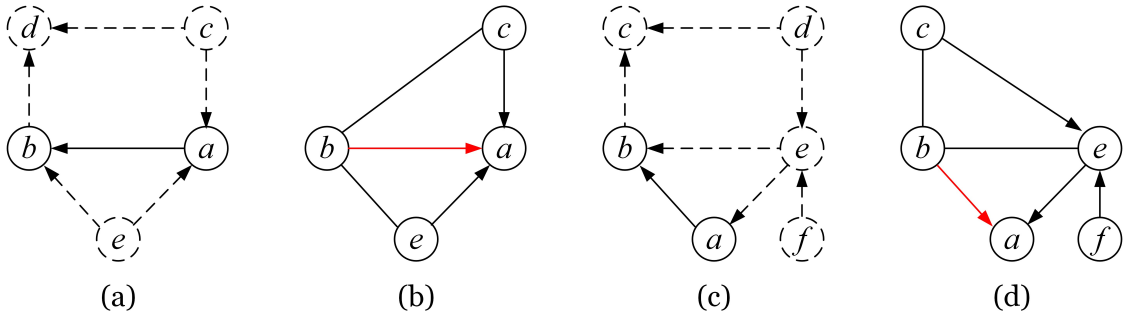



Figure 17: Examples: direct wrong link persuasion in non-simple world.

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 easen the receiver's causal discovery problem. We find this to be an appealing avenue for future research.[12]

**Rich worlds.** Some of our results (specifically, Theorem 2) assumed the world is rich—namely that the true graph can be uniquely identified. If a world is not rich, then debunking is more difficult, since a link then needs to be falsifiable—on top of being defective—to be debunked. It follows that persuasion about something that is truthful becomes more difficult. At the same time, conditional on debunking the receiver's model, deception may be simpler because the receiver would have less evidence in favor of the true model.

**Simple worlds.** All of our results are phrased for simple worlds—those where all V-structures are obvious in the sense of not requiring any additional control variables. If the true model is not simple, then any defective model can still be debunked using the same "causal variables", but now also require revealing some controls. Most of our other results continue to hold in this case, with the caveat that controls must be included.

One exception, however, is Theorem 1: in non-simple worlds the receiver can be persuaded of something that is false (as opposed to simply accepting a spurious correlation). Figure 17 presents two examples of this. Suppose first the true model is as shown in Figure 17(a). The receiver correctly believes that $a \to_r b$, and the sender wants to persuade them that $b \to_s a$. To do so, the sender can reveal variables $c$ and $e$. The IC algorithm will identify the V-structure $e, a, c$, and then will apply R3 to redirect $b \to_s a$, resulting in the partially directed graph as shown in Figure 17(b). Similarly, suppose the true model is as shown in Figure 17(c) and the sender is interested in flipping the truthful link $a \to_r b$. The sender is able to achieve this by revealing variables $c$, $e$ and $f$. Then the IC algorithm will identify the V-structure $c, e, f$ and then it will apply R1 to $f \to e \to a$. Finally, it will apply R4 to obtain the link $b \to a$. The resulting partially directed graph will be as shown in Figure 17(d). Note that both examples rely on non-simple true models: in Figure 17(a), $c, d, b$ is a V-structure but only conditionally on $a$, i.e., $(b \perp c \mid a)$; whereas in Figure 17(c), $b, c, d$ is a V-structure but only conditionally on $e$, i.e., $(b \perp d \mid e)$.

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

# 8 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 causality can be established and debunked.

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 only communicating truthfully. We show that the scope for deception exists in this case, with the sender being sometimes able to replace the receiver's defective (incorrect) model with another defective model. This may sometimes involve flipping the direction of a causal link between two variables from a correct to an incorrect direction, meaning that the sender can use reverse causality to trick the receiver (albeit this requires a more substantial mistake elsewhere in the receiver's subjective model). Conversely, if a sender wants to persuade the 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, like discussed above.

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

# A Appendix

## A.1 Proof of Theorem 1

Suppose $x \Leftarrow_t y$ and $x \leftarrow_r y$ for some consistent receiver's model. We show that this link cannot be debunked by disclosing new variables, i.e., $x \rightarrow_s y$ cannot be uniquely identified by the IC algorithm.

