

# **Causal Narratives**

By

**Constantin Charles**  
**Chad Kendall**

**FINANCIAL MARKETS GROUP DISCUSSION PAPER NO. 929**

**June 2025**

**Any opinions expressed here are those of the authors and not necessarily those of the FMG. The research findings reported in this paper are the result of the independent research of the authors and do not necessarily reflect the views of the LSE.**

# Causal Narratives

Constantin Charles and Chad Kendall\*

May 27, 2025

## Abstract

We study causal narratives – narratives which describe a (potentially incorrect) causal relationship between variables. In a series of experiments across a range of data-generating processes, we show that externally provided causal narratives influence decisions in ways inconsistent with rational theory. Instead, decisions are generally consistent with a behavioral theory, but there is significant heterogeneity in responses driven by the fact that subjects construct their own causal models of the data-generating process. To examine these homegrown models directly, we ask subjects to observe a dataset and offer advice to future subjects. The resulting homegrown narratives reveal that many subjects mistake correlation for causation, leading both themselves and those who receive their advice to make systematic errors.

---

\*Constantin Charles: Department of Finance, London School of Economics (e-mail [c.charles@lse.ac.uk](mailto:c.charles@lse.ac.uk)). Chad Kendall: Department of Economics, University of Miami and National Bureau of Economic Research (e-mail [cxk757@miami.edu](mailto:cxk757@miami.edu)). We thank Kai Barron, Jean-Pierre Benoit, Pedro Bordalo, Alexander Coutts, Kfir Eliaz, Cary Frydman, David Hirshleifer, Ryan Oprea, Ran Spiegler, and seminar participants at the Berlin Behavioral Economics Seminar, the BRIQ Beliefs Conference, Carnegie Mellon University, the CESS Mental Models Workshop, Chapman University, Claremont Graduate University, the ESA World Meetings, the Frankfurt School of Finance & Management, Georgia State University, Indiana University, the LSE Behavioral Political Economy Conference, the IIPF Annual Conference, the Online Seminar in Economics and Data Science at ETH Zurich, Penn State Smeal, SAFE, Texas A&M, the University of Alabama, the UCI Finance Conference, the University of Miami, the University of Ottawa, the University of Toronto, the University of Pittsburgh, the University of Southern California, the University of Texas at Dallas, the University of Utah, and the University of Virginia for valuable input. Supplementary material is available [here](#).

# 1 Introduction

Causal narratives – narratives that tell a causal story about the relationship between variables of interest – are ubiquitous. We see them in economics (“the pandemic interrupted the supply chain, causing inflation”), in politics (“immigration leads to job losses for locals”), in medicine (“social media causes depression”), and elsewhere in everyday life. Some of these narratives reflect true causal relationships; others mislead by presenting mere correlations as causation. Understanding when and how such narratives are effective is critical for understanding how people come to form beliefs and opinions on a vast range of topics.

While causal narratives may influence people through various channels – such as emotion, memory, or salience – recent theoretical work focuses on one in particular: the idea that narratives shape beliefs by providing causal interpretations of correlations in the data (Spiegler (2016); Eliaz and Spiegler (2020); Eliaz, Galperti, and Spiegler (2024)). If this assumption is correct, it offers a tractable means of incorporating narratives into economic theory, allowing us to study their impacts on issues ranging from financial bubbles to political polarization (Shiller (2017, 2019)).

In this paper, we conduct a series of controlled experiments, both online and in an experimental laboratory, to understand which types of causal narratives are most effective and why. In our experiments, we pit standard theory against a behavioral theory that predicts precisely how beliefs can be manipulated through causal narratives. Controlled experiments are ideal for this purpose because they allow us to vary both the narratives and the data-generating processes (DGPs) that subjects are exposed to, thereby isolating the effect of narratives themselves from confounding factors – such as prior beliefs or a narrative’s source – that might influence the effectiveness of narratives in the real world.

To see how causal narratives can influence beliefs, consider a concerned parent who hears the claim that social media causes depression in teenagers. This narrative provides a model for interpreting data: it posits a causal chain from a parental ban (an action), to social media use (an auxiliary variable), to depression (an outcome). Suppose that, for the sake of illustration, the true relationship reflects reverse causality – depression increases social media use, not the other way around. A parent who understands this would recognize that banning social media would not improve their

child’s well-being, and would have no reason to do so. But a parent who accepts the narrative and forms beliefs accordingly may instead choose to impose a ban.

The problem is that the correlations in the data, when interpreted through the lens of the narrative, can distort beliefs. Using directed acyclic graphs (DAGs), where arrows represent causal links, the true model in this hypothetical example is:  $ban \rightarrow social\ media\ use \leftarrow depression$ . This structure implies that banning social media has no causal effect on depression; the two variables are independent. In contrast, the narrative suggests a different model:  $ban \rightarrow social\ media\ use \rightarrow depression$ . Given the positive correlation between social media use and depression, this narrative may lead parents to mistakenly believe that banning social media will reduce depression.

In our experiment, we test the effectiveness of several types of causal narratives but primarily focus on two types: *Chain* narratives imply a causal chain from action, to auxiliary variable, to outcome, as in the social media and depression narrative. *Collider* narratives, by contrast, imply that the action and auxiliary variable both directly affect the outcome, rather than being linked in a chain. To give an illustrative example, consider a gun rights activist who argues that criminals have guns, and that we need to arm citizens to counteract this threat ( $arm\ citizens \rightarrow public\ safety \leftarrow criminals$ ).

In our first series of treatments, we seek to understand how such causal narratives influence subjects’ beliefs and actions across a range of DGPs. Importantly, we use generic variable labels and values rather than a specific context like social media, in order to shut down any prior beliefs about the relationships among the variables that subjects may have. Subjects observe several compact datasets in sequence, each of which describes the joint correlation between three binary variables: an action, an outcome, and an auxiliary variable. For each dataset, a rational subject can identify the causal effect of the action on the outcome simply by counting the frequencies of each outcome conditional on each action in the dataset.

After observing a dataset, subjects form beliefs about the relationship between actions and outcomes, and then choose an initial *policy* – a probability distribution over the two actions. Subjects are paid more for good outcomes than bad, and pay a cost that increases for policies further from one-half. After choosing initial policies, subjects observe a narrative and make a second policy choice. Finally, subjects observe

a competing narrative alongside the first narrative and make a third policy choice.

We begin with datasets that are designed such that the action has no causal effect on the outcome (“Independent datasets”). Within these datasets, we vary how the auxiliary variable is generated: in some cases, it is independent of both the action and the outcome; in others, it is correlated with both. These correlations are irrelevant to a rational subject, who would always choose the least costly policy of one-half. However, a causal narrative that leverages these correlations may lead a behavioral subject to deviate from one-half.

Recent theoretical models of causal narratives (Spiegler (2016); Eliaz and Spiegler (2020); Eliaz, Galperti, and Spiegler (2024)) rest on the behavioral assumption that people form beliefs according to the Bayesian-network factorization formula (BNFF). This formula makes a point prediction for beliefs, taking only the joint distribution described by a dataset and the DAG implied by a causal narrative as inputs (i.e., it has no free parameters). Under the Chain narrative, the BNFF predicts that beliefs will be distorted such that policies above one-half are expected to produce the good outcome more frequently. In contrast, under the Collider narrative, policies below one-half are expected to yield better outcomes. The fact that two different narratives are predicted to have opposite effects on beliefs and policy choices for the *same* dataset is a key prediction of interest.

When causal narratives can leverage correlations in the data, we find that they robustly induce policies different from the rational prediction of one-half in the directions predicted by the BNFF, both in our online sample and in the lab. These effects persist even as we make the correlations noisier or less salient. In our online sample, we further test narratives in a dataset in which the action *does* have a causal effect on the outcome, so that a rational subject would choose a policy of *less* than one-half. Even for these datasets, we find that Chain narratives cause many subjects to instead choose policies *greater* than one-half, policies which produce the good outcome less often. By leveraging policy choices in datasets with no correlations - where causal narratives should have no effect - we show that these effects cannot be driven by inattention or potential experimenter demand effects.

While the BNFF organizes the overall results reasonably well, we find two key departures from assumptions in existing theoretical work on narratives (e.g., Eliaz and

Spiegler (2020); Schwartzstein and Sunderam (2021)). First, initial policy choices suggest that many subjects infer erroneous causal relationships from the correlations in the data *even without being exposed to a narrative*. Thus, any externally provided narrative must implicitly compete with a subject’s homegrown model. These homegrown models vary widely across subjects, leading to substantial heterogeneity in how narratives influence behavior. In particular, because roughly half of subjects’ initial policies align with the direction predicted by the Chain narrative, policy choices after Chain narratives tend to deviate more from the rational benchmark than those after Collider narratives.

Second, when subjects are presented with two competing narratives, their average policy choices tend to fall between those observed when each narrative is presented in isolation. This pattern, like the competition with subjects’ homegrown models, suggests that participants do not simply adopt a single narrative, but instead weight multiple causal models simultaneously. We confirm this interpretation in a follow-up treatment: when asked to report their beliefs about the underlying causal model, 92% of participants assign positive probability to both models.

To better understand subjects’ homegrown models, our second pair of treatments directly examines the models subjects construct for themselves. As before, subjects view datasets, but now they are incentivized to generate their own narratives by offering free-form advice to future participants (Schotter (2023)). The resulting advice reveals substantial variation, with roughly half of all narratives indicating false causal relationships. About 15% closely mirror the Chain narratives we designed, while very few resemble Collider narratives, supporting our hypothesis that subjects are more prone to infer mistaken Chain structures on their own. Notably, subjects who produce homegrown Chain narratives not only make mistakes themselves, but also mislead others. These findings show that false causal narratives can emerge, spread, and shape beliefs, even when there is no intent to deceive.

In our final treatment, we provide evidence of the external validity of two of our key results by leaving behind the abstract nature of our primary treatments. First, we find that when asked to generate real-world examples of Chain and Collider narratives, subjects produce more Chain narratives and report finding them easier to construct. This suggests that Chain narratives come more easily to subjects, consistent with our

finding that subjects are more likely to infer causal Chain models from abstract data. Second, we show that even in real-world contexts, where one might expect strong prior beliefs to dominate, subjects remain willing to entertain multiple causal models at once.

In the literature, the importance of causal narratives in politics is highlighted by Stone (1989), who argues that political actors deliberately associate events with *causal* stories in order to motivate partisan support for their side. Within economics, Shiller (2017, 2019) sparked interest in formally modeling narratives. For our purposes, the most directly relevant contribution is Eliaz and Spiegler (2020), building on Spiegler (2016), who propose a conceptual framework in which narratives are represented as causal graphs. We test key assumptions of their innovative framework. Related work by Schwartzstein and Sunderam (2021), Izzo, Martin, and Callander (2023), and Aina (2025) examines how a principal can persuade an agent through a narrative represented as a model of the underlying DGP.<sup>1</sup> Although we don’t test these other models explicitly, our experiment provides some of the first available evidence (together with Barron and Fries (2024)) that persuasion via models (as opposed to signals or Bayesian persuasion (Kamenica and Gentzkow (2011))) can be effective.

In a related experiment, Andre et al. (2025) survey people about the causes of recent inflation, map their responses to DAGs, and test the power of narratives to influence self-reported inflation expectations. We complement their work by testing the power of causal narratives in a setting in which we control the true DGP, allowing us to tightly engage with theory and compare different types of causal narratives. Ambuehl and Thyssen (2024) study competing causal narratives in a setting that, by design, forces subjects to choose one narrative or another, ruling out the ‘weighting’ behavior we identify. Barron and Fries (2024), Graeber, Roth, and Zimmermann (2024), and Morag and Loewenstein (2024) study other conceptions of narratives in experimental settings.

Given that causal narratives can be thought of as mental models that are used to interpret data, our work also connects to a recent experimental literature studying how people form and get stuck in mental models (Enke (2020); Esponda, Vespa, and Yuksel (2024); Kendall and Oprea (2024); Graeber (2023); Fan (2024)). In partic-

---

<sup>1</sup>Benabou, Falk, and Tirole (2018) theoretically study narratives as they relate to morality norms.

ular, Frechette, Yuksel, and Vespa (2024) study whether subjects form the correct statistical (not necessarily causal) model when presented with datasets generated by different DAGs.

Cognitive scientists have formalized how causal and statistical processes differ (Pearl (2009); Sloman (2009)), and conducted experiments to study how people perceive (and misperceive) causal relationships (see Waldmann and Hagmayer (2013) and Matute et al. (2015) for recent reviews). We draw on this literature to identify potential mechanisms that guide our experimental design (Section 3.1). The psychology literature on illusory correlation (Chapman (1967)) and apophenia or patternicity (Conrad (1958); Shermer (2008)) documents how people often perceive patterns in random data (typically in visual contexts). Due to our focus on causality, our setting is conceptually different and perhaps more closely related to the hot hand or gambler’s fallacies in which people think streaks of independent draws will continue or reverse (Rabin (2002); Asparouhova, Hertzels, and Lemmon (2009)). Our findings provide further evidence that people have difficulty understanding random processes, often times seeking to explain such randomness through a causal story.

## 2 Conceptual Background

### 2.1 Datasets and Rational Benchmark

We consider environments in which there are only three variables involved in the construction of a narrative: an action ( $a$ ), an auxiliary variable ( $z$ ), and an outcome ( $y$ ). All of the variables are binary, taking values 0 and 1. We describe a joint distribution function,  $p(a, z, y)$ , via a *dataset*. Figure 1 displays a pair of datasets from our experiment as subjects saw them ( $a \in \{BLUE(1), GREEN(0)\}$ ,  $z \in \{\blacktriangle(1), \circ(0)\}$ ,  $y \in \{ON(1), OFF(0)\}$ ). In both datasets,  $a$  and  $y$  are statistically independent, and both values of  $a$  and  $y$  are equally likely. In the left dataset, which we call *Strong*,  $z$  is generated as the logical AND of  $a$  and  $y$ , so that each is strongly correlated with  $z$ .<sup>2</sup> In the right dataset, which we call *Null*,  $z$  is statistically independent of  $a$  and  $y$  (and

---

<sup>2</sup>In a previous experiment (documented in Appendix C), we also tested a dataset in which  $z = 1$  when  $a = 0$  and  $y = 1$ . We found symmetric results (swapping  $a = 0$  for  $a = 1$ ) relative to the Strong dataset.



Figure 1. Strong and Null Datasets

Choice	Z	Y
BLUE	○	OFF
GREEN	○	ON
BLUE	▲	ON
BLUE	○	OFF
GREEN	○	ON
GREEN	○	ON
GREEN	○	OFF
BLUE	○	OFF
BLUE	▲	ON
GREEN	○	ON
BLUE	▲	ON
BLUE	▲	ON
GREEN	○	OFF
GREEN	○	OFF
BLUE	○	OFF
GREEN	○	OFF

Choice	Y	Z
GREEN	OFF	▲
GREEN	OFF	○
BLUE	ON	○
GREEN	OFF	○
BLUE	ON	▲
BLUE	OFF	○
BLUE	ON	○
BLUE	ON	▲
BLUE	OFF	○
BLUE	OFF	▲
GREEN	ON	▲
GREEN	ON	▲
GREEN	ON	○
GREEN	OFF	▲
GREEN	ON	○
BLUE	OFF	▲

Notes: In the Strong dataset on the left, Z is generated as the logical AND of Choice and Y. In the Null dataset on the right, Z is statistically independent of Choice and Y. The ordering of rows, and of columns Y and Z, was randomized across subjects.

each value is equally likely) so that all correlations are null. With the understanding that (i) a dataset exhaustively describes all possible combinations of the variables and (ii) each row in the dataset is equally likely, a dataset completely describes  $p(a, z, y)$ .

The decision-maker (DM) is interested in the causal effect of the action variable, Choice, on the outcome, Y, because she would like to obtain the good outcome, ON, as often as possible. In general, observing the joint distribution is insufficient to determine this causal effect. However, as we discuss in Section 2.3, all of the datasets in our experiment are constructed such that the causal relationship can be inferred when the DM knows that the action variable is exogenous.<sup>3</sup> Given this fact, and the fact that each row in a dataset is equally likely, a rational subject can simply count the frequencies of each outcome to determine how often ON occurs for each of the two choices. For either of the datasets in Figure 1, doing so gives  $p(Y = ON|BLUE) = p(Y = ON|GREEN) = \frac{1}{2}$ .

In addition to the Strong and Null datasets, we use five more datasets in the experiment. The *Noise* dataset weakens the correlations in the Strong dataset by perturbing some of the  $z$  values (see Figure A1 in Appendix A). The *Masked* dataset masks the correlations by using more symbols (see Figure 2 in Section 3.1). The final three datasets are *Causal* datasets that map one-to-one to the Strong, Noise,

<sup>3</sup>We inform subjects of this exogeneity by saying that their choice will have the same effect on the other variables as it does in the dataset. We also tell subjects that no other (hidden) variables impact any of the observed variables in any way.

and Null (together, *Independent*) datasets with one key difference: in the Causal datasets, a rational subject *should* infer a causal relationship from  $a$  to  $y$  because  $p(y = ON|a = BLUE) = \frac{1}{3}$  while  $p(y = ON|a = GREEN) = \frac{2}{3}$ .

## 2.2 Narratives

Suppose a DM is presented with the following *Chain* narrative when studying the Strong dataset:

*“Z is a  $\blacktriangle$  only when the choice is BLUE. Further, when Z is a  $\blacktriangle$ , Y is always ON. So, choose BLUE more often than GREEN.”*

The pattern highlighted in this narrative is completely factual – it can be verified with the data at hand. But, it suggests a false causal relationship in which the DM’s choice influences Z, which in turn influences Y. A DM that hears this narrative might come to believe that she can increase the probability of the good outcome, ON, by choosing BLUE more often.

