# Causal Narratives

Constantin Charles and Chad Kendall*

March 25, 2024

**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 exogenously generated causal narratives manipulate decisions in ways inconsistent with rational theory. Instead, decisions are generally consistent with a behavioral theory, but with important exceptions, including when subjects face multiple narratives with contradictory recommendations. To study the generation and transmission of causal narratives, we show that they arise endogenously when subjects observe a dataset and provide advice to future subjects. These homegrown causal narratives mislead both the sender and receiver.

## 1 Introduction

Causal narratives – narratives that tell a causal story about the relationship between variables of interest – are ubiquitous. Examples abound in economics ("the

*Constantin Charles: Department of Finance, London School of Economics (e-mail c.charles@lse.ac.uk). Chad Kendall: Department of Finance and Business Economics, Marshall School of Business, University of Southern California and National Bureau of Economic Research (e-mail chadkend@marshall.usc.edu). We thank Kai Barron, 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, Chapman 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 UCI Finance Conference, the University of Ottawa, the University of Toronto, the University of Southern California, the University of Texas at Dallas, and the University of Utah for valuable input. Supplementary material is available here.

pandemic interrupted the supply chain, causing inflation"), in politics ("immigration leads to job losses for locals"), in medicine ("social media use causes depression"), and elsewhere in everyday life. These narratives can be truthful, describing true causal relationships in the data, or misleading, misrepresenting correlations in the data as causal. Understanding how and when these types of narratives are effective is critical for understanding how people come to form beliefs and opinions on a vast range of topics.

Although causal narratives may operate through many channels (e.g., emotions, recalling certain examples, etc.), one fundamental possibility is that they operate through *beliefs*: by providing a causal interpretation of correlations in the data, a causal narrative can potentially alter beliefs. Recently, economic theory has assumed causal narratives operate this way, suggesting a novel, tractable means of incorporating narratives into economic theory (Spiegler (2016); Eliaz and Spiegler (2020); Eliaz, Galperti, and Spiegler (2022)). If correct, it opens up a potentially fascinating research agenda, allowing us to understand the impacts of causal narratives on economic issues from financial bubbles to political polarization (Shiller (2017, 2019)).

In this paper, we conduct a series of controlled experiments in which we vary the narratives and data-generating processes (DGPs) that subjects are exposed to. Our primary goal is to understand which types of causal narratives are most effective, and why. Controlled experiments are ideal for this purpose because we require tight control over subjects' knowledge of the DGP to be able to isolate the effects of different narratives from other factors that might affect their influence in the real world (their source, peoples' priors, etc.). Our experiments pit standard theory against a behavioral theory in which subjects' beliefs can be manipulated through causal narratives. After testing exogenous narratives that we construct, we then go beyond either theory to understand the types of narratives that arise when subjects generate them, using these findings to bolster our understanding of why certain types of narratives are most effective.

To understand how causal narratives can potentially influence beliefs, consider a concerned parent that hears the narrative that social media use causes depression in teenagers. This narrative provides a model through which to interpret data: it implies a causal chain from an action (parental ban on social media use), to an auxiliary variable (social media use), to an outcome (depression). Let us suppose, however, that this example represents a case of reverse causality: depression increases social media use and not the other way round. Parents who understand this relationship would realize that banning social media use would have no impact on the well-being of their children, and therefore rationally not impose such bans. But, parents who accept the narrative, and update their beliefs accordingly, may instead believe a ban is appropriate.

The problem here is that correlations in the data *together with* the narrative can distort parents' beliefs. Using the notation of directed acyclic graphs (DAGs), we can state the problem as follows (an arrow indicates a causal relationship between variables, pointing in the direction of causality). If the true model is a case of reverse causality, *ban* $\rightarrow$ *social media use* $\leftarrow$ *depression*, it is clear that a ban has no effect on depression: the two variables are independent. Instead, the narrative implies the causal model, *ban* $\rightarrow$ *social media use* $\rightarrow$ *depression*, which, because of the positive correlation between social media use and depression in the data, may induce parents to impose a ban.