**Direct case.** We start with the case $x \leftarrow_t y$ and show that the IC algorithm cannot produce a model on a subset of variables with $x \rightarrow_s y$. The link can be incorrectly oriented in Step 2 or Step 3 of the IC algorithm. We show below that any such flip requires one of the following:

1. that at least one arrow already was incorrectly identified, 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", the first arrow to be incorrectly identified in the IC algorithm, since any such mistake requires that another mistake has already been made.

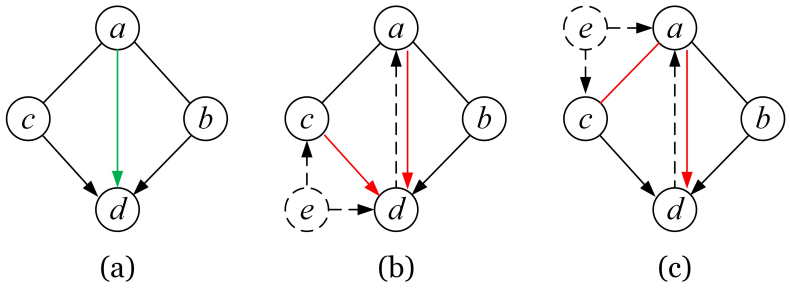



Figure 18: R3 and required non-simple models to generate an incorrectly directed link.

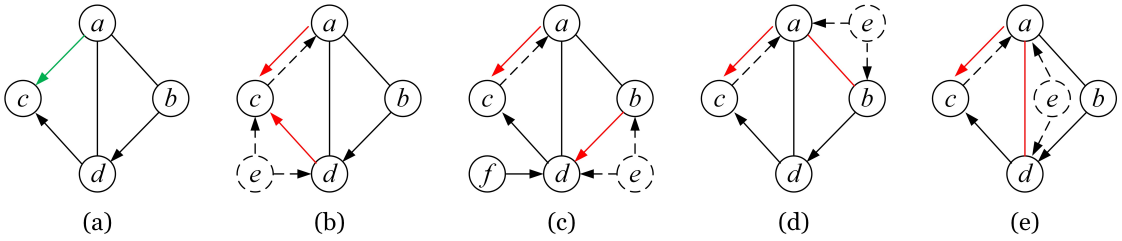



Figure 19: R4 and required non-simple models to generate an incorrectly directed link.

To flip a direction of a link (i.e., to identify $x \to_s y$ when $x \leftarrow_t y$) in Step 2, we need to create a new direct V-structure $x \to_s y \leftarrow_s a$ for some $a$, but this will not be a V-structure if $y \leftarrow_t a$, since then $x \not\perp a$ in the data/true model. Hence this requires that link $y \leftarrow_s a$ has been mis-identified (then we have traced back to an already existing incorrect link) or comes from a hidden latent variable (in which case still $x \not\perp a$).

To flip a direction of a link in Step 3, we need to do it via one of Meek's rules. Rule R1 cannot do such a flip. R1 identifies $x \to_s y$ if $a \to_s x$ for some $a$ not adjacent to $y$. But if $a \to_t x$ then $a \to_t x \leftarrow_t y$ form a direct V-structure. Hence mis-identifying $x \to_s y$ via R1 requires that $a \to_s x$ has also been mis-identified. (Note that we cannot mis-identify a lack of adjacency between $a$ and $y$: if they are adjacent in $(\Omega_t, C_t)$, they must be adjacent in $(\Omega_s, C_s)$, hence non-adjacency in the sender's model means they are not adjacent in the true model.) Also, another possibility is that $a \not\to_t x$, and rather there's a latent variable $b \in \Omega_L(a, x)$. In this case, however, $b, x, y$ still form a V-structure.

Rule R2 can obviously not do such a flip unless one of the other link directions has been mis-identified. For either of rules R3 and R4 to identify an incorrect link direction, one of the following cases must apply:

1. At least one of two other required link directions have been mis-identified. In this case, we have traced back to an already existing incorrect link.