Instead, suppose that the DM is presented with the *Collider* narrative:

*“If Z is a  $\circ$ , Y is always OFF when the choice is BLUE. To counteract this, choose GREEN more often than BLUE so that Y can be ON even if X is a  $\circ$ .”*

Again, the pattern highlighted in this narrative is factual. But, this narrative implies a causal relationship in which the DM’s choice and Z both influence Y directly. A DM believing it may try to increase the probability of ON by choosing GREEN more often. Thus, for the same dataset, different narratives can potentially cause a DM to make different choices. To understand the importance of the auxiliary variable, consider the Null dataset instead. In this dataset, the patterns highlighted by the previous two narratives do not exist, so it is not possible to construct causal narratives that exploit correlations in the data.

The Chain and Collider narratives are key narratives we test in our experiment.<sup>4</sup> Together we refer to them as *elaborate* narratives: they use the third, auxiliary variable to tell a story. We compare elaborate narratives with two other types of narratives, *simple* narratives and *Summary* narratives.<sup>5</sup> Simple narratives provide only a

<sup>4</sup>For the Noise and Masked datasets, we modify each slightly to reflect the fact that the patterns no longer hold deterministically (i.e., we replace ‘always’ with ‘more often’, etc.).

<sup>5</sup>The elaborate, simple, and Summary narratives are exhaustive in the sense that the theoretically-predicted beliefs for any other DAG would be the same as for one of them.

prescription: “Choose BLUE more often than GREEN” (the *Up* narrative) or “Choose GREEN more often than BLUE” (the *Down* narrative). Summary narratives summarize the relationship between  $a$  and  $y$  in a 2x2 table, thus emphasizing the true relationship. In addition, they explicitly tell subjects to choose the rational policy of one-half for Independent datasets, and to choose “more GREEN” for Causal datasets. We discuss optimal policies for our setting in Section 2.4.

### 2.3 Directed Acyclic Graphs and Beliefs

In our framework, narratives convey causal models that suggest how the world works. To describe these causal models, we follow Pearl (2009) and use directed acyclic graphs (DAGs). Within economics, Spiegler (2016) suggests the use of DAGs as a way to describe the subjective (behavioral) beliefs of a DM who faces a joint probability distribution, and Eliaz and Spiegler (2020) use DAGs as a means of describing narratives.<sup>6</sup> DAGs are parameter-free descriptions of causal models that use directed links to describe the direction of the causal relationships between variables, but not the associated ‘strengths’ of the links (i.e., conditional probabilities).

For the Strong dataset, the true DAG is given by  $a \rightarrow z \leftarrow y$ . A rational DM who infers this DAG from the dataset would understand that  $a$  has no influence on  $y$  because there is no causal pathway between  $a$  and  $y$ . Understanding this, the DM can readily infer  $p(y|a = 1) = p(y|a = 0) = \frac{1}{2}$ . Pearl (2009) shows that two DAGs are observationally equivalent (*compatible* with the same set of joint distributions) if and only if they share the same *skeleton* (the set of undirected links) and *v-structures* (converging arrows whose tails are not connected by an arrow). Importantly, this implies that a rational DM can uniquely determine the DAG from the Strong dataset because no other DAGs have the same skeleton and v-structures: flipping the link from  $a$  to  $z$  or from  $z$  to  $y$  (or both) breaks the original v-structure that converges on  $z$ . In fact, we construct all of our datasets such that they imply unique DAGs.<sup>7</sup>

The Chain and Collider narratives from the previous section imply causal models (DAGs) that are *incompatible* with the joint distribution and therefore potentially lead to incorrect beliefs. The implied DAG and the joint distribution together determine

<sup>6</sup>Glazer and Rubinstein (2021) use DAGs to describe (not necessarily causal) stories.

<sup>7</sup>For the Causal datasets, the DM’s knowledge of  $a$  being exogenous is necessary in addition to Pearl’s criteria.

the conditional probabilities, which can be calculated using the Bayesian-network factorization formula (BNFF). The BNFF provides a normative description of how a DM *should* form beliefs given knowledge of the true causal model and the joint distribution. Of course, if the causal model the DM believes is wrong, incorrect beliefs may follow.<sup>8</sup>

For example, consider the Chain narrative which implies the causal relationship,  $a \rightarrow z \rightarrow y$ . Here, the BNFF prescribes  $p(y|a) = \sum_{z=0,1} p(y|z)p(z|a)$  (see Spiegler (2016) for the general form of the BNFF). When compared to the conditional law of total probability,  $p(y|a) = \sum_{z=0,1} p(y|z, a)p(z|a)$ , which must hold for any joint distribution, the conditioning of  $y$  on  $a$  is dropped, which is how the BNFF can lead to incorrect beliefs. In particular, for the Strong dataset in Figure 1, the BNFF prediction for the Chain narrative is  $p(y = 1|a = 1) = \frac{2}{3}$ , which is greater than the true probability,  $p(y = 1|a = 1) = \frac{1}{2}$ .

The Collider narrative, by contrast, implies the causal relationship,  $a \rightarrow y \leftarrow z$ . The BNFF prescribes,  $p(y|a) = \sum_{z=0,1} p(z)p(y|a, z)$ , which again can lead to incorrect beliefs. For the Strong dataset in Figure 1, the BNFF predicts  $p(y = 1|a = 1) = \frac{1}{4}$ . In contrast, for the Null dataset, the BNFF prediction for both the Chain and Collider narrative is the truth,  $p(y = 1|a = 1) = \frac{1}{2}$ , because there are no correlations in the dataset that can be wrongly interpreted as causation. As a result, both causal narratives are predicted to produce null effects in the Null dataset.

Simple narratives also suggest a causal relationship between  $a$  and  $y$ , but one that is ambiguous. A DM hearing such a narrative could infer one of the elaborate narrative DAGs or simply the direct relationship,  $a \rightarrow y$  (with  $z$  unrelated). In the latter case, the BNFF leads to the truth, so that beliefs under a simple narrative depend upon the DM’s interpretation of the narrative. Finally, for the datasets in Figure 1, Summary narratives also imply a causal model - one in which there is no causal relationship between the action and outcome. For Causal datasets, they instead imply the direct causal relationship,  $a \rightarrow y$ .<sup>9</sup>

---

<sup>8</sup>Cognitive scientists have put forth other models of causal reasoning (Waldmann and Hagmayer (2013)), but the BNFF is the only theory of which we are aware that provides quantitative behavioral predictions in our environment.

<sup>9</sup>Importantly, by implying distinct DAGs, narratives provide different information - namely, different models of the world. Thus, there is no sense in which we can compare causal and non-causal narratives that provide the same ‘information’. Instead, our interest is in assessing whether

## 2.4 From Beliefs to Actions

The BNFF provides behavioral predictions about conditional beliefs. To map beliefs to observable actions, we adopt the setup of Eliaz and Spiegler (2020). We incentivize subjects according to

$$u(y, d) = y - c(d - d^*)^2 \quad (1)$$

where  $d$  is the policy choice variable that determines the frequency at which  $a = 1$  is played (i.e.,  $d = p(a = 1)$ ),  $d^*$  is a policy from which deviations are costly, and  $c$  is a scale variable that determines the cost of deviating from  $d^*$ . This incentive scheme is similar to a belief elicitation mechanism such as the quadratic or binarized scoring rules, except that both beliefs,  $p(y = 1|a = 1)$  and  $p(y = 1|a = 0)$ , affect policy choices.

Given subjective beliefs,  $p_G(y|a)$ , induced by a narrative,  $G$ , a DM chooses the policy,  $d$ , to maximize

$$\max_d d \cdot p_G(y = 1|a = 1) + (1 - d) \cdot p_G(y = 1|a = 0) - c(d - d^*)^2$$

where the conditional probabilities are held fixed based on the historical joint distribution, as in our experiment. For example, for the Strong dataset, a rational DM would simply choose  $d = d^*$  because she would realize  $p_G(y = 1|a = 1) = p_G(y = 1|a = 0) = \frac{1}{2}$ . By contrast, for a DM who believes the Chain narrative, we can solve the DM's problem using  $p_G(y = 1|a = 1) = \frac{2}{3}$  and  $p_G(y = 1|a = 0) = \frac{1}{3}$ . The optimal policy is  $d = d^* + \frac{1}{6c}$ , which is strictly above the rational policy,  $d^*$ . For the Collider narrative,  $p(y = 1|a = 0, z = 1)$  is indeterminate because the combination of  $a = 0$  and  $z = 1$  never occurs in the joint distribution.<sup>10</sup> We handle this case empirically by allowing for any subjective belief,  $\gamma = p(y = 1|a = 0, z = 1) \in [0, 1]$ , so that  $p_G(y = 1|a = 0) = \frac{\gamma}{4} + \frac{3}{8}$ . The optimal policy is then given by  $d = d^* + \frac{1}{2c} \left( -\frac{1}{8} - \frac{\gamma}{4} \right)$  which lies on the opposite side of the rational policy compared to the Chain narrative, for any subjective belief,  $\gamma$ .

In the experiment, subjects received \$1.00 for  $y = ON(1)$  and \$0 otherwise. We

---

causal narratives can influence beliefs, and if so, which types of narratives do so more effectively.

<sup>10</sup>The Noise and Masked datasets instead have full support.

chose  $d^* = \frac{1}{2}$ , so that we can observe deviations in either direction for Independent datasets. We set  $c = \frac{2}{3}$ , to achieve separation between the behavioral and rational predictions, while ensuring that flat incentives are not responsible for the results. With  $c = \frac{2}{3}$ , a subject who deviates to one of the most extreme policies (0 or 1) earns one third less (on average) than a subject who chooses rationally for an Independent dataset. Given these parameter choices, a rational subject would choose  $d = 0.5$  for Independent datasets and  $d = 0.25$  for Causal datasets. In Table A1 of Appendix A, we summarize the predictions for the narrative and dataset combinations for both the rational and behavioral (BNFF) theories. We omit simple narratives because of their ambiguity about the causal relationship.

### 3 Receiving Narratives

Here, we describe three treatments (BASELINE, SALIENCE, and BELIEFS) in which subjects observe externally provided narratives. Each of these treatments broadly consists of three main steps that are repeated for each of three datasets (which vary across treatments). Subjects first observed only a dataset and chose an initial policy. They then observed the same dataset alongside a narrative and made a second policy choice. Finally, subjects observed an additional narrative and made a third policy choice while facing the two competing narratives simultaneously.

#### 3.1 Experimental Design

In the BASELINE treatment, we placed subjects into one of two arms. Subjects in the first arm observed the Independent datasets (Strong, Noise, and Null), while subjects in the second arm observed the corresponding Causal datasets (Causal Strong, Causal Noise, and Causal Null), in random order. The Independent datasets consisted of 16 rows, while the Causal datasets consisted of 12 rows. We described the datasets as summarizing thousands of historical observations such that each row occurred an equal number of times. We also explicitly told subjects that (i) the variables will maintain the same relationships (if any) they had in the past and (ii) no other ‘hidden’ variables influence the observed variables in any way. In both treatment arms, for each dataset, we asked subjects to complete the following tasks in order:

1. We presented the dataset. We told subjects that the variables may or may not be related and asked them to study the dataset to identify any relationships.
2. We asked subjects to choose a policy (the probability with which each action would be taken) using a slider. The slider had no default – subjects had to make a choice. The outcome,  $y$ , then realized and subjects received a payoff according to equation (1), in dollars. We gave subjects no feedback on the realizations of  $y$  or their payoffs until the end of the experiment to ensure that their beliefs remain fixed, as assumed in the theory.
3. We asked subjects to rate (on a scale from 0-100) how certain they were that their chosen policy maximizes their earnings. These questions were not incentivized.
4. We provided subjects with a narrative alongside the dataset. Importantly, we framed all narratives as advice that may or may not be useful, and asked subjects to assess the advice for themselves. Subjects could review the dataset and advice simultaneously, allowing them to form subjective conditional expectations,  $p_G(y|a)$ . Subjects then made a second policy choice, rated how certain they were that their policy choice maximizes their earnings, and indicated whether they found the advice helpful or unhelpful.
5. We provided subjects with a second narrative alongside the narrative from step 4 (randomizing which appears first), except in the case of the Null and Causal Null datasets. For these datasets, we instead provided a second narrative on its own. In all cases, subjects could review the dataset and the piece(s) of advice together. We then asked subjects to make a third and final policy choice, to rate how certain they were that their policy choice maximizes their earnings, and to indicate whether they found the piece(s) of advice helpful.

















We assigned the narratives from Section 2.2 based on the dataset subjects viewed, as follows.

**Strong and Noise datasets:** In step 4, we presented either a simple or elaborate narrative, randomized across subjects.<sup>11</sup> If we presented a simple narrative in step 4, we presented it again in step 5 together with the Summary narrative. If we instead presented an elaborate narrative in step 4, we presented it again in step 5 with either

---

<sup>11</sup>We presented simple narratives to 40% of subjects and elaborate narratives to 60% of subjects. We oversampled elaborate narratives to allow for more observations of these narratives in step 5.

Figure 2. Masked Dataset

Choice	Z	Y
GREEN		ON
BLUE		OFF
BLUE		OFF
BLUE		ON
GREEN		ON
BLUE		OFF
BLUE		ON
GREEN		OFF
GREEN		OFF
BLUE		ON
BLUE		ON
BLUE		OFF
GREEN		ON
GREEN		OFF
GREEN		OFF
GREEN		ON

the other elaborate narrative or with the Summary narrative, randomized across subjects.

**Null datasets:** We always presented one of the two simple narratives in step 4. In step 5, we always presented the Chain narrative on its own.

Given three policy choices per dataset and three datasets, subjects made a total of nine incentivized policy choices in the BASELINE treatment. We paid one randomly selected choice.

Our second treatment, the SALIENCE treatment, is virtually identical to the BASELINE treatment except that we replaced the Noise dataset with the Masked dataset (displayed in Figure 2). In the SALIENCE treatment, we also split the instructions across several screens and reminded subjects of the cost function on each decision screen.

Our third treatment, the BELIEFS treatment, is identical to the SALIENCE treatment except that after presenting a single narrative, we elicited subjects’ beliefs that the causal relationship suggested by the narrative is true. For example, for the Chain narrative, we stated *“This advice suggests that the Choice (BLUE or GREEN) has an effect on the Z variable, which in turn has an effect on the Y variable. How likely do you think it is that this relationship is true?”*. Further, after presenting two competing narratives, we elicited subjects’ beliefs about each of the two possible causal relationships. We elicited subjects’ beliefs prior to their policy decisions, and told subjects that one of their belief responses or one of their policy choices would be



randomly selected for payment, with equal probability. We incentivized beliefs using the binary scoring rule as implemented by Danz, Vesterlund, and Wilson (2022), with a maximum payoff of \$1.00.<sup>12</sup>

### 3.1.1 Understanding the Design

We designed the experiment to achieve several goals.

First, we purposefully framed the datasets as neutrally as possible to reduce the chances that subjects import prior beliefs about the DGP into the experiment.<sup>13</sup> In the BASELINE treatment, we labeled the variables ‘Choice’, ‘X’, and ‘Payoff’ to try avoid any preconceived notions between the variables. In the SALIENCE and BELIEFS treatments we went further, labeling the auxiliary variable ‘Z’ and the outcome variable ‘Y’. This relabeling addresses the potential concern that subjects in BASELINE may not believe that the ‘Payoff’ variable can affect another variable (even though it does so in the true DGP). Further, in all of our treatments, the values of all variables are labeled using neutral colors and symbols.

Second, to avoid deception, we constructed narratives that point out patterns in the data, rather than explicitly stating a causal model. The narratives also give policy recommendations, as many narratives do in practice. If the narratives do change beliefs, it implies that subjects both (i) form a causal model from the narrative and (ii) use that causal model to update their beliefs. One purpose of the BELIEFS treatment is to separate this joint hypothesis by providing direct evidence for (i).

Third, we took seriously the possibility that subjects may blindly respond to narratives independently of the information contained in the dataset, either because they are not paying attention or because they are trying to do what the experimentalist desires (i.e., a demand effect). To identify and exclude such subjects in robustness tests, we presented each subject with the Chain narrative alongside the Null dataset. Because the pattern highlighted by the narrative does not exist in this dataset, the narrative is *inconsistent* (Pennington and Hastie (1993)), and should only influence

---

<sup>12</sup>Because of the additional elicitations in the BELIEFS treatment, we dropped the questions about certainty and narrative helpfulness. We did so in the SALIENCE treatment as well to make it more comparable to the BELIEFS treatment.

<sup>13</sup>In a previous experiment (documented in Appendix C), we used a less neutral frame, labeling the action, ‘Manager Action’, the outcome, ‘Firm Profits’, and the auxiliary variable, ‘Employee Action’. The results are very similar.

choices of inattentive subjects or those responding to demand effects.

Fourth, we use both Independent and Causal datasets to test the efficacy of narratives under two extremes: when the action has no causal effect on the outcome and when it has a strong causal effect. The Causal datasets allow us to test whether (and which types of) narratives work when they oppose a true causal relationship in the data. The comparison across Independent and Causal datasets also allows us to test for illusion of control (Langer (1975); Fan (2024)). Narratives may work in Independent datasets because they give subjects false hope that they can control the outcome, even though policies have no effect on the outcome. In Causal datasets, policy choices do control the outcome, placing all narratives – including the true Summary narrative – on a level playing field.