We test the efficacy of several such narratives across a range of different DGPs. Our experiment is split into two main sets of treatments – one in which we provide subjects with controlled narratives that we construct, and another in which we ask subjects to construct narratives on their own, as advice for future subjects. In both treatments, we consider a baseline environment in which the action and outcome are independent, as well as a second environment in which the action has a true causal effect on the outcome. Our baseline environment closely mirrors the social media and depression example, except that it is free of any context, instead using generic variable labels and values. Using a context-free environment is critical for testing the predictions of theory because in any specific context, subjects would likely have preconceptions about the DGP, potentially overriding any information we provide.

In both of our environments, subjects observe several compact datasets in sequence. Each dataset completely describes the joint correlation between three binary variables: an action, an outcome, and an auxiliary variable. Because subjects know the action is exogenous, a rational subject can easily form correct beliefs about the effect of each action on the outcome by simply counting the frequencies of each outcome (conditional on each action) in the DGP. Across the datasets, we vary how the auxiliary variable is generated: in some datasets, the auxiliary variable is independent of the action and outcome. In other datasets, it is correlated with both. This variation is key to understanding the extent to which causal narratives rely on correlations in the data to distort beliefs.

In the treatment in which we construct the narratives, we generate *elaborate* narratives that suggest a causal story by pointing out specific correlations in the dataset. These elaborate narratives come in two distinct types. The first is a *Lever* narrative: it implies a causal chain from action, to auxiliary variable, to outcome, as in the social media and depression narrative. The second is a *Threat* narrative: it implies that the action and auxiliary variables have direct effects on 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 addition to the elaborate narratives, we also construct *simple* narratives, which imply a direct causal effect between action and outcome.

After forming beliefs about the relationship between actions and outcomes, subjects choose a *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. A rational subject, after studying the baseline dataset, would understand the independence of actions and outcomes and therefore choose a policy of one-half (implying equal probabilities of each action).

A key assumption for incorporating causal narratives into theoretical models (Spiegler (2016); Eliaz and Spiegler (2020); Eliaz, Galperti, and Spiegler (2022)) is that people form beliefs according to the Bayesian-network factorization formula. Critically, 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). For the Lever narrative, it predicts beliefs will be distorted so that one action will be believed to produce the good outcome more often. For the Threat narrative, the prediction is the opposite: the other action will be believed to lead to the good outcome more often. 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.

In both our baseline environment in which actions and outcomes are unrelated, and in our second environment where the action has a causal effect on the outcome, we find that narratives generally induce policies different from the rational prediction and in the directions predicted by the Bayesian-network factorization formula. In the second environment, the Lever narrative even induces subjects to choose the action that produces the good outcome *less* often. Importantly, these predicted effects of narratives are robust to controlling for potential demand effects.

While the Bayesian-network factorization formula organizes the overall results reasonably well, we find several interesting departures from its predictions. First, the Threat narrative is much less robust than the Lever narrative, suggesting that some qualitative difference between the two that is not captured by the formula is important for the efficacy of narratives (we discuss some possibilities after presenting the results). Second, the mere presence of an auxiliary variable that is correlated with the action and outcome causes subjects to erroneously infer the causal relationship suggested by the Lever narrative themselves: prior to receiving any narrative, subjects' policy choices deviate from the rational policy in the direction consistent with this narrative.

We also test for the effects of competing narratives and whether or not narratives work even when we make the *true* relationship between actions and outcomes salient and explicitly recommend the rational policy. When confronted with two competing narratives, subjects choose policies that lie between the policies they choose when provided with either narrative on its own. Similarly, when jointly viewing a narrative and a summary of the true relationship, subjects choose policies that lie between the policy they chose with the narrative on its own and the rational policy. Importantly, we can show that subjects do not simply become confused by the two contradictory recommendations, but instead engage in a somewhat more sophisticated approach in which they weight both recommendations. This 'averaging' behavior seems somewhat intuitive, but contradicts the common theoretical assumption that only the 'best' narrative according to some criterion will be adopted (e.g., Eliaz and Spiegler (2020); Schwartzstein and Sunderam (2021)).

To further substantiate our finding that subjects infer the causal relationship suggested by the Lever narrative on their own, we ask whether subjects generate causal narratives when simply provided with a dataset of correlated variables. We provide a new group of subjects with datasets and again ask them to choose policies, but instead of providing them with narratives, we incentivize them to construct their own. We ask subjects to give free-form advice to future subjects (Schotter (2023)), and pay them according to how often their advice is rated as helpful by these subjects. In order to earn the right to share their advice, subjects must win a first-price auction.