2. At least one of two other required link directions come from a hidden latent variable.

   (a) Consider R3 in Figure 18(a). If link $c \to d$ was incorrectly identified due to omitted variable $e$, then it must have come from the underlying true model shown by black arrows in Figure 18(b). If $a \to_t c$, then we have a non-simple V-structure $a, c, e$ conditional on $d$. Alternatively, if $c \to_t a$, then note that we must have $b \to_t a$, so we have a V-structure $b, a, c$ that would have been identified by the receiver, yet it was not. So we either have a non-simple true model or a contradiction.

(b) Consider R4 in Figure 19(a). Then we have two directed links to examine:

i. If link $d \to_s c$ was incorrectly identified due to omitted variable $e$, then it must have come from the underlying true model shown in Figure 19(b). In this case, the link would be identified in the opposite direction (as $c \to_s d$) since $c, d, b$ form a V-structure conditional on $a$ in the data. So we would have a contradiction.

ii. If link $b \to_s d$ was incorrectly identified due to omitted variable $e$, then it must have come from the underlying true model shown in Figure 19(c). For example, the sender would reveal variable $f$ to support a V-structure $b, d, f$, which would also imply the link $d \to_s c$ from R1. So, generally, this would be an example of a situation that could flip a truthful link. However, this is not possible in a simple world, since the true model in Figure 19(c) is not simple: either $c, a, b$ form a V-structure conditional on $d, e$ or $a, b, e$ form a V-structure conditional on $c, d$.

3. At least one of two other required blank links are not identified in $\Omega_s$.

(a) Consider R3 in Figure 18(a). If the "empty" edge $a - c$ has been incorrectly identified due to omitted variable $e$, then it must have come from the underlying true model shown by black arrows in Figure 18(c). Then we have a non-simple V-structure $e, a, d$ conditional on $c$.

(b) Consider R4 in Figure 19(a). Then we have two blank links to consider.

i. If the blank edge $a-b$ has been incorrectly identified due to omitted variable $e$, then it must have come from the underlying true model shown by black arrows in Figure 19(d). However, then we have a non-simple V-structure $c, a, e$ conditional on $b, d$.

ii. If the blank edge $a-d$ has been incorrectly identified due to omitted variable $e$, then it must have come from the underlying true model shown by black arrows in Figure 19(e). However, then we have a non-simple V-structure $c, a, e$ conditional on $d$.

If $x \leftarrow_t y$, then we have shown that to flip a correct link requires either non-simple world or another existing incorrect link, but because it has to trace back to the original source of mistake, this finishes the proof (since it also follows from Lemma 1 that $x$ and $y$ must then be adjacent in any consistent subjective model).

**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 aware of all variables that are between $x$ and $y$ in the true model, then a link direction can never be mis-identified, hence $x \Rightarrow_s y$ can never be identified. Suppose then that there exists variable $a$ such that $x \leftarrow_t a \leftarrow_t y$ and $a \notin \Omega_r$. We want to show that the IC algorithm can not identify $x \to_s y$ in this situation. (The case with more than one hidden variable in between $x$ and $y$ then follows by induction.)

Note that if $x$ is the only child of $a$, or if $a$ is the only parent of $x$, then we can treat them 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 $x \to_s y$ cannot be identified as a direct V-structure in Step 2 of the IC algorithm using the same argument as above.

We now show that the link $x \to_s y$ cannot be identified in Step 3 of the IC algorithm. In doing so, we only need to consider the cases not covered above.

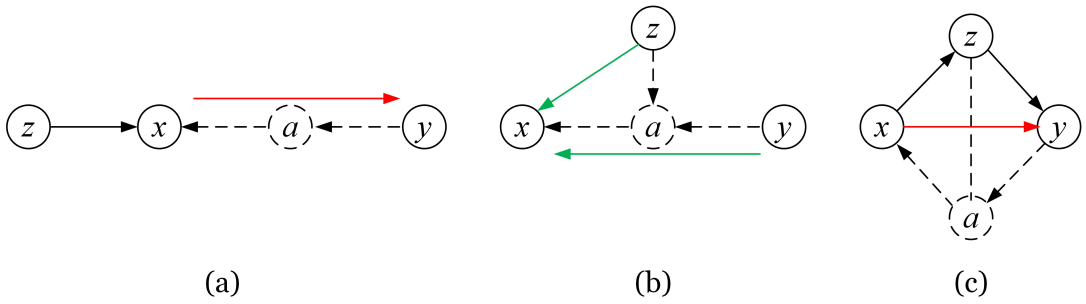



Figure 20: R1 and R2 and required models to generate an incorrectly directed link.