Fifth, we use the Noise datasets to test whether narratives are robust to weaker correlations (more noise) in the data. Crucially, the Noise dataset also provides the most direct comparison between the Chain and Collider narratives, as the BNFF predicts similarly sized belief distortions under both, unlike in the Strong dataset.

Sixth, we directly pit narratives against one another for two reasons. First, to examine what drives subjects to adopt one narrative over another. One hypothesis, proposed by Eliaz and Spiegler (2020), is that when narratives compete, individuals gravitate toward the more ‘hopeful’ one – that is, the narrative that provides the highest expected utility given the subjective beliefs it induces. Anticipatory utilities for each dataset–narrative combination used in our experiment are reported in Table A2 of Appendix A. Second, competing narratives offer an additional test of the hypothesis that subjects blindly follow elaborate or simple narratives. When one of these narratives competes with the Summary narrative, there is no reason to believe subjects should favor one or the other.

Seventh, we compare simple and elaborate narratives to test for what cognitive scientists call *coverage* (Pennington and Hastie (1993)). While both types of narratives provide the same recommendation, elaborate narratives provide coverage by offering an explanation for the relationships between all of the variables in the dataset, while simple narratives do not.

Eighth, we randomized the row order of each dataset across subjects to ensure that our results are not driven by some idiosyncrasy of a particular ordering. We also

randomized the column order of the auxiliary and outcome variables across subjects to test whether, for example, the Chain narrative is more likely to be adopted when the data is presented in the same order as the implied causal chain ( $a \rightarrow z \rightarrow y$ ). As we show in Appendix B.2, column ordering has little effect, so we pool the data in the analyses that follow.

Lastly, the SALIENCE and BELIEFS treatments serve two purposes. First, they replace the Noise dataset with the Masked dataset to test a hypothesis we discuss in Section 3.2.2. Second, by including the Strong datasets, they test the robustness of our BASELINE treatment findings to variations in variable labeling, narrative wording, and changes in the instructions.<sup>14</sup>

### 3.1.2 Implementation

We conducted both arms of the BASELINE treatment in April and May of 2023. The SALIENCE and BELIEFS treatments followed in November and December of 2024.<sup>15</sup> All three treatments were conducted online with a sample of the U.S. population, balanced by gender, using Prolific. Each session began with detailed instructions (replicated in the Supplementary Material along with the decision screens), followed by a set of comprehension questions that subjects were required to answer correctly before proceeding. We recruited approximately 500 subjects in each arm of the BASELINE treatment, and in each of the other two treatments. Subjects in the BASELINE treatments earned an average of \$3.32 for about 13.5 minutes of participation (\$14.76/hour). In the SALIENCE treatment, average earnings were \$3.31 for 16.0 minutes (\$12.41/hour), and in the BELIEFS treatment, \$4.50 for 24.6 minutes (\$10.97/hour).

In March of 2025, we also conducted 14 sessions (for a total of 103 subjects; 76% female) of the SALIENCE treatment in the LSE experimental laboratory. Subjects

---

<sup>14</sup>In BASELINE, the simple narratives were phrased as “Choose BLUE more often” and “Choose GREEN more often”. One concern with this wording is that it leaves the benchmark for “more often” ambiguous. To address this, the SALIENCE and BELIEFS treatments used slightly revised versions: “Choose BLUE more often than GREEN” and “Choose GREEN more often than BLUE”.

<sup>15</sup>To view the treatments directly, visit BASELINE, SALIENCE, and BELIEFS. A software bug resulted in incorrect initial bonuses for the BASELINE treatment. When we discovered the bug, we immediately corrected the issue by paying additional bonus payments (average of \$0.05) in June of 2023. Importantly, the bug did not affect the data collected because the bonus was only reported to subjects at the end of the experiment.

were given printed copies of the instructions and they were read aloud. This lab sample is academically very sophisticated: all subjects were university students and 68% were studying at the graduate level. Subjects earned an average of £13.19 (about \$17.41 USD) in the form of Amazon vouchers for 45 minute sessions (£17.59/\$23.22 per hour). To increase statistical power for the elaborate narratives, we omitted the simple narratives in these lab sessions due to the smaller sample size.

## 3.2 Results

We first show results for the Strong and Causal Strong datasets, establishing that causal narratives, particularly Chain narratives, have robust effects on policy choices in both datasets. We then explore why the Collider narrative appears to be less effective, using the Noise and Masked datasets to demonstrate that many subjects independently infer the opposing Chain relationship. Finally, we examine policy choices under competing narratives and find that subjects tend to place weight on multiple causal models simultaneously. Whenever available, we present the lab results alongside the online results. All statistical results refer to the larger online sample, but they are very similar in size for the lab sample.

### 3.2.1 Strong and Causal Strong Datasets

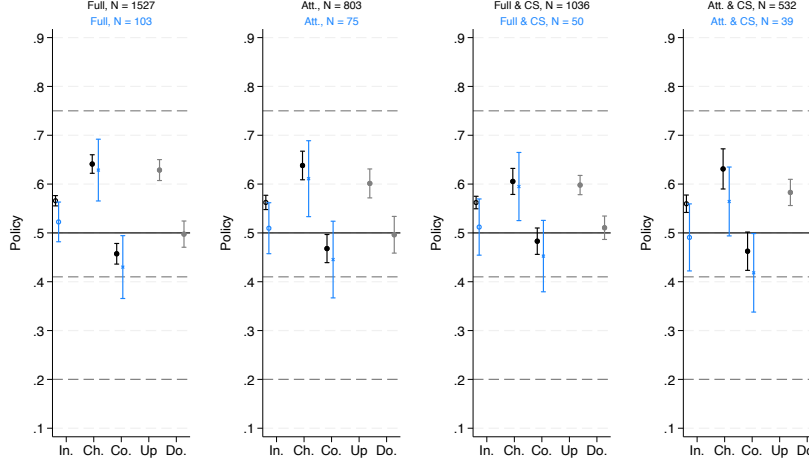
We begin with the Strong datasets in which the action and outcome are independent. Figure 3 plots average initial policies, as well as average policies following exposure to each of the elaborate and simple narratives, across four subsamples of subjects. These averages reflect policy choices across all subjects in a subsample, regardless of when they observed the Strong dataset (first, second, or third).<sup>16</sup> We pool results from the BASELINE, SALIENCE, and BELIEFS treatments because we found no systematic differences across treatments (indicating that changes to the instructions, variable labeling, and narrative wording made little difference). Figure A3 in Appendix A replicates Figure 3 separately for each treatment.

The first key takeaway from Figure 3 is that the lab and online samples provide

---

<sup>16</sup>The effects of narratives are slightly larger when restricting the sample to subjects who saw the Strong dataset first, though standard errors are also larger due to reduced statistical power (see Figure A2 in Appendix A).

Figure 3. Policies in the Strong Dataset



Notes: Average initial policy choices (In.), and those after Chain (Ch.), Collider (Co.), Up, and Down (Do.) narratives. Error bars represent 95 percent confidence intervals. We show results for both the online sample and lab sample (in blue). The first panel includes all subjects in each sample. The second panel restricts to attentive subjects that do not respond to inconsistent narratives. The third panel is for policy choices when the narrative competed directly with the Summary narrative. The fourth panel is for policy choices of attentive subjects when the narrative competed directly with the Summary narrative. The solid line indicates the rational prediction while the dashed dark gray lines indicate the predictions of the BNFF for the Chain narrative (upper) and the Collider narrative (lower two lines indicate the predicted range). Data is pooled from BASELINE, SALIENCE, and BELIEFS.

very similar results, establishing that the effects of narratives are robust across different samples. Focusing first on the elaborate narratives in the full sample (first panel), in both the lab and online, we see that the Chain and Collider narratives result in different policy choices: the difference between the average policy choices under each narrative is highly significant (0.18,  $p < 0.001$ , two-sample t-test). Both narratives move policies in the directions predicted by the behavioral theory and, since neither confidence interval contains 0.5, we can reject the predictions of rational theory. When we compare policies to the BNFF predictions, indicated by the dashed dark gray horizontal lines, we see that they undershoot the prediction on average. In Appendix B.1, we show that this undershooting is consistent with cognitive uncertainty (Enke and Graeber (2023)): subjects who report greater certainty in their policy choices tend to deviate *more* from the rational policy.

One immediate concern is that these effects might be driven by subjects blindly

following elaborate narratives, either because of inattention or due to an experimenter demand effect. Results in the second through fourth panels, however, show that this is not the case. In the second panel, we restrict the sample to *attentive* subjects – those who do not follow the inconsistent Chain narrative when presented alongside the Null dataset, where the specified pattern does not exist. Specifically, we exclude any subject who adjusted their initial policy in the direction of the inconsistent narrative by any amount. We adopt this very strict criterion to eliminate any possibility of demand effects, but the results are virtually identical if we adopt a weaker criterion.<sup>17</sup> Under this strict criterion, 47% of subjects in the online sample are classified as inattentive, compared to 27% in the lab sample. As expected, the difference indicates that laboratory subjects tend to be more attentive.

In the third panel, we examine policy choices made when subjects were shown both an elaborate narrative and the Summary narrative that explicitly recommends choosing 0.5. These results further alleviate concerns that subjects are blindly following narratives, as there is no clear reason to favor one narrative over the other, particularly given that we randomized the order in which the recommendations were presented. This test also helps rule out confirmation bias: if subjects were selectively seeking support for a favored narrative, either narrative could serve that purpose equally well. Finally, in the fourth panel of Figure 3, we look at the choices of only attentive subjects when they see both an elaborate narrative and the Summary narrative. This test is a fairly extreme robustness check in that it rules out inattention and demand effects via two methods simultaneously. Overall, we see very similar results across all four panels of Figure 3.

**Result 1:** *The Chain and Collider narratives result in different policy choices for the same (Strong) dataset. Both narratives distort policies away from the rational policy as predicted by the behavioral theory, though not always significantly so for Collider narratives. These distortions cannot be driven by inattention, demand effects,*

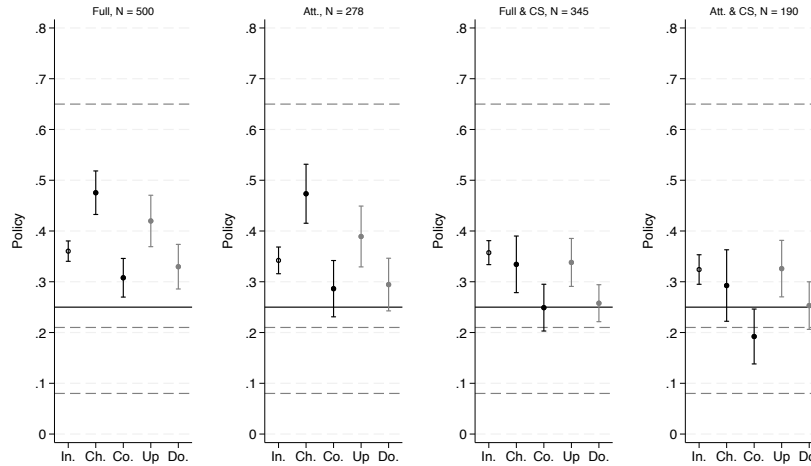
---

<sup>17</sup>As an alternative to excluding subjects who respond to the inconsistent narrative, we can maintain the full sample and compare average policy choices under the Chain narrative for Strong datasets versus Null datasets. This comparison provides a direct test of whether subjects are simply following the narrative regardless of the underlying data. We implement this test by regressing average policy choices when subjects observed the Chain narrative on a dummy indicating the Strong dataset with the Null dataset as the omitted category, clustering standard errors at the subject level. The difference is highly significant (0.07,  $p < 0.001$ ), indicating that Chain narratives are more effective when they can leverage correlations in the data.

or confirmation bias, and are very similar among the online and lab samples (though with larger standard errors in the lab sample).

One potential explanation for these results is that subjects are inclined to believe the false narratives because these narratives suggest subjects' actions influence outcomes - a form of illusion of control. However, we find similar narrative effects in the Causal Strong datasets (available for the BASELINE treatment only), where choosing a policy below 0.5 *does* in fact increase the probability of a good outcome. Figure 4 shows the average initial policies and the average policies after each narrative, following the same structure as Figure 3.

Figure 4. Policies in the Causal Strong Dataset



Notes: Average initial policy choices (In.), and those after Chain (Ch.), Collider (Co.), Up, and Down (Do.) narratives. Error bars represent 95 percent confidence intervals. The first panel is for all subjects. The second panel restricts to attentive subjects that do not respond to inconsistent narratives. The third panel is for policy choices when the narrative competed directly with the Summary narrative. The fourth panel is for policy choices of attentive subjects when the narrative competed directly with the Summary narrative. The solid line indicates the rational prediction while the dashed dark gray lines indicate the predictions of the BNFF for the Chain narrative (upper) and the Collider narrative (lower two lines indicate the predicted range). Data is from BASELINE.

In Figure 4, we see that the Chain narrative is effective even when it contradicts the true causal relationship. Average policies under the Chain narrative are far from the rational policy prediction of 0.25, and in every panel except the fourth – which focuses on attentive subjects (56% of the sample) who also received the Summary narrative – we can reject the null hypothesis of no deviation from the rational policy. In fact,

when we analyze subject-level data, we find that many subjects choose policies above 0.5, policies that reduce the likelihood of a good outcome. Moreover, across all panels, average policy choices under the Chain and Collider narratives differ significantly, with differences ranging from 0.085 to 0.187 ( $p < 0.025$  in all cases, two-sample t-tests).

**Result 2:** *The Chain and Collider narratives result in different policy choices for the same (Causal Strong) dataset. The Chain narrative has a significant effect even when it opposes the true causal relationship.*

### 3.2.2 Implicit Models

Although both elaborate narratives shift policies in the directions predicted by the behavioral theory, the Chain narrative consistently moves choices farther from the rational benchmark than the Collider narrative, in both the online and lab samples. To investigate why, we focus on the Strong datasets, where the symmetry around 0.5 provides a clean setting for comparison.

One possible explanation is that the behavioral theory allows for small deviations from the rational policy for Collider narratives, given the range of predicted policies. However, this explanation cannot fully account for the pattern we observe. In the Noise dataset, the BNFF predicts a slightly larger deviation for the Collider narrative, yet we still see that the Chain narrative produces larger deviations (see Figure A4 in Appendix A).<sup>18</sup>

The key to understanding the larger apparent effectiveness of Chain narratives lies in the heterogeneity of subjects' *initial* policy choices. Figure 5 splits the full sample of subjects into three groups based on initial choices: those who lean down (policy  $< 0.48$ ), those who choose central policies (0.48-0.52), and those who lean up (policy  $> 0.52$ ). For each group, we plot average initial policies and average policies after each narrative, following the format of Figure 3. We also indicate the size of each group in the panel headings. Figure A6 of Appendix A replicates this figure for Attentive subjects, showing very similar results.

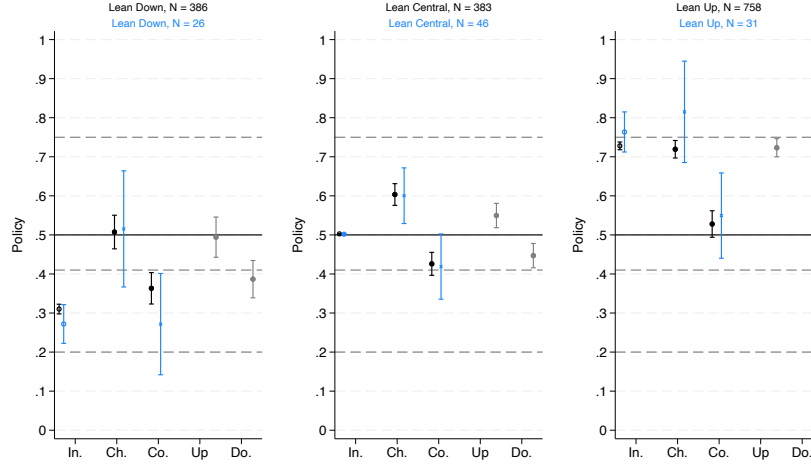
The first important takeaway from Figure 5 is the substantial heterogeneity in initial policy choices, before subjects are exposed to any narrative. Only 25% of

---

<sup>18</sup>More broadly, the results for the Noise and Causal Noise datasets closely resemble those from the Strong and Causal Strong datasets (see Figures A4 and A5 in Appendix A), suggesting that deterministic patterns in the data are not necessary for the Chain (and Up) narratives to be effective.



Figure 5. Heterogeneity in the Strong Dataset



Notes: Average initial policy choices (In.), and those after Chain (Ch.), Collider (Co.), Up, and Down (Do.) narratives. Error bars represent 95 percent confidence intervals. We show results for both the online sample and lab sample (in blue). The first panel is for subjects with initial policies less than 0.48. The second panel is for subjects with initial policies between 0.48 and 0.52. The third panel is for subjects with initial policies greater than 0.52. The solid line indicates the rational prediction while the dashed dark gray lines indicate the predictions of the BNFF for the Chain narrative (upper) and the Collider narrative (lower two lines indicate the predicted range). Data is pooled from BASELINE, SALIENCE, and BELIEFS.

subjects choose policies near the rational benchmark of 0.5; the rest lean toward higher or lower policies. Crucially, these initial choices strongly predict how subjects respond to narratives. Those who initially choose central policies tend to respond nearly symmetrically to the Chain and Collider narratives ( $-0.08$  for the Collider versus  $0.10$  for the Chain). In contrast, subjects who lean up or down respond little when the narrative aligns with their initial inclination, but shift their policies by similar amounts when the narrative runs counter to it ( $-0.21$  for the Collider versus  $0.21$  for the Chain). Thus, in terms of *movements* in policies, Chain and Collider narratives are similarly effective.