We find that subjects produce all kinds of advice, with rational advice being the most common. But, strikingly, we find that many subjects produce elaborate narratives that contradict the true relationships in both our baseline environment (in which actions and outcomes are unrelated) and in our second environment (in which the two are, in fact, related). These subjects deviate more from the rational policy than other subjects, indicating that they believe their own advice. And, in some cases, they bid more than subjects that produce rational advice, thereby demonstrating a stronger preference to share their narratives. The vast majority of the elaborate narratives that subjects generate are Lever narratives, confirming that they erroneously infer the associated causal relationship on their own. On the receiving end, of all homegrown narratives, Lever narratives are most often rated as helpful and alter beliefs and actions in ways very similar to our constructed narratives. Thus, we see that false causal narratives can arise, be transmitted, and persuade, even absent any incentive to mislead.

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, narratives are receiving increased attention (Shiller (2017, 2019)) and being formally modeled. For our purposes, Eliaz and Spiegler (2020), building on Spiegler (2016), is most relevant, as we test key assumptions of their innovative conceptual framework that represents narratives as causal graphs that weave in auxiliary variables. Schwartzstein and Sunderam (2021), Izzo, Martin, and Callander (2021), and Aina (2022) consider how a principal can persuade an agent through a narrative represented as a model of the underlying DGP.[1] Although we don't test these other models explicitly, our experiment provides some of the first available evidence (together

[1] Benabou, Falk, and Tirole (2018) theoretically study narratives as they relate to morality norms.

with Barron and Fries (2023)) that persuasion via models (as opposed to signals or Bayesian persuasion experiments (Kamenica and Gentzkow (2011)) can be effective.

A handful of recent experimental papers study narratives from different perspectives. Perhaps most related, Andre et al. (2022) 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 this 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. Morag and Loewenstein (2021) show that people who tell stories about items they own, as opposed to simply describing them, ask for higher selling prices. Barron and Fries (2023) experimentally test persuasion via narratives using the theoretical framework of Schwartzstein and Sunderam (2021). Graeber, Roth, and Zimmerman (2024) show that stories are easier to recall than statistics, leading to larger impacts on beliefs. Ambuehl and Thysen (2024) study competing causal narratives in a setting that rules out, by construction, the averaging behavior that we find.

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 (2021); Kendall and Oprea (2022); Graeber (2023)). In particular, Frechette, Yuksel, and Vespa (2023) study whether subjects form the correct model when presented with datasets generated by different DAGs.

Outside of economics, narratives are mainly thought to be important because of their appeal to emotion (Fryer (2003); Quesenberry and Coolsen (2014)), a fact demonstrated by neuroscientists (e.g., Wallentin et al. (2011); Song, Finn, and Rosenberg (2021)). Narratives can also leverage peoples' abilities to identify with characters in the narrative (Jenni and Loewenstein (1997)). We complement these findings by showing that narratives can operate through beliefs by conveying a causal model.

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 use findings from this literature to guide our choices of parameters (Section 2.6) and to identify potential mechanisms that guide our experimental design (Section 3.1). Relatedly, the psychology literature on illusory correlation (Chapman (1967)) and apophenia / patternicity (Conrad (1958); Shermer (2008)) identifies instances in which people misperceive patterns in random data (typically images), similar to our finding that people erroneously infer causal relationships from data. Perhaps the most closely related work is that on the hot hand and gambler's fallacies – misperceptions of correlations in statistically independent events (Rabin (2002); Asparouhova, Hertzel, and Lemmon (2009)). Our work differs in that actual correlations are misinterpreted as causal relationships.

# 2 Conceptual Background

## 2.1 Environment and Rational Benchmark

We consider environments in which there are only three variables involved in the construction of a narrative: an action ($a$), an outcome of interest ($y$), and an auxiliary variable ($z$). All of the variables are binary, taking values 0 and 1. We describe a joint distribution function, $p(a, z, y)$, via a *dataset*. Table 1 illustrates a pair of datasets with different auxiliary variables. In both, $a$ and $y$ are statistically independent and both values of $a$ and $y$ are equally likely. In the left dataset ($I^+$), $z$ is generated as the logical AND of $a$ and $y$.[2] In the right dataset ($I^{NEU}$), $z$ is statistically independent of $a$ and $y$ (and each value is equally likely). 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 determining $p(y|a)$, knowing that $a$ is the choice variable and $y$ is the outcome of interest. A rational DM would use the conditional version of the law of total probability,

$$p(y|a) = \sum_{z=0,1} p(y|z, a)p(z|a) \tag{1}$$

a statistical formula which must hold for any joint distribution, $p(a, z, y)$.