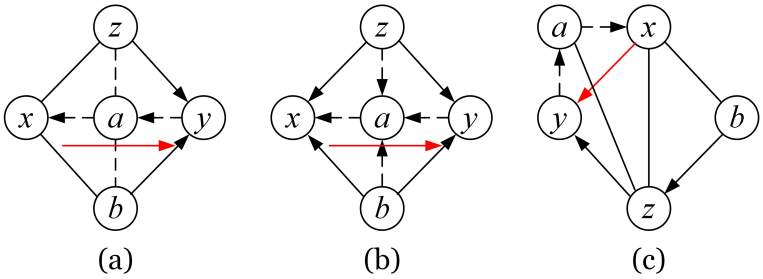



Figure 21: Cases R3 and R4 for the proof of Theorem 1.

Suppose rule R1 identifies the link $x \to_s y$ from some $z \to_s x$ non-adjacent to $y$. The logic from the above direct case shows that this cannot happen when $z$ is adjacent to $x$ in the true model (regardless of whether it is adjacent to $a$), see Figure 20(a) for illustration. Suppose then that $x \leftarrow_t a \leftarrow_t z, y$. But then $z, x, y$ form a direct V-structure in $\Omega_s$, as shown in Figure 20(b), which is a contradiction. Hence, again it must be that the link $z \to_s x$ has been already mis-identified.

Suppose rule R2 identifies the link $x \to_s y$ from some identified links $x \to_s z \to_s y$. Then $x, a, z, y$ form a rhombus in the true model as shown in Figure 20(c). Then there must either be a cycle (so contradiction) or one of the R2 implying links has already been mis-identified.

Suppose rule R3 identifies the link $x \to_s y$ from some identified links $z \to_s y \leftarrow_s b$ for some $z$ and $b$ adjacent to $x$ in the sender's model, see Figure 21(a) for illustration. To avoid cycles in the true model, we must have $z \to_t x$ and $b \to_t x$. If there is no relationship between $z, a$ and $b$, then we get that $z, x, b$ is a V-structure that should have been identified but it was not. If $z$ and $b$ are adjacent with $a$, then by acyclicity, we have $z \to_t a$ and $b \to_t a$. So the true model must be as it is shown in Figure 21(b). However, in this case, we still obtain a V-structure with $z, x, b$, which must have been identified in Step 2, meaning R3 does not apply unless one of already existing links has been mis-identified earlier.

Suppose rule R4 identifies the link $x \to_s y$ from some identified links $b \to_s z \to_s y$ for some $b$ and $z$ adjacent to $x$, see Figure 21(c) for illustration. To avoid cycles, we must have $b \to_t x$ in the true model. However, then we get a non-simple V-structure $a, x, b$ conditional on $z$. Hence, this implies that a past mistake must have been already present.

Again by applying the argument that a mis-identified link must trace back to some original source of a mistake, but it does not. This completes the proof.

## A.2 Proof of Theorem 2

We begin by establishing a supplementary Lemma.

**Lemma 2.** *Suppose the true model is simple and there exist $x, y \in \Omega_t$ such that $x \Rightarrow_t y$ and $\Omega_L(x, y)$ is nonempty. Then $x, y, z$ must be adjacent for any $z \in \Omega_L(x, y)$. We call such $x, y, z$* a triangle with root $z$.

E: This lemma is incorrect, see Figure 12 for a counterexample.

*Proof.* Proceed by contradiction: if $z \Rightarrow_t x$ but not $z \rightarrow_t x$, then there exists $\omega_i$ s.t. $z \Rightarrow_t \omega_i \Rightarrow_t x$, but then $\omega_i$ becomes a control variable for V-structure $x \Rightarrow_t y \Leftarrow_t z$: $(x \perp z \mid \omega_i)$ and $(x \not\perp z \mid \omega_i, y)$. If this V-structure is direct, this contradicts $(\Omega_t, C_t)$ being simple. Otherwise, there exists a direct V-structure $\omega_j \rightarrow_t y \leftarrow_t \omega_k$ with $x \Rightarrow_t \omega_j$, $\omega_k \Leftarrow_t z$, for which $\omega_i$ is also a control variable, which contradicts $(\Omega_t, C_t)$ being simple. Similar contradictions arise in all other cases. □

**Lemma 3.** *Suppose the true model $(\Omega_t, C_t)$ is simple and rich and there exist $x, y \in \Omega_t$ such that $x \Rightarrow_t y$, there exists no obvious cause of $y$ given $x$, and $\Omega_L(x, y)$ 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 $(\Omega_t, C_t)$ 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, $\Omega_t$. 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.