The second key takeaway from Figure 5 helps explain why final policy choices differ across the two narratives: nearly twice as many subjects lean up as lean down. This asymmetry is unlikely to arise from random variation or from subjects miscounting how often each outcome occurs for each action. Instead, we hypothesize that subjects construct their own mental models of the data-generating process (DGP) based on the dataset. In this case, any externally provided narrative must compete with a

subject’s homegrown model. If a larger share of subjects naturally infer the causal Chain structure, rather than the Collider structure, it would explain why the Chain narrative appears more effective in the aggregate.

In Section 4.2, we directly elicit subjects’ mental models and confirm that they are more likely to infer the causal Chain relationship than the Collider relationship when presented with the Strong dataset. Here, we provide preliminary evidence consistent with this hypothesis, leveraging differences across the Strong and Null datasets. First, Figure 3 shows that initial policy choices tend to lean up on average. If this upward bias reflected subjects inferring the Chain relationship from the data, then we should observe a difference between the Strong dataset, which contains correlations, and the Null dataset, which does not. This is indeed what we find: initial policies deviate more from the rational benchmark in the Strong dataset than in the Null dataset (0.57 vs. 0.53,  $p < 0.001$ ).

Second, the hypothesis can also explain why the Chain and Up narratives result in similar final policies, even though only the Chain narrative provides coverage – that is, an explanation of how all three variables are related in the dataset. If subjects independently infer the Chain structure, or are nudged toward it by the Up narrative, we would expect final policy choices under the Up narrative to deviate more from the rational policy in the Strong dataset relative to the Null dataset. To see if this is the case, we regress average policy choices following an Up narrative on a dummy for the Strong dataset, using the Null dataset as the omitted category and clustering standard errors at the subject level. We find that policy choices are significantly higher in the Strong dataset (difference = 0.05,  $p < 0.001$ ).<sup>19</sup>

Taken together, these findings suggest that the Chain narrative appears more effective, not because it is inherently more persuasive, but because it resonates with the mental model that many subjects infer from the data.<sup>20</sup> If we can prevent subjects

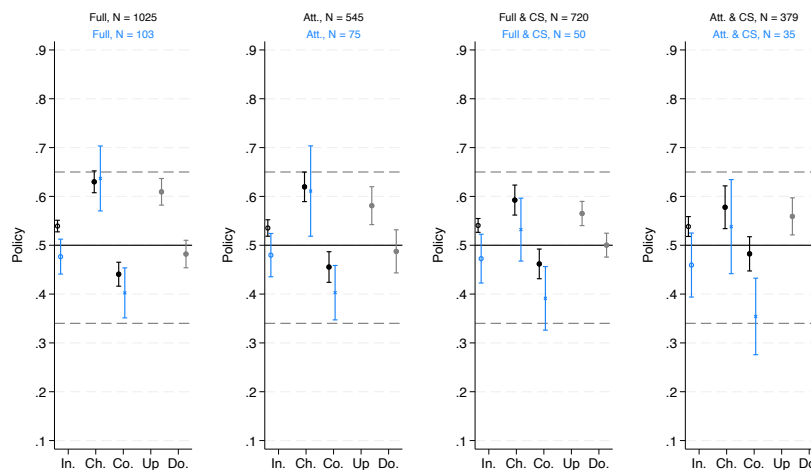
---

<sup>19</sup>One potential concern is that this result may simply reflect spillover effects from subjects having previously seen the Chain narrative in another dataset. However, the effect is even stronger when we restrict the sample to subjects who encountered the Strong dataset first (0.10,  $p < 0.001$ ).

<sup>20</sup>An alternative explanation is that subjects simply anchor on their initial policy, even if they do not uncover a causal model on their own. However, in Figure A7 in the Appendix, we use data from the BELIEFS treatment to show that initial policies predict subjects’ beliefs in the causal models implied by the different narratives – a finding we would not expect under the anchoring explanation. Furthermore, the results from the ELICIT treatment, which we discuss in Section 4.2, directly show that subjects construct their own models from the data.

from forming their own causal models, it would potentially place the Chain and Collider narratives on more equal footing. This is the purpose of the Masked dataset (Figure 2), introduced in Section 3.1. It masks the correlations in the data, making it harder for subjects to infer a causal structure on their own, while preserving the correlations, so that narratives can still point to them to potentially influence beliefs. Figures 6 and 7 present the results for these datasets (pooled across the SALIENCE and BELIEFS treatments).

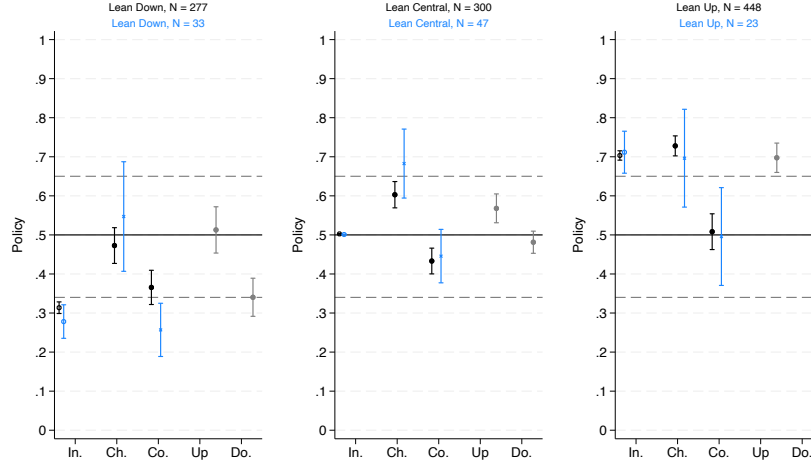
Figure 6. Policies in the Masked Dataset



Notes: Average initial policy choices (In.), and those after Chain (Ch.), Collider (Co.), Up, and Down (Do.) narratives. Error bars represent 95 percent confidence intervals. We show results for both the online sample and lab sample (in blue). The first panel includes all subjects in each sample. The second panel restricts to attentive subjects that do not respond to inconsistent narratives. The third panel is for policy choices when the narrative competed directly with the Summary narrative. The fourth panel is for policy choices of attentive subjects when the narrative competed directly with the Summary narrative. The solid line indicates the rational prediction while the dashed dark gray lines indicate the predictions of the BNFF for the Chain narrative (upper) and the Collider narrative (lower). Data is pooled from SALIENCE and BELIEFS.

We find some evidence that obscuring correlations dampens subjects' tendency to form homegrown narratives. First, initial policies deviate less from 0.5 in Figure 6 than in Figure 3 (difference = 0.027, both in the full sample ( $p < 0.001$ ) and for attentive subjects ( $p = 0.007$ )). Second, a smaller share of subjects lean up in Figure 7 compared to Figure 5 (44% vs 50%;  $p = 0.001$ ). Nevertheless, despite our attempt to mask the correlations, some online subjects still appear to infer a causal Chain relationship, even though doing so requires considerable effort to detect the

Figure 7. Heterogeneity in the Masked Dataset



Notes: Average initial policy choices (In.), and those after Chain (Ch.), Collider (Co.), Up, and Down (Do.) narratives. Error bars represent 95 percent confidence intervals. We show results for both the online sample and lab sample (in blue). The first panel is for subjects with initial policies less than 0.48. The second panel is for subjects with initial policies between 0.48 and 0.52. The third panel is for subjects with initial policies greater than 0.52. The solid line indicates the rational prediction while the dashed dark gray lines indicate the predictions of the BNFF for the Chain narrative (upper) and the Collider narrative (lower). Data is pooled from SALIENCE and BELIEFS.

correlations.<sup>21</sup> Because of this, the Collider narrative has only a slightly stronger, and not statistically different, effect on final policy choices in the Masked dataset compared to the Strong dataset ( $p = 0.294$ ).

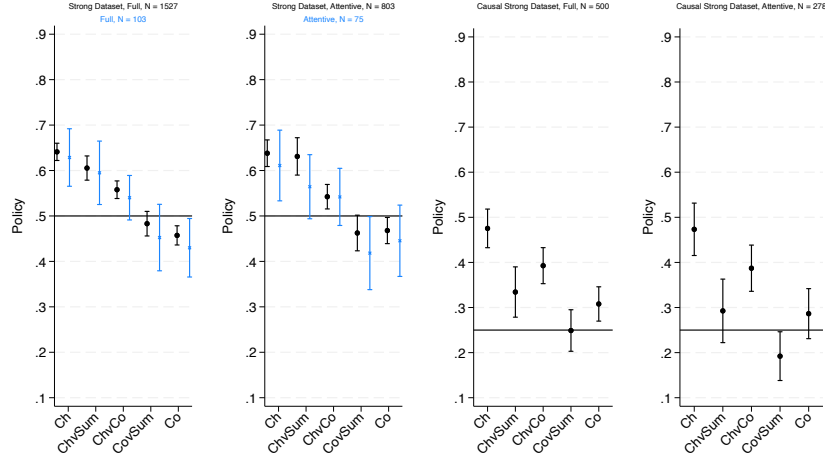
**Result 3:** *The Chain and Collider narratives are equally effective at moving policies. However, because subjects tend to construct implicit causal Chain models from correlations in the data: (i) the Chain narrative appears more effective than the Collider narrative in the aggregate, and (ii) the Up narrative – despite lacking coverage – appears nearly as effective as the Chain narrative in the aggregate.*

### 3.2.3 Competing Narratives and Beliefs

The previous results suggest that policy choices reflect both the externally provided narrative as well as subjects' homegrown models – implying that narratives face im-

<sup>21</sup>Average initial policies remain significantly different from 0.5 even among those who encountered the Masked dataset first, ruling out the possibility that these patterns are driven by spillovers from earlier exposure to the Strong dataset.

Figure 8. Policies after Single and Competing Narratives



Notes: Average policy choices and 95 percent confidence intervals. We show results for both the online sample and lab sample (in blue), where available. ChvSum indicates the average policy choice when subjects observed both the Chain narrative and the Summary narrative. CovSum indicates the Collider narrative and the Summary narrative. ChvCo indicates both the Chain and the Collider narrative. The left pair of graphs is for the Strong dataset and the right pair of graphs is for the Causal Strong dataset. In each, we plot the average for the full sample of subjects and for the subset of attentive subjects that do not respond to inconsistent narratives. Data is pooled from BASELINE, SALIENCE, and BELIEFS for the left pair of graphs, and from BASELINE for the right pair of graphs.

plicit competition. In this section, we examine beliefs and policy choices when subjects are presented with two explicitly competing narratives, providing direct evidence that subjects do, in fact, place weight on multiple models simultaneously.

Beginning with policy choices, Figure 8 plots average choices in the Strong and Causal Strong datasets under several conditions: when subjects see only the Chain or Collider narrative, when each is paired with the Summary narrative (ChvSum and CovSum), and when the Chain and Collider narratives are presented together (ChvCo). For the Strong dataset, we continue to pool data across the BASELINE, SALIENCE, and BELIEFS treatments.

Consistent with the results of the previous section, Figure 8 shows that subjects do not appear to adopt one narrative over the other. Instead, when presented with competing narratives, they tend to choose policies that lie between those chosen under each narrative individually. For both datasets, we can reject the null hypothesis that average policy choices under joint exposure to Chain and Collider narratives are equal

to those under either narrative alone. Similarly, when subjects see both the Chain and the Summary narrative, average policy choices fall between those under the Chain narrative alone and the rational policy (though among attentive subjects we can't reject the null that they overlap with one or the other).<sup>22</sup> The only case in which the policies under competing narratives do not lie between the two competing predictions is when the Collider narrative competes with the Summary narrative in the Causal Strong datasets, which is likely a consequence of the two narratives reinforcing each other, amplifying the overall effect.

While the above results strongly suggest that subjects weight both models, the BELIEFS treatment allows us to demonstrate this explicitly. In particular, for each narrative – whether presented alone or in competition – we elicit subjects' beliefs that the model implied by the narrative is the true DGP. This approach allows us to (i) disentangle the joint hypothesis that subjects infer a causal model from the narrative and then choose policies according to the model, and (ii) test whether subjects entertain two models simultaneously.

Figure 9 presents binscatters of subjects' policy choices against the relative belief they assign to each of the two competing narratives.<sup>23</sup> The upper three panels are for the full sample and the lower three for attentive subjects only.

The plots in Figure 9 reveal two key findings. First, while subjects' beliefs about the likelihood of each model are highly heterogeneous, most belief shares lie away from the extremes: only 8% of subject-narrative pairs assign relative beliefs of exactly 0 or 1. This result implies that the average policy choices observed in Figure 8 are not driven by some subjects fully adopting one model and others fully adopting the other; rather, most subjects appear to weight both. Second, policy choices are strongly correlated with beliefs.<sup>24</sup> The more weight a subject places on the Chain narrative (relative to either the Collider or Summary narrative), the higher their policy choice. Likewise, the more weight on the Collider narrative (relative to the Summary narrative), the lower the policy choice, though this effect is weaker and

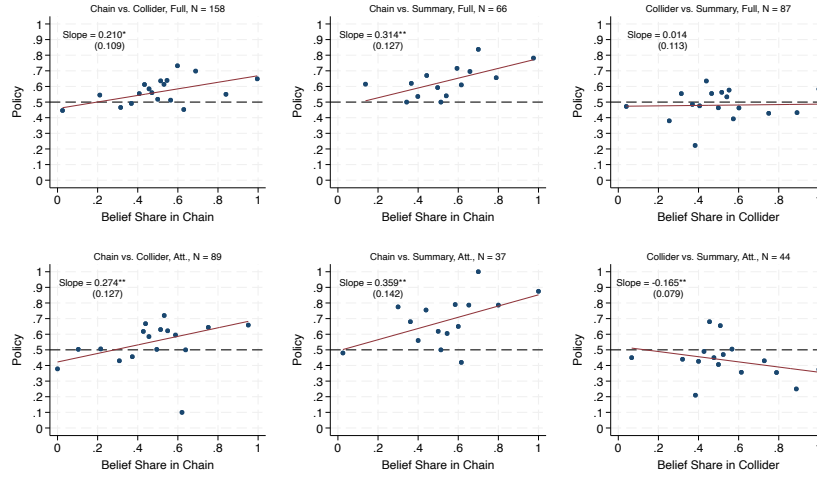
---

<sup>22</sup>In Appendix B.2, we show that this result is not driven by subjects anchoring on the Chain narrative because they see it first.

<sup>23</sup>We did not constrain the two beliefs to add up to a maximum of 100%. 40% of subject-narrative pair beliefs violated this consistency requirement. The results are similar if we drop these observations.

<sup>24</sup>In Figure A8, we show that subjects' beliefs after receiving a single narrative also strongly predict their policy choices.

Figure 9. Policies vs. Beliefs after Competing Narratives



Notes: Binscatters of policy choices versus relative belief share of the narrative indicated. The upper three panels are for the full sample and the lower three for attentive subjects only. Slopes and standard errors are shown for each panel. \* $p < .1$ ; \*\* $p < .05$ ; \*\*\* $p < .01$ . Data is from BELIEFS.

statistically significant only among attentive subjects.

**Result 4:** *Subjects assign positive weight to the causal models implied by narratives. When confronted with competing narratives, they generally believe both narratives are plausible and make policy choices that reflect the relative weight they place on each.*

## 4 Creating and Transmitting Narratives

The previous results strongly suggest that the impacts of narratives are shaped by the heterogeneous mental models that subjects construct from the data. In the following pair of treatments (ELICIT and NATURAL), we make this heterogeneity explicit by asking subjects to generate their own narratives. We then incentivize subjects to transmit these narratives to other subjects, allowing us to test whether self-generated narratives influence beliefs and choices in ways comparable to the structured narratives we provided in the previous treatments.

## 4.1 Experimental Design – ELICIT and NATURAL

The first three steps of the ELICIT treatment are identical to those of the BASELINE treatment: subjects observed a dataset, chose a policy, and stated their certainty. After completing these steps, we provided a free-form text box and asked subjects to give specific advice to future subjects. We endowed each subject with \$1.00, which they could use to bid in a first-price auction against a group of nineteen other subjects. We broke ties randomly and showed the winner’s advice to an average of 40 future subjects. We paid the winner \$0.025 for each future subject who rated the advice as helpful (rather than unhelpful).<sup>25</sup> We also informed subjects that if their advice was not specific, meaning it didn’t explicitly or implicitly imply a policy choice, it would be excluded from the auction. Subjects completed these tasks for four datasets presented in random order: Strong, Null, Causal Strong, and Causal Null. For payment, we randomly selected one of the four policy choices and one of the four auctions.

The NATURAL treatment is virtually identical to the BASELINE treatment except that we replaced our constructed narratives with those generated by subjects who won the auctions in the ELICIT treatment. We informed subjects in NATURAL that the narratives they saw came from other subjects who had viewed the same dataset and had won the auction. For the Strong and Causal Strong datasets, we always paired the subject-generated advice with the Summary narrative before asking for a third policy choice. For the Null and Causal Null datasets, we instead presented only the Chain narrative we had constructed.