[2] In a previous experiment (documented in Appendix C), we also tested an $I^-$ dataset in which $z = 1$ when $a = 0$ and $y = 1$. We found symmetric results (swapping $a = 0$ for $a = 1$) relative to the $I^+$ dataset.

Table 1. Dataset Examples

| $a$ | $z$ | $y$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |

| $a$ | $z$ | $y$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 1 |
| 0 | 1 | 0 |
| 0 | 1 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 1 |

Notes: In the dataset on the left ($I^+$), $z$ is generated as the logical AND of $a$ and $y$. In the dataset on the right ($I^{NEU}$), $z$ is statistically independent of $a$ and $y$.

Of course, the statistical formula on its own does not pin down the *causal* effect of $a$ on $y$. However, if one also knows that $a$ is exogenous, the law of total probability *is* sufficient to determine the causal effect of $a$ on $y$.[3] For example, for either of the datasets in Table 1, a rational DM would simply count the conditional outcomes to determine $p(y = 1|a) = p(y = 1) = \frac{1}{2}$, realizing that the auxiliary variable is irrelevant.

## 2.2 Narrative Examples

Suppose now a DM is presented with the following *Lever* narrative (Eliaz and Spiegler (2020)) when studying the $I^+$ dataset:

*"$z = 1$ only when $a = 1$. Further, when $z = 1$, $y = 1$ always. So, choose $a = 1$."*

The pattern highlighted in this narrative is completely factual – it can be verified with the data at hand. But, it suggests the false causal relationship in which $a$ influences $z$, which in turn influences $y$. A DM that hears this narrative might come to believe that she can increase the probability of $y = 1$ by choosing $a = 1$.

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

*"If $z = 0$, $y = 0$ whenever $a = 1$. To counteract this, choose $a = 0$ so that $y = 1$ is possible even if $z = 0$."*

Again, the pattern highlighted in this narrative is factual. But, this narrative implies

[3] We inform subjects of this exogeneity by saying that their choice of action 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.

a causal relationship in which $z$ and $a$ both influence $y$ directly. A DM believing it may try to increase the probability of $y = 1$ by choosing $a = 0$. Thus, for the same dataset, different narratives can potentially cause a DM to take different actions.

To understand the importance of the auxiliary variable, consider the $I^{NEU}$ 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.

## 2.3 Directed Acyclic Graphs and Beliefs

In our framework, narratives are distinguished by the causal model they convey. In contrast, the non-causal narrative corresponding to the $I^+$ dataset would simply convey the truth – that the action, $a$, has no effect on the outcome, $y$. Because narratives convey a discrete causal relationship, it is not possible to construct both a causal and a non-causal narrative that convey the same 'information': by definition, the information they convey (the model) must be different.

To describe the causal relationships that narratives imply, we employ directed acyclic graphs (DAGs), an idea first introduced by Pearl (1985). Within economics, Spiegler (2016) suggested the use of DAGs as a way to describe the subjective (behavioral) beliefs of a DM faced with a joint probability distribution, and Eliaz and Spiegler (2020) used DAGs as a means of describing narratives.[4]

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 conditional probabilities. Importantly, however, a DAG and a joint distribution *together* determine the conditional probabilities: they 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 joint distribution and her (potentially incorrect) belief in a causal model.

For example, consider the Lever narrative described previously. We refer to it as an *elaborate* narrative – it leverages the auxiliary variable into the narrative to imply 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

[4] Glazer and Rubinstein (2021) use DAGs to describe (not necessarily causal) stories.

compared to the conditional law of total probability, the conditioning of $y$ on $a$ is dropped, which can lead to incorrect beliefs. In particular, for the $I^+$ dataset in Table 1, the BNFF prediction for the Lever narrative is $p(y = 1|a = 1) = \frac{2}{3}$, which is greater than the true probability, $p(y = 1|a = 1) = \frac{1}{2}$. Because of the positive (upward) shift in beliefs under the Lever narrative, we refer to $I^+$ as a 'positive' dataset.