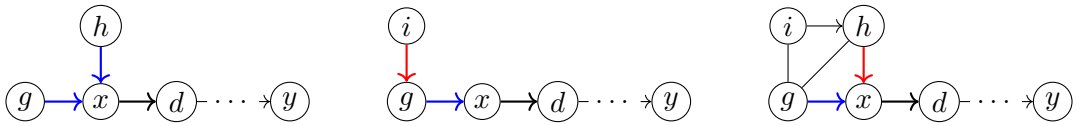
Consider a path from $x$ to $y$ and take some variable $d \in \Omega_t$ such that $x \rightarrow_t d \Rightarrow_t y$ (possibly with $d = y$ if $x$ and $y$ are adjacent). Focus on the link $x \rightarrow_t d$. Proceed according to cases outlined above.

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 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 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:

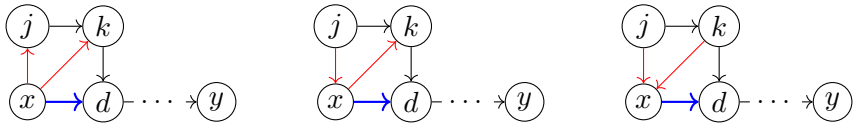
   (a) If $g \rightarrow_t x$ was oriented in Step 2 of the IC algorithm as a part of some direct V-structure $g, x, h$ then this is exactly the V-structure required by the lemma, so we are done. This case is depicted in Figure 22(a).

   (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 22(b).




(a) (b) (c)

Figure 22: Options for orienting link $g \to_t x$.




(a) (b) (c)

Figure 23: Options for orienting link $x \to_t d$ via R4.

(c) If $g \to_t x$ was oriented in Step 3 by rule R4, then there must exist another link $h \to_t x$ that was oriented earlier in the algorithm. This case is depicted in Figure 22(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 $x \to_t d$ is oriented in Step 3 by Meek rule R4. Then there exist $j, k \in \Omega_t$ adjacent to $x$ such that $j \to_t k \to_t d$, with these latter links having already been oriented by the IC algorithm. Focus on the link $j \to_t k$ and consider different cases for how it could have been oriented and how $x, j, k$ are connected (see Figure 23):

   (a) If $x \leftarrow_t k$, as in Figure 23(c), then $k$ is a latent variable for $x, y$, which violates the premise of the lemma requiring that $\Omega_L(x, y)$ is empty.

   (b) Link $j \to_t k$ 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 $x \to_t k$ (we are in cases (a) or (b) of Figure 23), then $x \perp l$, so $l$ is an obvious cause of $y$ given $x$—a contradiction to the premise of the lemma.

   (c) Link $j \to_t k$ was oriented in Step 3 by Meek rule R1. Then there exists $m \in \Omega_t$ s.t. link $m \to_t j$ was oriented earlier in the algorithm. If $x \to_t j$, then such $m$ is an obvious cause of $y$ given $x$, which is a contradiction. If $x \leftarrow_t j$, then $m \to_t j$ 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 $j \to_t k$ was oriented in Step 3 by Meek rule R4. Then there exist $n, o \in \Omega_t$ adjacent to $j$ such that $n \to_t o \to_t k$, with these latter links having already been oriented by the IC algorithm. We can then repeat the analysis in case 3, focusing on link $n \to_t o$. Proceeding by induction, we will either find a V-structure upstream from $x$, or arrive at a contradiction.

This concludes the proof of Lemma 3. □

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 accordign to the cases from the theorem.