### 4.1.1 Implementation

We ran the ELICIT and NATURAL treatments online in May of 2023.<sup>26</sup> We recruited a sample of the U.S. population, balanced by gender, using Prolific. Each session began with detailed instructions (replicated in the Supplementary Material along with the decision screens), followed by a set of comprehension questions that subjects were required to answer correctly before proceeding. We recruited 201 subjects in the ELICIT treatment, who earned an average of \$4.31 for about 17.3 minutes of

---

<sup>25</sup>Being paid for helpfulness is akin to garnering ‘likes’ on social media and means that the subjects providing advice had no explicit incentive to manipulate the beliefs of future subjects.

<sup>26</sup>To view the experiments directly, visit ELICIT and NATURAL.



Table 1. Elicited Narratives

Classification	Strong	Null	Causal Strong	Causal Null
Up	11.5	18	5.5	3.5
Chain	14.5	1.5	7	0.5
Down	8.5	5	2.5	6
Collider	2	0	3.5	0
Rational	37	48	54.5	55.5
Other	0	0.5	5.5	10.5
Multiple	1	0	2	0
Reject	25.5	27	19.5	24

Notes: Classification of elicited narratives (percentages) in each dataset. ‘Multiple’ indicates advice that described both a Chain narrative and Collider narrative or a Chain narrative and rational advice. ‘Other’ consists mainly of advice that says the process is random or to choose 0.5 in the causal datasets. Data is from ELICIT.

participation (\$14.94/hour). In NATURAL, 401 subjects earned an average of \$3.25 for 16.0 minutes (\$12.16/hour).

## 4.2 Results – ELICIT

We categorized the elicited advice in the following way: each co-author independently assessed whether a piece of advice contained an explicit or implicit recommendation, and if so, classified it into one of several categories. We provide details of this classification in Appendix D. In Table 1, we summarize the resulting distribution of narrative types.

Overall, 76% of subjects followed our instructions and provided explicit advice. In the Strong dataset (first column), the most common advice is rational advice that argues for a policy of 0.5 (e.g., “*each color is listed 8 times. There is an equal amount of high and low in each color. chances are 50%*”). However, nearly as many subjects offered some form of causal advice, with most recommending higher policies, consistent with the Up or Chain narratives. Indeed, much of the causal advice explicitly pointed out the Chain narrative (e.g., “*The triangle in this trial alway (sic) had a High payoff. And the only time a triangle appeared was with the Blue choice, not the green. therefore, selecting Blue would maximize the chance of a triangle and therefore of a high payoff.*”). By contrast, only four subjects identified the Collider narrative on their own.<sup>27</sup> This asymmetry supports our hypothesis that more subjects implicitly

<sup>27</sup>We document an almost identical result in an earlier version of the same experiment in Appendix C.

inferred an erroneous causal Chain model than a causal Collider model in our other treatments. Overall, the distribution of advice in Table 1 dovetails nicely with the heterogeneity in initial policy choices observed in these treatments (see Figure 5).

The second column of Table 1 shows the distribution of elicited narratives in the Null dataset, where all three variables are statistically independent. Here, rational advice is more common, and elaborate narratives largely disappear; those that remain mostly come from subjects who previously encountered the Strong dataset. The near absence of elaborate narratives in the Null dataset suggests that correlations between  $z$  and  $a$ , and between  $z$  and  $y$ , are necessary for such narratives to emerge.

The third and fourth columns of Table 1 show the breakdown for the Causal Strong and Causal Null datasets, respectively. As with the Independent datasets, rational advice is most common in both cases. However, we still observe Chain narratives in the Causal Strong dataset. This is a particularly striking finding, as the Chain narrative recommends a higher policy, directly contradicting the true causal relationship in the data which favors a lower policy.<sup>28</sup>

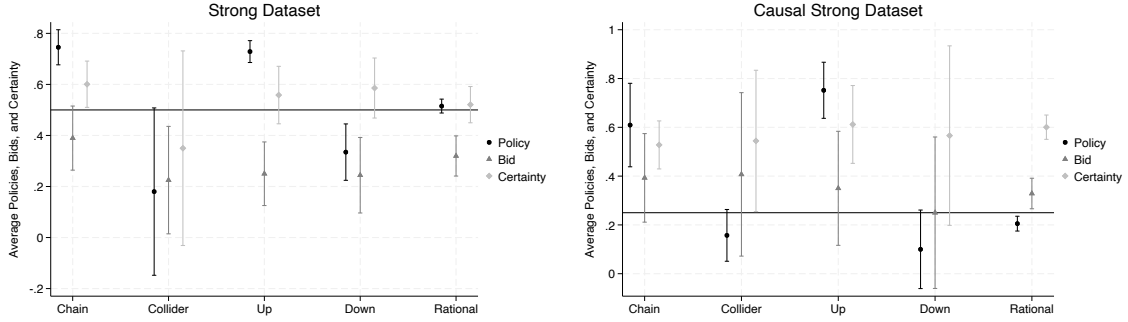
**Result 5:** *Subjects generate elaborate causal narratives after merely observing a dataset containing auxiliary variables, but almost exclusively when the data contain correlations that support such narratives. Subjects are much more likely to produce a Chain narrative than a Collider narrative, and do so even when the Chain narrative directly contradicts the true causal relationship in the data.*

We designed the ELICIT treatment such that it is in subjects’ best interest to provide advice that *appears* helpful, rather than necessarily being helpful. However, because we also elicit subjects’ policy choices, we can evaluate whether they believe their own advice. Figure 10 shows average policies for the Strong and Causal Strong datasets, broken down by narrative type (excluding the categories Multiple and Other for clarity). We find that subjects largely follow their own advice: those who offer rational advice choose policies near the rational benchmark, while those who give

---

<sup>28</sup>Subjects do not produce Chain narratives only after first encountering the Strong dataset; the proportion of Chain narratives is roughly the same when subjects see the Causal Strong dataset before the Strong dataset. Further, subjects that construct an elaborate narrative in either of the Strong or Causal Strong datasets spend slightly longer on the experiment overall than those that do not (19.3 vs. 16.8 minutes on average). The difference is not significant ( $p = 0.200$ , two-sample t-test), but suggests that subjects that construct elaborate narratives are paying at least as much attention as other subjects.

Figure 10. Policies and Bids in ELICIT



Notes: Average policies, bids, and certainty by narrative type. The error bars indicate 95 percent confidence intervals. The left graph is for the Strong dataset, and the right for the Causal Strong dataset. Solid lines indicate rational policies. Data is from ELICIT.

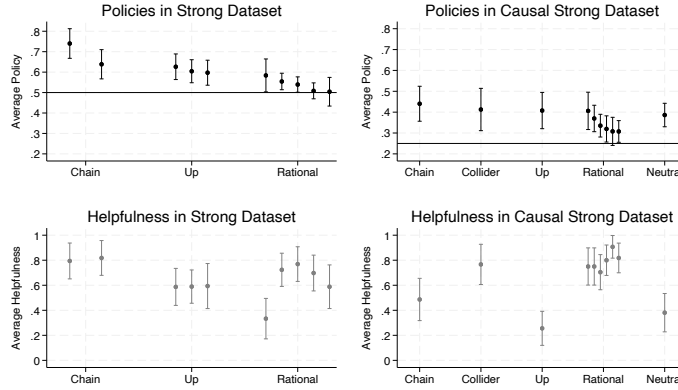
causal advice tend to choose policies that deviate in the direction suggested by their narrative.<sup>29</sup>

Figure 10 also shows subjects' average bids and certainty in their policy choices. In the Strong dataset, those who produce a Chain narrative are more certain on average, and bid slightly more than those who produce rational advice, providing additional evidence that subjects believe their own advice.<sup>30</sup> In the Causal Strong dataset, however, only the bids are higher. Even though the differences in bids are marginally significant, they are not large, and as a result, the narratives that won the auction and were passed on to subjects in NATURAL are fairly representative of the full sample of narratives that we collected in ELICIT. In terms of bid levels, we observe substantial underbidding on average: across datasets, average bids are \$0.31-\$0.34 for submissions that we did not reject, while 65-69 percent of subjects rate advice as helpful (equating to an expected value of \$0.65-\$0.69). There is a sizable winner's curse in most cases, however, with winning bids averaging \$0.70-\$0.97.

<sup>29</sup>When we compare initial policies in ELICIT to those in the treatments where subjects did not provide advice, we find that more subjects in ELICIT choose rational policies, suggesting that having to provide advice prompted subjects to engage more closely with the dataset (see Figure A9 in the Appendix).

<sup>30</sup>In the BASELINE treatment, we also observe that more certain subjects tend to lean up (see Figure A10 in Appendix A).

Figure 11. Policies and Helpfulness in NATURAL



Notes: Average policies and rated helpfulness by narrative type. The error bars indicate 95 percent confidence intervals. The left graphs are for the Strong dataset, and the right for the Causal Strong dataset. Data is from NATURAL.

### 4.3 Results – NATURAL

We now examine how the endogenously generated narratives from the ELICIT treatment were received by subjects in the NATURAL treatment. Figure 11 displays the effect of each individual narrative shown to subjects, for the Strong dataset (left panels) and the Causal Strong dataset (right panels). The upper two panels report average policy choices, while the lower two panels show how helpful subjects rated each narrative.

In the Strong dataset, subjects who receive a Chain or Up narrative consistently choose higher policies than those who receive rational advice, although not all pairwise differences are statistically significant. Chain and rational narratives are also rated as more helpful than simple narratives. In the Causal Strong dataset, subjects who receive rational advice select the lowest policies, in some cases approaching the rational policy of 0.25. In contrast, Chain and Up narratives lead to higher policy choices, deviating from the rational policy. While these narratives are followed, they are rated as less helpful than rational advice. Overall, these results show that even naturally-generated causal narratives, particularly Chain narratives, can have strong effects, with sizes comparable to those of the structured narratives we designed.

**Result 6:** *Endogenously generated Chain narratives have strong effects on policy choices. They are generally perceived as helpful, though less so when they contradict*

*a strong causal relationship in the data.*

## 5 Narratives in Context

To this point, we have relied on an abstract experimental environment designed to eliminate subjects’ prior beliefs, thus allowing us to test different types of narratives on a level playing field. While this approach gives us tight control over the DGP and provides clean identification, it may raise questions about external validity. In this section, we show that two of our key findings generalize to more naturalistic settings: subjects find it easier to generate Chain narratives than Collider narratives when asked for real-world examples, and they continue to place weight on multiple models when presented with competing real-world narratives.

### 5.1 Not all Narratives are Created Equally

A key takeaway from our results thus far is that subjects tend to infer causal Chain relationships much more frequently than Collider relationships. Is this result driven by some feature of how we constructed the datasets, or do Chain relationships come more naturally to people in general?

To answer this question, we conducted an additional treatment, CONTEXT, in which we first introduced subjects to the concept of a DAG. We then presented subjects with three DAG structures and asked them to provide three real-world examples for each, using separate open-ended text boxes. The DAGs included a Chain DAG, a Collider DAG, and a Two Outcome DAG in which a single action variable directly affects two outcome variables. For each DAG, we also asked subjects to rate the difficulty of coming up with examples using a Likert scale. Subjects were paid \$0.10 for each valid (non-empty) example they provided.<sup>31</sup>

We use two measures to study which causal structures come more naturally to subjects: the number of examples that subjects generate and their self-reported diffi-

---

<sup>31</sup>We instructed subjects not to use artificial intelligence (AI) and informed them that any example we suspected of being generated by AI would be rejected. Importantly, our experimental design is robust to potential AI use. Because our goal is to compare the number of examples and self-reported difficulty across DAGs, AI usage would bias against finding differences: AI tools are equally capable of generating examples for all three DAG types and would likely also lead to uniform difficulty ratings across them.

culty ratings. Both measures suggest that Chain DAGs come more easily to subjects than Collider DAGs. On a scale from 0 (easiest) to 7 (hardest), subjects report an average score of 4.61 for the Two Outcome DAG, 4.62 for the Chain DAG, and 4.96 for the Collider DAG. The difference between the Chain DAG and the Collider DAG is highly significant ( $p < 0.001$ ). In terms of the number of examples, subjects provide an average of 2.33 for the Chain DAG, 2.23 for the Collider DAG, and 2.21 for the Two Outcome DAG (none of these differences are statistically significant).<sup>32</sup>

Why might causal Chain relationships come more naturally than causal Collider relationships? One possibility, suggested by evidence from psychology, is that people tend to prefer explanations involving a single exogenous cause rather than multiple causes, particularly in contexts where the question is whether one or two causes led to two outcomes (e.g., Vratsidis and Lombrozo (2022)). Chain structures may also be cognitively simpler because they require reasoning about only conditional distributions,  $p(z|a)$  and  $p(y|z)$ . Collider structures involve the joint distribution,  $p(y|z, a)$ , which may be more difficult to process. Identifying the precise psychological mechanism is challenging because DAGs are discrete objects, and there is no obvious way to continuously modify the Chain DAG to make it ‘closer’ to the Collider DAG.

We included the Two Outcome DAG because it is the only other way to construct a fully-connected, three-variable DAG. While self-reported difficulty and the number of examples generated were similar for the Chain and Two Outcome DAGs, subjects almost never produced a Two Outcome narrative in the ELICIT treatment.<sup>33</sup> Moreover, if subjects were frequently inferring the Two Outcome model in our other treatments, it would not account for the observed upward bias in initial policy choices: the optimal policy under the Two Outcome DAG is the rational one. One possible reason subjects gravitate toward the Chain DAG rather than the Two Outcome DAG is that the latter is intuitively incompatible with the data. If the action variable  $a$

---

<sup>32</sup>We also measured the time it takes for subjects to come up with each example, but given that subjects come up with more Chain examples, response time is a difficult measure to interpret. Conditional on providing an example, subjects take longer on average to generate a Chain example than a Collider example. However, this measure is biased because it excludes cases where subjects struggled to come up with a Collider example, which are precisely the cases that indicate greater difficulty.

<sup>33</sup>Some of the advice in ELICIT can be interpreted as a Two Outcome DAG, but the policy recommendations and choices associated with this advice are generally inconsistent with this interpretation.

has no causal effect on the outcome  $y$ , it becomes hard to explain the observed correlation between the auxiliary variable,  $z$ , and the outcome,  $y$ . In contrast, the Chain narrative offers a coherent (albeit incorrect) explanation for the correlations between  $a$  and  $z$ , and between  $z$  and  $y$ , by mistakenly reversing the causality between  $z$  and  $y$ .

## 5.2 Competing Narratives

Contrary to the assumption in much of the theoretical literature that individuals adopt a single model (e.g., Eliaz and Spiegler 2020; Schwartzstein and Sunderam 2021), our experimental results strongly suggest that subjects are often willing to entertain multiple models simultaneously. Does this pattern extend to real-world settings, where people hold stronger priors, or do those priors lead individuals to overwhelmingly adopt one model over another?

To answer this question, we included a second part in the CONTEXT treatment described above. In this part, we presented subjects with Chain and Collider narratives in ten real-world contexts, covering a diverse range of topics, including gun rights, space exploration, and taxation (see Table A3 of Appendix A for all ten contexts and narratives). For each context, we organized the same three real-world variables into Chain and Collider narratives. For example, in the gun rights context, the Chain narrative was “*Arming citizens with guns inevitably leads to guns in the hands of criminals, which then decreases public safety*”, while the Collider narrative was “*Arming citizens with guns improves public safety by protecting against the threat of criminals with guns*”. For each context, we asked subjects to report how frequently they believed each DAG applied in the real world. Subjects received \$1.00 for completing all ten tasks. We recruited a politically-balanced sample of 200 self-identified Republicans and Democrats via Prolific in January 2025. Including earnings from the first part of the CONTEXT treatment (described in the previous section), subjects earned an average of \$5.18 for 33.8 minutes of participation (\$9.19/hour).<sup>34</sup>

The results of the second part of the CONTEXT treatment provide strong evidence that people do entertain multiple models even when thinking about real-world scenarios. In 94% of the 2000 subject-narrative pairs, subjects placed positive weight

---

<sup>34</sup>To view the experiment directly, visit CONTEXT.

on both the Collider and Chain narratives. Furthermore, *no* subject placed weights of 0 and 1 on the two narratives in all ten contexts.<sup>35</sup> While one might worry that subjects were simply guessing or defaulting to 50/50 beliefs, Figure A11 in Appendix A shows otherwise. The figure displays kernel density plots of the relative weight placed on the Chain narrative across all ten contexts, separated by political affiliation. The distributions reveal systematic variation across contexts: in some cases, subjects favor the Chain narrative; in others, the Collider. Importantly, there is little difference between Democrats and Republicans in most contexts, suggesting that the responses are not driven by political preferences (our rationale for recruiting a politically-balanced sample). The one clear exception is gun rights: Democrats tend to place more weight on the Chain narrative, while Republicans lean toward the Collider narrative. This pattern further supports the interpretation that subjects are reporting meaningful, context-sensitive beliefs rather than responding at random.

**Result 7:** *In real-world contexts, subjects (i) find it easier to construct Chain narratives than Collider narratives, and (ii) believe both Chain and Collider narratives are possible within the same context.*

## 6 Discussion

### 6.1 Mistaking Correlation for Causation

The key mechanism driving our results is that subjects are prone to infer and believe misleading causal relationships when correlations are present in the data. While prior work in cognitive science has documented a general tendency to overinfer causality (Waldmann and Hagmayer (2013); Matute et al. (2015)), we believe we are the first to show that the mere presence of a correlated auxiliary variable is sufficient to induce mistaken beliefs in a causal model. This result is important: in reality, actions and outcomes are inevitably correlated with some irrelevant variable, for example through reverse causality or omitted variables.

---

<sup>35</sup>Because we asked subjects to report how often they believe each DAG holds in the real world, it is possible they are allowing for variation across time or place, believing that one DAG applies in some contexts and the other in others. However, we view this interpretation as unlikely for many of the scenarios. In several cases, such as space exploration, it is hard to imagine that subjects perceive meaningful variation across time or place that would justify endorsing both models on that basis.



## 6.2 Implications for Theory

We show that causal narratives consistently produce costly deviations in policy choices in the directions predicted by the Bayesian-network factorization formula. While we can reject the exact point predictions of the formula in most cases, we would encourage researchers to continue to use it to model causal narratives. Compared to the rational model, it performs notably better: it gets the directions right and is a very tractable way of incorporating narratives into theoretical models.

On the other hand, our finding that subjects place weight on multiple narratives, whether provided explicitly or constructed implicitly, suggests that the way competing narratives are modeled in existing theory may warrant revision. In Schwartzstein and Sunderam (2021), for example, narratives are assumed to be supplied only by experts, whereas our results show that individuals often generate their own interpretations of the data. Moreover, both Eliaz and Spiegler (2020) and Schwartzstein and Sunderam (2021) assume that agents adopt a single narrative at a time, in contrast to our evidence that subjects often weight multiple, competing models. With that said, our study represents only one data point, and further work is needed to assess whether the pattern of weighting multiple models generalizes to other contexts.

In our setting, one possible reason subjects weight multiple models is that the Chain, Collider, and Summary narratives each highlight patterns that are factually present in the data, making it difficult to decisively favor one over the others. Although subjects may understand that multiple causal models cannot simultaneously be true, they are still required to choose a single action. Faced with this ambiguity, it is perhaps natural that they hedge their bets, averaging the models in a semi-Bayesian manner, by forming posteriors about the likelihood that each model is correct and choosing an action according to these posteriors. Note, however, that subject behavior cannot be truly Bayesian, as a true Bayesian would use the available information (the dataset) to reject the false Chain and Collider narratives in favor of the true DGP.

In other settings, distinguishing between models may be more straightforward. Consider, for example, a setting with a single action and a single outcome. If one model implies no causal effect and the other implies a clear causal effect, then – given sufficient data – it would likely be much easier to identify and adopt the correct model.

### 6.3 False Narratives

The results of the ELICIT and NATURAL treatments demonstrate how misspecified models of the world can emerge, be transmitted as narratives, and mislead both senders and receivers, even in the absence of any intent to deceive. In light of these results, it is perhaps not surprising that false narratives and conspiracy theories are so pervasive. In fact, our results may understate the problem because narratives and statistical information (Summary narratives) are exogenously assigned in our experiment. In reality, as Bursztyn et al. (2023) show, people often prefer opinion programs (i.e., narratives) over straight news. This tendency to self-select into persuasive but potentially misleading narratives may further amplify the reach and impact of false narratives.

Our findings raise the thorny question of what can be done to counteract the effects of false narratives. We explore several possibilities, but show that causal narratives are very robust: they remain effective when they reflect only noisy relationships, when they contradict the true causal structure in the data, and when they compete directly with clear statistical information that should invalidate them. An open question for future research is whether subjects can learn the true relationship. Although we make the full joint distribution available to subjects, so that there is technically nothing to be learned, the cognitive psychology literature suggests that people are more likely to learn causal relationships when they are actively making decisions rather than passively observing data (see Waldmann and Hagmayer (2013) for a survey).

### 6.4 Future Directions and Conclusion

Our study focuses on understanding how causal narratives operate when the joint correlations in the data are fully known. This controlled setting allows us to establish a theoretical understanding of how narratives influence beliefs and decisions, and to compare the effectiveness of different types of narratives. We also extend two of our key findings to real-world contexts: (i) subjects more naturally generate Chain narratives, and (ii) they tend to weight multiple narratives simultaneously.

One possible direction for future research is to study environments where correlations are unknown or only partially known. In such settings, subjects may be less inclined to form their own models. In this case, elaborate narratives may be more

persuasive than simple ones, either because they provide coverage (Pennington and Hastie (1993)) or justification (Bursztyn et al. (2023)). When the data-generating process is ambiguous, the effectiveness of a narrative may also depend on how it cues specific memories or draws attention to certain observations (Kahana (2012); Bordalo, Gennaioli, and Shleifer (2020)).

Recent empirical work has begun to explore how causal narratives operate in real-world domains (e.g., Andre et al. (2025); Angrisani, Samek, and Serrano-Padial (2024); Espín-Sánchez, Gil-Guirado, and Ryan (2023); Goetzmann, Kim, and Shiller (2024)), and new methods have emerged for identifying narratives using textual analysis (e.g., Ash, Gauthier, and Widmer (2024); Lange et al. (2022); Flynn and Sastry (2024); Hüning, Mechtenberg, and Wang (2022)). Given the abundance and potential influence of causal narratives across domains such as politics, financial markets, macroeconomics, and health, we look forward seeing more work, both in laboratory settings and in the field.

## References

- [1] Aina, Chiara. 2025. “Tailored Stories.” working paper.
- [2] Ambuehl, Sandro and Heidi Thysen. 2024. “Choosing Between Causal Interpretations: An Experimental Study.” working paper.
- [3] Andre, Peter, Ingar Haaland, Christopher Roth, and Johannes Wohlfart. 2025. “Narratives About the Macroeconomy.” *Review of Economic Studies*, forthcoming.
- [4] Angrisani, Marco, Anya Samek, and Ricardo Serrano-Padial. 2024. “Competing Narrative in Action: An Empirical Analysis of Model Adoption Dynamics.” working paper.
- [5] Ash, Elliott, Germain Gauthier, and Philine Widmer. 2024. “RELATIO: Text Semantics Capture Political and Economic Narratives.” *Political Analysis*, 32 (1): 115-132.
- [6] Asparouhova, Elena, Michael Hertzel, and Michael Lemmon. 2009. “Inference from Streaks in Random Outcomes: Experimental Evidence on Beliefs in Regime

- Shifting and the Law of Small Numbers.” *Management Science*, 55 (11): 1766-1782.
- [7] Barron, Kai and Tilman Fries. 2024. “Narrative Persuasion.” working paper.
  - [8] Benabou, Roland, Armin Falk, and Jean Tirole. 2018. “Narrative, Imperatives, and Moral Reasoning.” working paper.
  - [9] Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2020. “Memory, Attention, and Choice.” *Quarterly Journal of Economics*, 135 (3): 1399-1442.
  - [10] Bursztyn, Leonardo, Aakaash Rao, Christopher Roth, and David Yanagizawa-Drott. 2023. “Opinions as Facts.” *Review of Economic Studies*, 90 (4): 1832-1864.
  - [11] Bursztyn, Leonardo, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth. 2023. “Justifying Dissent.” *Quarterly Journal of Economics*, 138 (3): 1403-1451.
  - [12] Chapman, Loren. 1967. “Illusory Correlation in Observational Report.” *Journal of Verbal Learning and Verbal Behavior*, 6 (1): 151-155.
  - [13] Conrad, Klaus. 1958. “Die beginnende Schizophrenie. Versuch einer Gestaltanalyse des Wahns [The onset of schizophrenia: an attempt to form an analysis of delusion].” Georg Thieme Verlag.
  - [14] Danz, Daniel, Lise Vesterlund, and Alistair Wilson. 2022. “Belief Elicitation and Behavioral Incentive Compatibility.” *American Economic Review*, 112 (9): 2851-2883.
  - [15] Eliaz, Kfir and Ran Spiegler. 2020. “A Model of Competing Narratives.” *American Economic Review*, 110 (12): 3786-3816.
  - [16] Eliaz, Kfir, Simone Galperti, and Ran Spiegler. 2024. “False Narratives and Political Mobilization.” *Journal of the European Economic Association*, forthcoming.
  - [17] Enke, Benjamin. 2020. “What You See is All There is.” *Quarterly Journal of Economics*, 135 (3): 1363-1398.
  - [18] Enke, Benjamin and Thomas Graeber. 2023. “Cognitive Uncertainty.” *Quarterly Journal of Economics*, 138 (4): 2021-2067.
  - [19] Espín-Sánchez, José-Antonio, Salvador Gil-Guirado, and Nicholas Ryan. 2023. “Praying for Rain: On the Instrumentality of Religious Belief.” working paper.

- [20] Esponda, Ignácio, Emanuel Vespa, and Sevgi Yuksel. 2024 “Mental Models and Learning: The Case of Base-Rate Neglect.” *American Economic Review*, 114 (3): 752-782.
- [21] Fan, Tony. 2024. “Choice-induced Misspecified Mental Models.” working paper.
- [22] Flynn, Joel P. and Karthik A. Sastry. 2024. “The Macroeconomics of Narratives.” working paper.
- [23] Frechette, Guillaume, Emanuel Vespa, and Sevgi Yuksel. 2024. “Extracting Models From Data Sets: An Experiment Using Notes-to-Self”. working paper.
- [24] Goetzmann, William N., Dasol Kim, and Robert Shiller. 2024. “Crash narratives.” working paper.
- [25] Glazer, Jacob and Ariel Rubinstein. 2021. “Story Builders.” *Journal of Economic Theory*, 193.
- [26] Graeber, Thomas. 2023. “Inattentive Inference.” *Journal of the European Economic Association*, 21 (2):560-592.
- [27] Graeber, Thomas, Christopher Roth, and Florian Zimmerman. 2024. “Stories, Statistics, and Memory.” *Quarterly Journal of Economics*, 139 (4): 2181-2225.
- [28] Hüning, Hendrik, Lydia Mechtenberg, and Stephanie Wang. 2022. “Using Arguments to Persuade: Experimental Evidence.” working paper.
- [29] Izzo, Federica, Gregory J. Martin and Steven Callander. 2023. “Ideological Competition.” *American Journal of Political Science*, 67 (3): 687-700.
- [30] Kahana, Michael J. 2012. “Foundations of Human Memory.” OUP USA.
- [31] Kamenica, Emir and Matthew Gentzkow. 2011. “Bayesian Persuasion.” *American Economic Review*, 101 (6): 2590-2615.
- [32] Kendall, Chad and Ryan Oprea. 2024. “On the Complexity of Forming Mental Models.” *Quantitative Economics*, 15(1): 175-211.
- [33] Lange, Kai-Robin, Matthis Reccius, Tobias Schmidt, Henrik Müller, Michael Roos, and Carsten Jentsch. 2022. “Towards Extracting Collective Economic Narratives from Texts.” working paper.
- [34] Langer, Ellen J. 1975. “The Illusion of Control.” *Journal of Personality and Social Psychology*, 32(2), 311–328.
- [35] Matute, Helena, Fernando Blanco, Ion Yarritu, Marcos Diaz-Lago, Miguel A. Vadillo, and Itxaso Barberia. 2015. “Illusions of Causality: How They Bias Our

- Everyday Thinking and How They Could be Reduced”. *Frontiers in Psychology*, 6:888.
- [36] Morag, Dor and George Loewenstein. 2024. “Narratives and Valuations.” *Management Science*, forthcoming.
  - [37] Pearl, Judea. 2009. “Causality: Models, Reasoning, and Inference.” Cambridge University Press.
  - [38] Pennington, Nancy and Reid Hastie. 1993. “Reasoning in Explanation-Based Decision Making.” *Cognition*, 123-163.
  - [39] Rabin, Matthew. 2002. “Inference by the Believers in the Law of Small Numbers.” *Quarterly Journal of Economics*, 117 (3), 775-816.
  - [40] Schotter, Andrew. 2023. “Advice, Social Learning, and the Evolution of Conventions.” Cambridge University Press.
  - [41] Schwartzstein, Joshua and Adi Sunderam. 2021. “Using Models to Persuade.” *American Economic Review*, 111 (1): 276-323.
  - [42] Shermer, Michael. 2008. ”Patternicity: Finding Meaningful Patterns in Meaningless Noise”. *Scientific American*, 299 (6): 48.
  - [43] Shiller, Robert. 2017. “Narrative Economics.” *American Economic Review*, 107 (4): 967-1004.
  - [44] Shiller, Robert. 2019. “Narrative Economics: How Stories Go Viral and Drive Major Economic Events.” Princeton University Press.
  - [45] Sloman, Steven. 2009. “Causal Models: How People Think About the World.” Oxford University Press.
  - [46] Spiegel, Ran. 2016. “Bayesian Networks and Boundedly Rational Expectations.” *Quarterly Journal of Economics*, 131 (3): 1243-1290.
  - [47] Stone, Deborah A. 1989. “Causal Stories and the Formation of Policy Agendas”, *Political Science Quarterly*, 104 (2): 281-300.
  - [48] Vrantzidis, Thalia H. and Tania Lombrozo. 2022. “Simplicity as a Cue to Probability: Multiple Roles for Simplicity in Evaluating Explanations.” *Cognitive Science*. 46 (7).
  - [49] Waldmann, Michael and York Hagmayer. 2013. “Causal Reasoning.” *The Oxford Handbook of Cognitive Psychology*. Oxford University Press.

# Online Appendix

## A Additional Figures and Tables

Table A1: Policy Predictions

	Strong	Noise	Null	Masked	Causal Strong	Causal Noise	Causal Null
Rational	0.5	0.5	0.5	0.5	0.25	0.25	0.25
Chain	0.75	0.62	0.5	0.65	0.65	0.53	0.5
Collider	[0.22,0.41]	0.35	0.5	0.34	[0.08,0.21]	0.21	0.25

Notes: Predicted policy for each dataset (column) and narrative (row). For the Collider narrative in the absence of noise, a range of policies is predicted because beliefs are not completely pinned down by the dataset.

Table A2: Anticipatory Utilities

	Strong	Noise	Null	Masked	Causal Strong	Causal Noise	Causal Null
Rational	0.5	0.5	0.5	0.5	0.54	0.54	0.54
Chain	0.54	0.51	0.5	0.52	0.52	0.5	0.5
Collider	[0.32,0.49]	0.52	0.5	0.50	[0.42,0.56]	0.56	0.54

Notes: Anticipatory utility for each dataset (column) and narrative (row), calculated using the expected utility equation from the main text and the subjective beliefs under each narrative. For the Collider narrative in the absence of noise, a range of utilities is predicted because beliefs are not completely pinned down by the dataset.

Table A3: Provided Narratives in CONTEXT

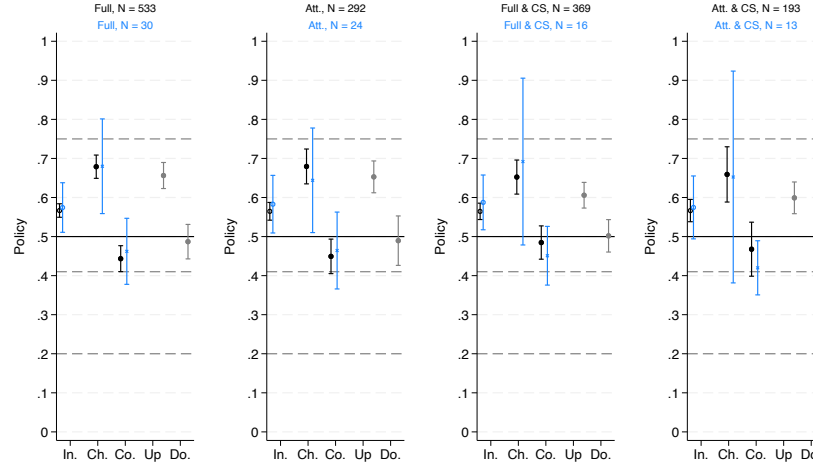
Context	Chain Narrative	Collider Narrative
Guns	Arming citizens with guns inevitably leads to guns in the hands of criminals, which then decreases public safety.	Arming citizens with guns improves public safety by protecting against the threat of criminals with guns.
Sanctions	Economic sanctions on a foreign enemy increase nationalism in the foreign enemy's country, which then increases the enemy's strength.	Economic sanctions on a foreign enemy decrease its strength by counteracting the threat of increasing nationalism in the foreign enemy's country.
Antibiotics	Antibiotic use causes diseases to become more dangerous, which then worsens health outcomes.	Antibiotic use improves health outcomes by counteracting the threat of dangerous diseases.
Social media	Social media use increases feelings of alienation from others, which then harms mental health.	Social media use improves mental health by reducing feelings of alienation from others.
Foreign aid	Providing aid to foreign countries discourages the use of local resources, which then hinders their growth.	Providing aid to foreign countries helps them grow by overcoming their lack of local resources.
Subsidies	Subsidizing domestic suppliers in key industries (e.g., semiconductors) encourages foreign countries to subsidize foreign suppliers, which then reduces the competitiveness of domestic suppliers.	Subsidizing domestic suppliers in key industries (e.g., semiconductors) ensures their competitiveness by counteracting the effects of foreign subsidies for foreign suppliers.
Taxes	Higher taxes encourage people not to pay their taxes, which then decreases government revenue.	Higher taxes increase government revenue by offsetting lost revenue from people that do not pay their taxes.
Space exploration	Funding for space exploration wastes earth's resources, which then reduces the welfare of future societies.	Funding for space exploration improves the welfare of future societies by counteracting the threat of limited resources on earth.
Government spending	Government spending decreases private investment by firms, which then reduces economic growth.	Government spending increases economic growth by counteracting low levels of private investment by firms.
Dietary supplements	The use of dietary supplements prevents natural nutrient absorption, which then worsens overall health.	The use of dietary supplements improves overall health by counteracting poor nutrient absorption from diet alone.