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

By assumption, $x$ and $y$ are adjacent in $(\Omega_r, C_r)$. Consider $\Omega_s \equiv \Omega_r \cup w$. Then $x$ and $y$ are still adjacent in any consistent $(\Omega_s, C_s)$, since $(x \not\perp y \mid S \cup w)$ for any $S \subset \Omega_r$. If $y$ and $w$ are adjacent in $(\Omega_s, C_s)$, then $x, y, w$ constitute a direct V-structure in $P|\Omega_s$ because $x \Rightarrow_t y \Leftarrow_t w$ (so $w \not\perp x \mid y$). The receiver must identify this V-structure in Step 2 of the IC algorithm and conclude that $x \rightarrow_s y$, which debunks $(\Omega_r, C_r)$. If $y$ and $w$ are not adjacent in $(\Omega_s, C_s)$, but there exists a path $y \Rightarrow_s w$ s.t. all variables $z \in \Omega_r$ along this path (i.e., such that $w \Rightarrow_s z \Rightarrow_s y$) are also obvious causes of $y$ given $x$, then we arrive at a contradiction: for some such $z$, $(x, y, z)$ constitute 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 $w$ are not adjacent in $(\Omega_s, C_s)$, and every path $w \Rightarrow_t y$ contains some $z \in \Omega_r$ such that $w \Rightarrow_s z \Rightarrow_s y$ and $z \not\perp x$. Fix one such path and take such $z$ which is closest to $w$ in $(\Omega_s, C_s)$ along that path. It is without loss to assume that $w$ and $z$ are adjacent. Then $z$ must be adjacent to $x$ in $(\Omega_r, C_r)$ and $(\Omega_s, C_s)$, since otherwise—if there existed $S \subset \Omega_r$ such that $(x \perp z \mid S)$—$x, y, z$ 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 $x \leftarrow_r y$.

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

**Case 2:** $x \Rightarrow_t y$ and there is a non-obvious cause $z \in C_t(x) \backslash \Omega_L(x, y)$ of $y$ given $x$. If there also exists an obvious cause $w$, then case 1 above applies and proves the statement, hence for the remainder of this proof assume there is no obvious cause of $y$ given $x$. If $\Omega_L(x, y)$ is non-empty, disclosing $z$ and any $l \in \Omega_L(x, y) \cap C_t(x)$ creates a direct V-structure $z, x, l$ 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 rule R1, debunking link $x \leftarrow_r y$ and the receiver's model $(\Omega_r, C_r)$.

If $\Omega_L(x, y) = \emptyset$, then by Lemma 3, there exists a direct V-structure $a \rightarrow_t b \leftarrow_t c$ with $b \in \bar{C}_t(x) \cup x$. 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 $(\Omega_s, C_s)$, then the receiver must identify a direct V-structure $a \rightarrow_s x \leftarrow_s c$ because: (i) $a \perp c$, but (ii) $(a \not\perp c \mid b)$ implies $(a \not\perp c \mid x)$. Then Meek rule R1 orients link $x \rightarrow_s y$ because $(a \perp y \mid x)$, debunking $(\Omega_r, C_r)$. Alternatively, if $x$ is not adjacent to $a, c$ in $(\Omega_s, C_s)$, 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 $(\Omega_s, C_s)$, which must be identified by the receiver in Step 2 of the IC algorithm. Meek rule R1 then orients all links along the path $d \Rightarrow_s x$ and, eventually, the link $x \rightarrow_s y$, debunking $(\Omega_r, C_r)$. 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.

Bergemann, D. and Morris, S. (2019). Information Design: A unified Perspective. *Journal of Economic Literature*, 57(1):44–95.

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. (2024). False Narratives and Political Mobilization. *Journal of European Economic Association*.

Eliaz, K. and Rubinstein, A. (2025). Wasonian Persuasion. Working paper.

Gottesman, A. (2025). A limitation of Evidence-Backed Communication. Working paper.

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.

Ispano, A. (2025). The perils of a coherent narrative. *Economic Theory*.

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.

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.

Meek, C. (1995). Causal Inference and Causal Explanation with Background Knowledge. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–410.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, pages 380–391.

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. _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1449.

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).

Schwartzstein, J. and Sunderam, A. (2021). Using Models to Persuade. *American Economic Review*, 111(1):276–323.

Spiegler, R. (2016). Bayesian Networks and Boundedly Rational Expectations. *The Quarterly Journal of Economics*, 131(3):1243–1290.

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.

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].