Figure A1: Noise Dataset

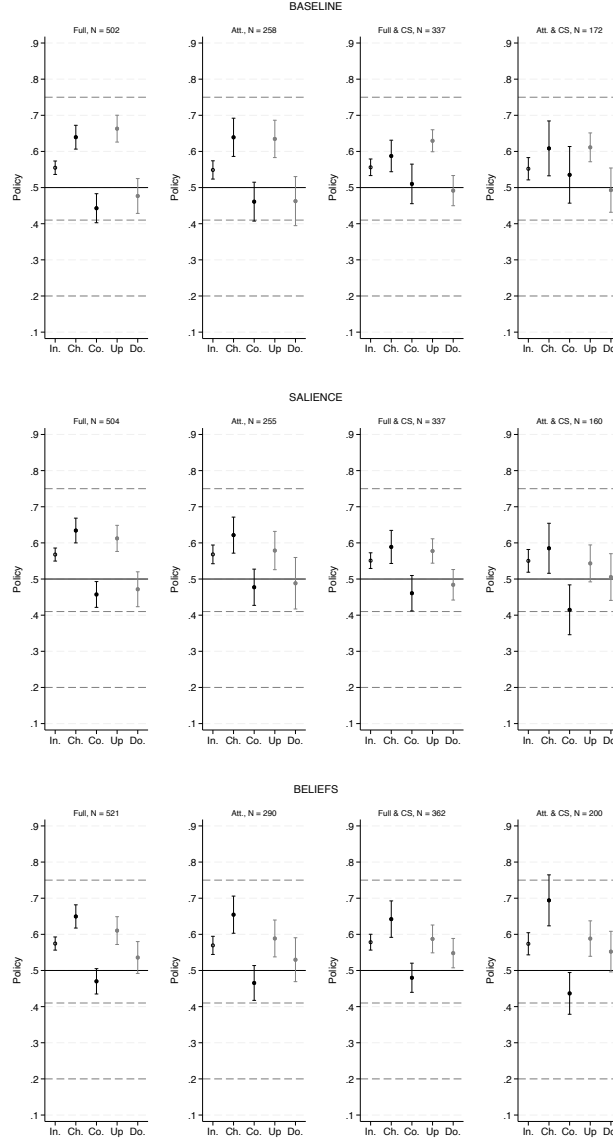
Choice	Payoff	X
GREEN	HIGH	○
BLUE	HIGH	▲
GREEN	HIGH	○
GREEN	HIGH	○
GREEN	LOW	○
GREEN	LOW	○
BLUE	LOW	○
BLUE	HIGH	▲
GREEN	LOW	○
BLUE	HIGH	○
BLUE	LOW	○
GREEN	HIGH	▲
BLUE	HIGH	▲
GREEN	LOW	○
BLUE	LOW	▲
BLUE	LOW	○

Figure A2: Policies in the Strong Dataset (First Dataset Only)



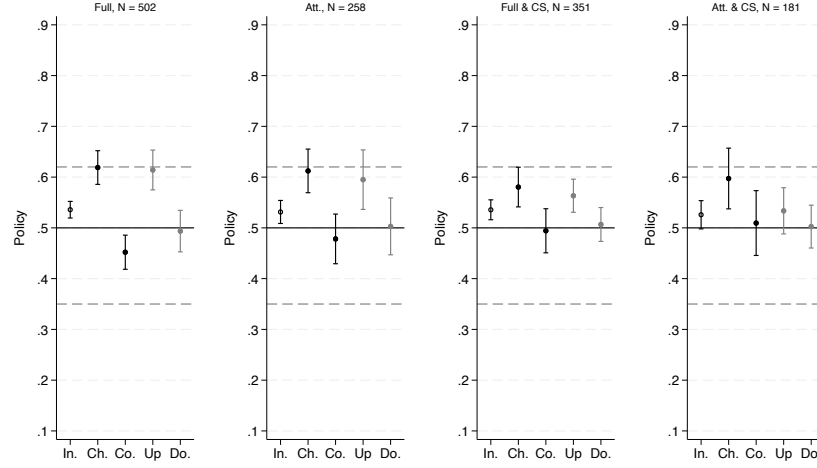
Notes: Average initial policy choices (In.), and those after Chain (Ch.), Collider (Co.), Up, and Down (Do.) narratives. Error bars represent 95 percent confidence intervals. We restrict the data to subjects who saw the Strong dataset first. The first panel includes all subjects in each sample. The second panel restricts to attentive subjects that do not respond to inconsistent narratives. The third panel is for policy choices when the narrative competed directly with the Summary narrative. The fourth panel is for policy choices of attentive subjects when the narrative competed directly with the Summary narrative. The solid line indicates the rational prediction while the dashed dark gray lines indicate the predictions of the BNFF for the Chain narrative (upper) and the Collider narrative (lower two lines indicate the predicted range). Data is pooled from BASELINE, SALIENCE, and BELIEFS.

Figure A3: Figure 3 for BASELINE, SALIENCE, and BELIEFS, respectively



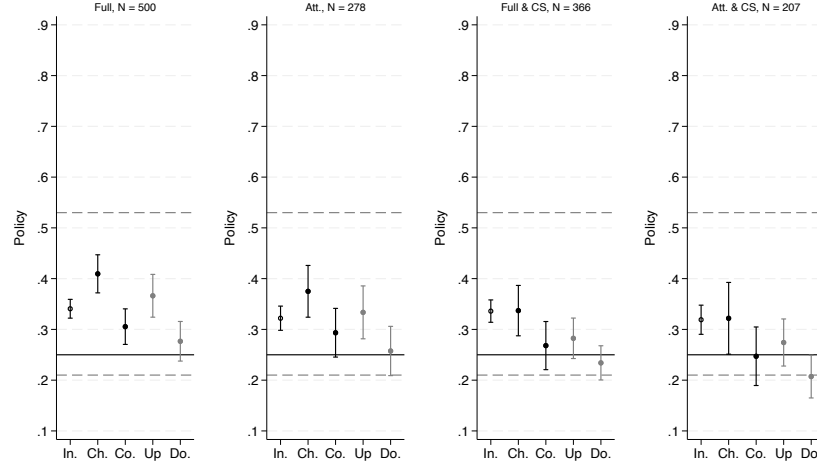
Notes: This figure replicates Figure 3 using only data from the BASELINE, SALIENCE, and BELIEFS treatment, respectively. Average initial policy choices (In.), and those after Chain (Ch.), Collider (Co.), Up, and Down (Do.) narratives. Error bars represent 95 percent confidence intervals. The first panel includes all subjects in each sample. The second panel restricts to attentive subjects that do not respond to inconsistent narratives. The third panel is for policy choices when the narrative competed directly with the Summary narrative. The fourth panel is for policy choices of attentive subjects when the narrative competed directly with the Summary narrative. The solid line indicates the rational prediction while the dashed dark gray lines indicate the predictions of the BNFF for the Chain narrative (upper) and the Collider narrative (lower two lines indicate the predicted range).

Figure A4: Policies in the Noise Dataset



Notes: Average initial policy choices (In.), and those after Chain (Ch.), Collider (Co.), Up, and Down (Do.) narratives. Error bars represent 95 percent confidence intervals. The first panel includes all subjects in each sample. The second panel restricts to attentive subjects that do not respond to inconsistent narratives. The third panel is for policy choices when the narrative competed directly with the Summary narrative. The fourth panel is for policy choices of attentive subjects when the narrative competed directly with the Summary narrative. The solid line indicates the rational prediction while the dashed dark gray lines indicate the predictions of the BNFF for the Chain narrative (upper) and the Collider narrative (lower two lines indicate the predicted range). Data is from BASELINE.

Figure A5: Policies in the Causal Noise Dataset



Notes: Average initial policy choices (In.), and those after Chain (Ch.), Collider (Co.), Up, and Down (Do.) narratives. Error bars represent 95 percent confidence intervals. The first panel includes all subjects in each sample. The second panel restricts to attentive subjects that do not respond to inconsistent narratives. The third panel is for policy choices when the narrative competed directly with the Summary narrative. The fourth panel is for policy choices of attentive subjects when the narrative competed directly with the Summary narrative. The solid line indicates the rational prediction while the dashed dark gray lines indicate the predictions of the BNFF for the Chain narrative (upper) and the Collider narrative (lower two lines indicate the predicted range). Data is from BASELINE.

Figure A6: Heterogeneity in the Strong Dataset (Attentive Subjects Only)

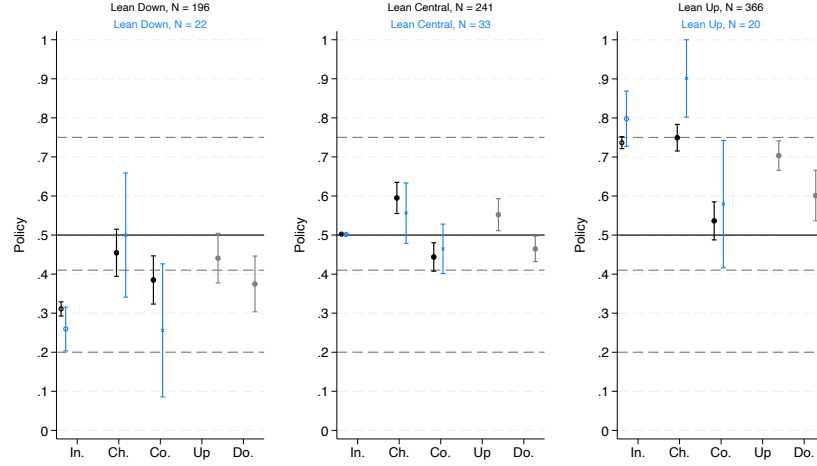


Figure A7: Heterogeneity in Beliefs after Single Narratives

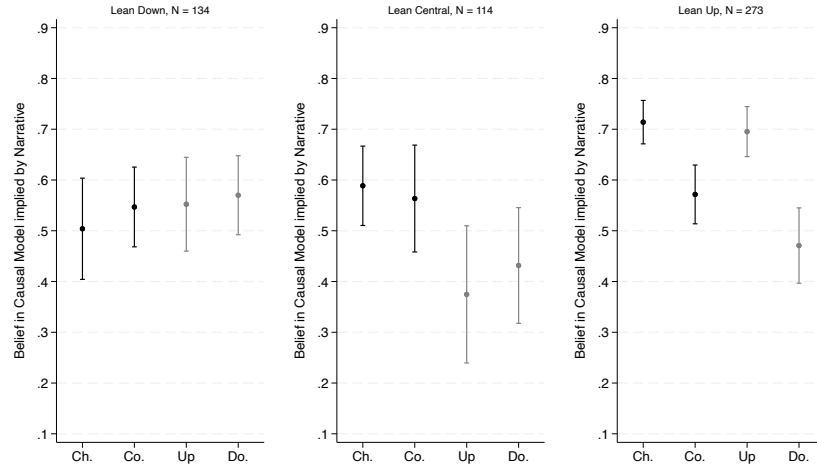
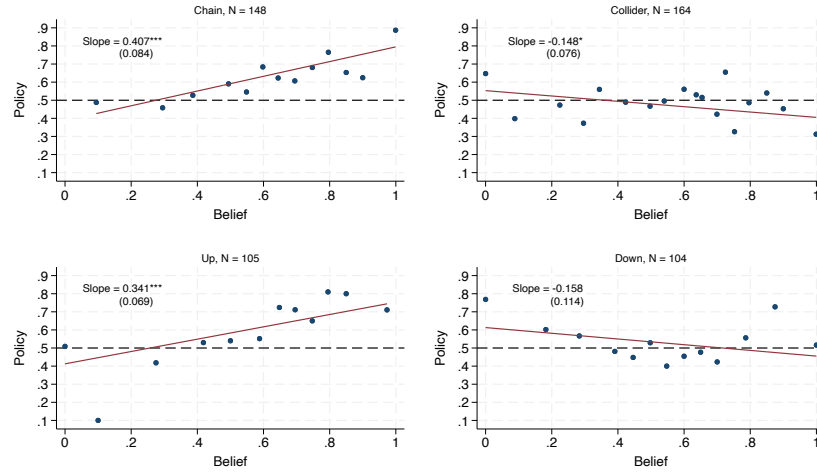
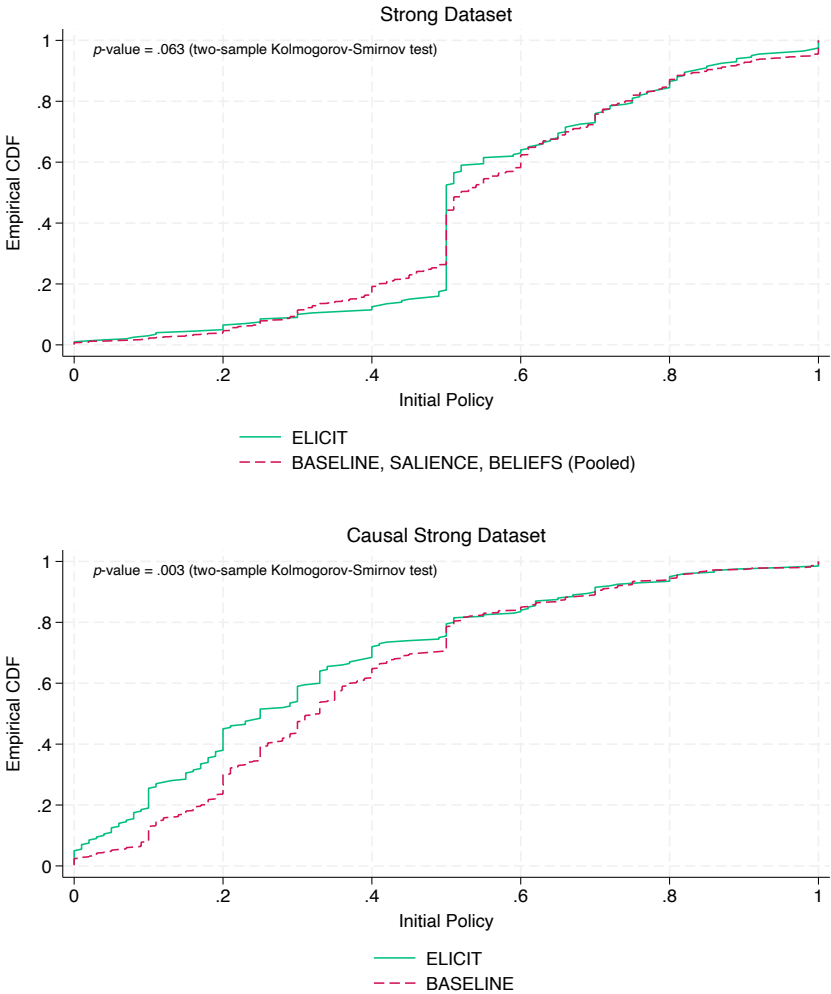


Figure A8: Policies vs. Beliefs after Single Narratives



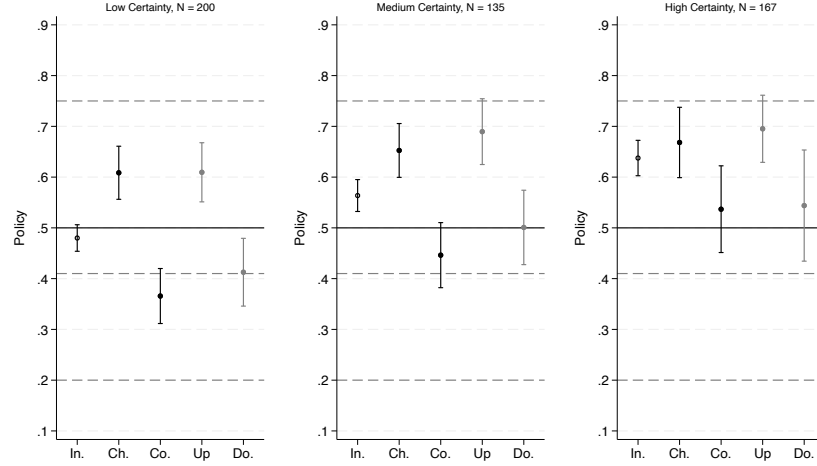
Notes: Binscatters of policy choices versus belief in the narrative indicated. Slopes and standard errors are shown for each panel. \* $p < .1$ ; \*\* $p < .05$ ; \*\*\* $p < .01$ . Data is from BELIEFS.

Figure A9: CDFs of Initial Policies in the Strong and Causal Strong Datasets



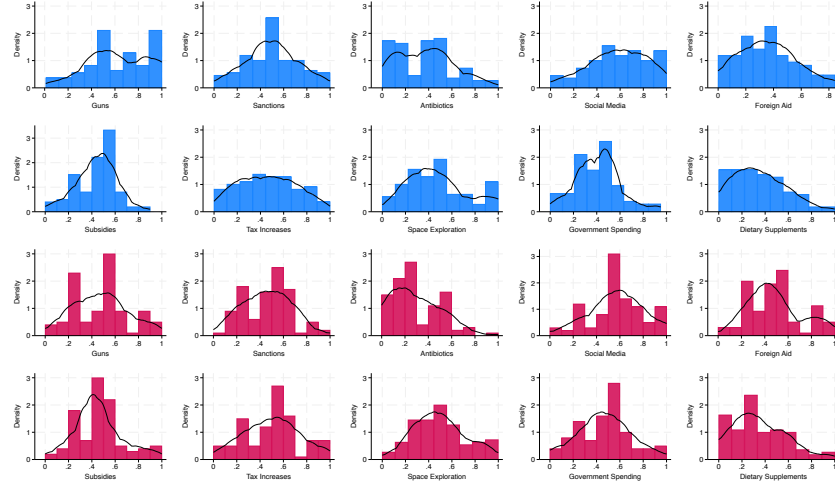
Notes: Empirical cumulative distribution functions of initial policies, shown separately for the Elicit treatment and other treatments, in the Strong and Causal Strong datasets.

Figure A10: Heterogeneity by Certainty in the Strong Dataset



Notes: Average initial policy choices (In.), and those after Chain (Ch.), Collider (Co.), Up, and Down (Do.) narratives. Error bars represent 95 percent confidence intervals. The first panel is for subjects whose self-reported certainty is in the first tercile. The second panel is for subjects with certainty in the second tercile. The third panel is for subjects with certainty in the third tercile. The solid line indicates the rational prediction while the dashed dark gray lines indicate the predictions of the BNFF for the Chain narrative (upper) and the Collider narrative (lower two lines indicate the predicted range). Data is from BASELINE.

Figure A11: Belief Share in the Chain Narrative in CONTEXT



Notes: Histograms and kernel densities for the belief share in the Chain DAG (relative to the Collider DAG) across ten different contexts. The upper ten plots are for subjects that self-identify as Democrats and the lower ten for those that self-identify as Republicans. Data is from CONTEXT.



## B Additional Results

### B.1 Undershooting of BNFF Predictions

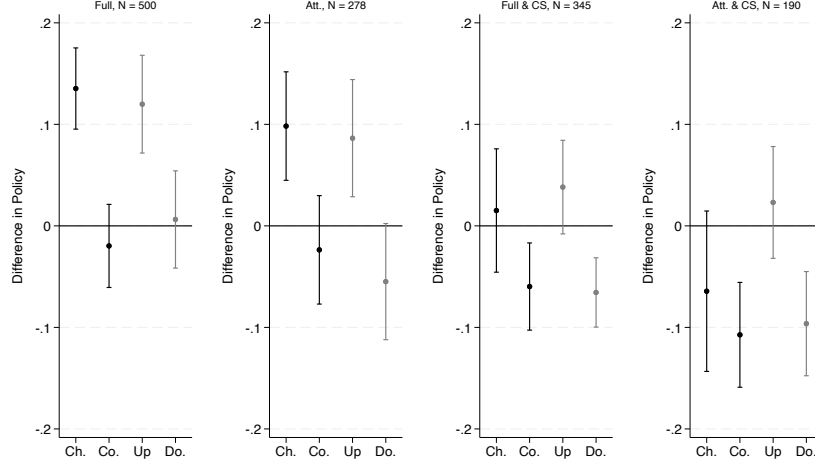
A common finding in belief elicitation tasks is that elicited beliefs tend to be compressed towards the middle (Danz, Vesterlund, and Wilson (2022)). Since the least costly policy in our setting is 0.5, it is possible that subjects' policies are compressed to 0.5, explaining why we find that subjects' policies tend to undershoot the predictions of the BNFF. Identifying compression to 0.5 in the Independent datasets, however, is challenging, since both the least costly policy as well as the rational policy coincide at 0.5. In contrast, in the Causal datasets, the least costly policy remains at 0.5, while the rational policy is at 0.25. This wedge allows us to show that compression is indeed occurring. Initial policies in the Causal Null dataset are compressed towards 0.5: the average policy choice is 0.33 which is significantly different from the rational policy of 0.25 ( $p < 0.001$ ).

This finding immediately raises a concern for results based on the Causal datasets: when we compare policies to the rational policy of 0.25, we might overstate the effects of Chain and Up narratives because subjects don't actually choose rational policies in the absence of a narrative. To rule out this possibility, we compute within-subject differences between policy choices that subjects make in Causal Strong datasets when they observe either the Chain or the Up narrative and their initial policy choices in Causal Null datasets. The idea is that initial policies in Causal Null datasets already capture the compression towards 0.5; any remaining movement towards 0.5 in Causal Strong datasets must be in response to the narrative. In Figure B1, we show that Chain and Up narratives continue to have positive effects by this measure, except when subjects see the Summary narrative simultaneously.<sup>36</sup>

---

<sup>36</sup>The tests when narratives compete with the Summary narrative are particularly strict because they ignore the fact that had a subject seen the Summary narrative when choosing their initial policy, the subject would likely have chosen a policy closer to the rational policy of 0.25 in response.

Figure B1: Policy Differences in BASELINE – Causal Datasets



Notes: Estimates of average differences in policy choices between choices made after seeing a narrative in the Causal Strong datasets and initial choices in the Causal Null datasets. Error bars indicate 95 percent confidence intervals with standard errors clustered at the subject level. The first panel includes all subjects in each sample. The second panel restricts to attentive subjects that do not respond to inconsistent narratives. The third panel is for policy choices when the narrative competed directly with the Summary narrative. The fourth panel is for policy choices of attentive subjects when the narrative competed directly with the Summary narrative. Data is from BASELINE.

A possible reason for the observed compression to the middle is cognitive uncertainty (Enke and Graeber (2023)).<sup>37</sup> Subjects might treat the least costly policy of 0.5 as a cognitive default, on which they lean when they are uncertain about the optimal policy. To investigate this possibility, we split the sample from the BASELINE treatment at the median reported certainty in policy choices in the Strong and Noise datasets. We find that subjects who are more certain deviate more from 0.5 for Chain and Up narratives. These differences are significant ( $p < 0.05$ ) except in the case of the Chain narrative in Noise datasets ( $p = 0.547$ ). On the other hand, we find no robustly significant differences for the Collider and Down narratives and the point predictions often go in the opposite direction, with more certain subjects being closer to 0.5. In the Causal datasets, subjects who are more certain are closer to the rational policy of 0.25 in their initial choices in Causal Null datasets ( $p = 0.078$ ).

Overall, the evidence for cognitive uncertainty is mixed, but we see stronger evidence for it when it is more likely to have bite. Specifically, for Collider and Down nar-

<sup>37</sup>Risk aversion doesn't straightforwardly result in compression because choosing an extreme policy reduces variability in the outcome.

ratives, it is challenging to test whether cognitive uncertainty modulates the amount that subjects deviate from 0.5, since these narratives do not lead to large deviations to begin with. In contrast, for Chain and Up narratives – narratives that cause the largest deviations from 0.5 in the Independent datasets – cognitive uncertainty does seem to modulate the amount of deviation.

## B.2 Column Ordering and Anchoring

Here, we leverage additional randomization in the design to rule out column ordering and anchoring as possible drivers of our results.

One reason the Chain narrative might be more robust than the Collider narrative is that the column ordering in the dataset naturally leads subjects to think of a causal chain moving from left to right. If this is the case, we would expect the Chain narrative to have a larger effect when the column ordering is  $a, z, y$  rather than  $a, y, z$ . To test for this possibility, we regress policy decisions when observing a narrative on a column ordering dummy. We find small and insignificant effects in all five datasets (Strong, Masked, Noise, Causal Strong, and Causal Noise) for all four types of narratives (Chain, Collider, Up, and Down), with one exception. In the Strong dataset with the Collider narrative, average policies are lower by 0.05 ( $p = 0.019$ ) when the column ordering is  $a, y, z$  (i.e., the Collider narrative is more effective). Overall, however, column ordering does not seem to have large effects.

One plausible reason that causal narratives are effective when they compete with the Summary narrative is that subjects may anchor their choices to the first narrative they see. To test for this possibility, we make use of the fact that some subjects see the Chain narrative and then both the Chain and Collider narratives, while others see the Collider narrative first. We regress choices when subjects see both elaborate narratives on a dummy that indicates that they saw the Chain narrative first, clustering standard errors at the individual level. If subjects anchor their choices, we would expect to see a positive coefficient. The results for each dataset are: Strong: 0.018 ( $p = 0.359$ ), Masked: 0.076 ( $p = 0.002$ ), Noise: 0.017 ( $p = 0.621$ ), Causal Strong: 0.043 ( $p = 0.291$ ), and Causal Noise: 0.049 ( $p = 0.165$ ). Thus, although each of the point estimates is consistent with anchoring, the effect is only significant for the Masked dataset. The lack of significance in the other datasets is unlikely due to a lack of power

because we have about 500 observations for the Strong dataset, 300 for the Masked dataset, and 150 each for the Noise, Causal Strong, and Causal Noise datasets. In sum, while we can’t rule out some amount of anchoring, we conclude that it is at most of second-order importance.

## C Prior Experimental Results

We circulated an older version of this paper in 2022 (Charles and Kendall, 2022). This “2022 version” contains the results of three treatments, which we have now removed from the current version of the paper. The previous working paper (available [here](#)) describes the prior experimental design and results in detail. Here, we discuss how the treatments in the main text differs from the treatments in the 2022 version and highlight some of the findings from that version. We also discuss how these results affected and motivated the design of the treatments reported in the main text. The treatments in the 2022 version were very similar to the BASELINE, ELICIT, and NATURAL treatments in the current version. Our prior treatments differed in the following main ways:

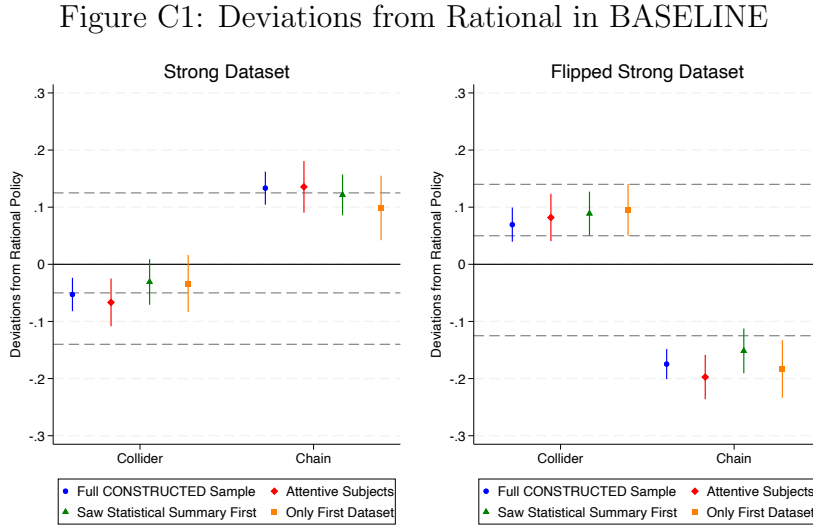
1. We framed the problem that subjects face as one of a manager choosing a policy, with the variables labeled as “Manager Action” ( $a$ ), “Employee Action” ( $z$ ), and “Firm Profits” ( $y$ ).
2. Subjects observed datasets containing 120 rows of observations.
3. Subjects observed three datasets in all treatments. These datasets were named slightly differently in the 2022 version of the paper compared to the current treatments. Specifically, the “positive” dataset corresponds to the Strong dataset, the “neutral” dataset corresponds to the Null dataset, and the “negative” dataset corresponds to a dataset that is symmetric to Strong, except that it swaps  $a = 0$  and  $a = 1$ . This yields a ‘flipped’ Strong dataset in which the Chain narrative supports a downward deviation from the rational policy of 0.5, and the Collider narrative supports an upward deviation.
4. Subjects only saw elaborate narratives and Summary narratives (i.e., they did not see simple narratives). They also never saw any competing narratives. Specifically, when making their second and third policy choices for each dataset,

subjects either saw an elaborate narrative or the Summary narrative (in randomized order). All narratives were framed as advice from a management consultant.

5. The cost parameter was  $c = \frac{4}{3}$ , double that of the experiments in the current version.

## C.1 BASELINE

Figure C1 plots deviations from the rational policy for several subsets of the data. The dashed dark gray lines in the graphs indicate the predictions of the Bayesian-network factorization formula (BNFF) for each type of narrative. The BNFF gives a point prediction for Chain narratives, while it only gives a range of predictions for Collider narratives. We find that, for the most part, subjects’ policies are remarkably close to the predictions of the BNFF.



Notes: Average policy deviations from the rational policy (0.5). Error bars indicate 95 percent confidence intervals. The left graph is for the Strong dataset, and the right is for the a ‘flipped’ version of the Strong dataset where Chain and Collider narratives are predicted to have opposite effects. Data is from an earlier version of BASELINE, which we ran in 2022.

We would like to highlight that across the Strong and ‘flipped’ Strong datasets, the movements in response to narratives are mirror images of each other. Specifically, in Figure C1, the deviations from the rational policy are almost perfect mirror images of each other across the two datasets. It is this symmetry that motivated us to focus

on the Strong dataset and omit the flipped version in the experiment reported in the main text.

In contrast to the experiment reported in the main text, subjects in our prior experiment saw Summary narratives for each dataset in isolation (i.e., without competing narratives). This allows us to check which policies subjects choose when they see only the Summary narrative. We find that Summary narratives move policy choices very close to the rational policy of 0.5: average policy choices after observing the Summary narrative are 0.53, 0.47, and 0.51, in the Strong, ‘flipped’ Strong, and Null datasets, respectively (not shown in a figure). The finding that subjects choose rational policies when provided with only the Summary narrative motivated us to omit isolated responses to Summary narratives from our current experiment. Finally, we chose to reduce the cost parameter in the current experiment, in order to generate more separation between the predictions of the various narratives, particularly in the Noise dataset.

## C.2 ELICIT

Similar to the treatment reported in the main text, subjects in our prior ELICIT treatment observed the Strong, ‘flipped’ Strong, and Null datasets in randomized order. For each dataset, they gave free-form advice, which could be shared with future subjects by bidding for the right to share it in a first-price auction. Of all the advice elicited for Strong or ‘flipped’ Strong datasets, we classified 18% as elaborate narratives, 51% as simple narratives, and 31% as rational narratives.<sup>38</sup> Of the 18% elaborate narratives that subjects identified, the vast majority (89%) are Chain narratives, providing further support for the result in the main text that subjects find it easier to identify Chain narratives in the raw data.

When we analyze bidding behavior, we find that subjects who identify an elaborate narrative are more bullish about their narrative compared to subjects who identify simple or rational narratives. As a result, elaborate narratives are more likely to be shared than narratives that (correctly) describe the independence of actions and outcomes. Specifically, of the narratives that were passed on from positive or neg-

---

<sup>38</sup>In the 2022 version, we labeled these categories slightly differently. Specifically, elaborate narratives were labeled as “causal” narratives and simple narratives as “other” narratives.

ative datasets, 25% are elaborate narratives (all Chain narratives), 55% are simple narratives, and 20% are rational narratives.

## D Narrative Classification

Each of the two co-authors independently classified each narrative into one of the categories shown in Table D1. In the case of disagreement (9.3% of cases), we first erred on the side of keeping the narrative: if only one co-author rejected, we kept it with the classification assigned by the other. This procedure resolved the vast majority of disagreements, but when it did not, we discussed until reaching agreement.

Table D1: Classification Descriptions

Classification	Code	Description
Reject	REJ	Does not contain an explicit or implicit (describes pattern) policy recommendation
Green Other	GO	Suggests green (policy $< 0.5$ ) but does not describe causal pattern
Green Chain	GL	Suggests green (policy $< 0.5$ ) and describes pattern for Lever narrative
Green Collider	GT	Suggests green (policy $< 0.5$ ) and describes pattern for Threat narrative
Green High	GH	Suggests green (policy $< 0.5$ ) and indicates that green more often leads to a HIGH payoff
Blue Other	BO	Suggests blue (policy $> 0.5$ ) but does not describe causal pattern
Blue Chain	BL	Suggests blue (policy $> 0.5$ ) and describes pattern for Lever narrative
Blue Collider	BT	Suggests blue (policy $> 0.5$ ) and describes pattern for Threat narrative
Blue High	BH	Suggests blue (policy $> 0.5$ ) and indicates that blue more often leads to a HIGH payoff
Neutral	N	Suggests a neutral policy either explicitly or by describing data as random
Rational	RAT	Suggests no policy direction but advises one to count HIGH and LOW payoffs for each choice

Table D2 provides the classification of all advice. In Table 1 of the main text, we aggregated the detailed categories as described in the notes of Table D2. All 804 elicited narratives and their classifications are available [here](#). Narratives that won the auction and were used in NATURAL are highlighted in yellow. Borders separate the groups for each auction (based on time of completion).

Table D2: Elicited Narratives – Detailed

Classification	Strong	Null	Causal Strong	Causal Null
Blue High	10	11	2.5	2.5
Blue Chain	14.5	1.5	7	0.5
Blue Collider	0	0	0	0
Blue Other	1.5	7	3	1
Green High	5.5	2	46.5	47
Green Chain	0	0.5	0.5	0.5
Green Collider	2	0	3.5	0
Green Other	3	3	2.5	6
Neutral	31	40	5	10
Rational	6	8	8	8.5
Blue Chain / Green High	0	0	1.5	0
Blue Chain / Green Collider	1	0	0.5	0
Reject	25.5	27	19.5	24

Notes: Classification of elicited narratives (percentages) in each dataset. Blue High and Green High suggest that the corresponding color leads to HIGH ( $y = 1$ ) outcomes more often. Blue Other and Green Other recommend the corresponding color, but do not provide a particular reason. Rational recommends counting the number of high outcomes under each action (color). Neutral recommends a policy of 0.5 explicitly or states that the outcome was random. To produce Table 1, we combined the advice into broader categories as follows. For all datasets, we combined Blue High and Blue Other into Up and the two categories indicating multiple narratives into Multiple. For the independent datasets, we combined Rational and Neutral into Rational, combined Green High and Green Other into Down, and relabeled Green Chain as Other. For the causal datasets, we combined Green High and Rational into Rational, relabeled Green Other as Down, and combined Neutral and Green Chain into Other.