In the Threat narrative introduced previously, $z$ is a potential threat to producing $y = 1$, one which must be counteracted by choosing $a = 0$. It implies the causal relationship (often referred to as a collider or common consequence DAG), $a \to y \leftarrow z$, and the BNFF prescribes, $p(y|a) = \sum_{z=0,1} p(z)p(y|a, z)$. Treating $z$ as exogenous can again lead to incorrect beliefs: for the $I^+$ dataset in Table 1, the BNFF results in $p(y = 1|a = 1) = \frac{1}{4}$.

Finally, consider the DAG, $a \to y\ z$, where the lack of links between $z$ and the other variables implies that $z$ is statistically independent of $a$ and $y$. The conditional BNFF in this case is simply the identity, $p(y|a) = \sum_{z=0,1} p(y|a)p(z)$: the fact that $z$ is independent allows it to be 'factored out'. We refer to such a narrative as a *simple* narrative, one that implies that the action is the sole determinant of the outcome.[5] We distinguish between two simple narratives, one that recommends choosing $a = 1$ more often (Simple Up), and one that recommends choosing $a = 0$ more often (Simple Down). A DM that views either of the $I^+$ or $I^{NEU}$ datasets and believes either simple narrative should have rational beliefs. Importantly then, for the $I^+$ dataset, *if* DMs form beliefs according to the BNFF, they will form different beliefs under simple, Threat, and Lever narratives. Of course, the alternative hypothesis is that subjects form rational beliefs, motivating our experiment.

In constructing narratives that can potentially lead to mistaken beliefs, it is critical that $a$ and $z$, as well as $z$ and $y$, are in fact correlated. If one applies the BNFFs associated with either the Lever or Threat narratives to the $I^{NEU}$ dataset, beliefs are not distorted. For this reason, we refer to the $I^{NEU}$ dataset as a 'neutral' dataset.

[5] Simple, Lever and Threat narratives together with the 'true' DAGs we use to generate the data (the DAG corresponding to $I^+$ is $a \to z \leftarrow y$, that for the $C^+$ dataset we describe below is the same but with an extra link between $a$ and $y$, and that for $I^{NEU}$ is one with no links between any of the variables) are not a completely exhaustive list of all the possible DAGs with $a$ exogenous and $y$ as the outcome of interest. However, the other DAGs, such as $a \to z\ y$, imply the same beliefs as one in the set we consider.

## 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

$$u(y,d) = y - c(d-d^*)^2 \tag{2}$$

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 a quadratic or binarized scoring rule except that both beliefs, $p(y = 1|a = 1)$ and $p(y = 1|a = 0)$, affect policy choices.[6]

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

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

A change in $d$ has the direct effect of changing the probability of $a$, which changes a DM's expected utility according to her beliefs. But, it could also have a more subtle indirect effect through learning – a change in $d$ will change the frequency of $a$ and therefore affect the DM's beliefs through changes in the new data generated. We shut down this indirect effect in our experiment by not providing subjects with feedback about the realizations of the variables. Thus, we treat beliefs as fixed objects that represent the beliefs of a DM that has observed a dataset in which the policy has been held constant at some policy, $d = \delta$ ($\delta = \frac{1}{2}$ for the examples in Table 1).

For example, taking the $I^+$ dataset, a rational DM would simply choose $d = d^*$ because she would realize $p_G(y = 1|a = 1) = p_G(y = 1|a = 0) = \frac{1}{2}$. For a DM that believes a Lever narrative instead, 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 $d = d^* + \frac{1}{6c}$ so that to the extent that narratives distort subjective beliefs, they will also distort policy choices away from the rational choice, $d^*$.

[6] In explaining the mechanism to subjects, we provide the details but also provide simple examples to illustrate that choosing any policy $d$ different from $d^*$ entails a cost.

For the same dataset and a Threat narrative, we run into a difficulty calculating beliefs because $p(y = 1|a = 0, z = 1)$ is indeterminate: the combination of $a = 0$ and $z = 1$ never occurs in the joint distribution. To handle this case empirically, we allow for any subjective belief, $\gamma = p(y = 1|a = 0, z = 1) \in [0, 1]$ so that $p_G(y = 1|a = 0) = \frac{\gamma}{4} + \frac{3}{8}$.[7] 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 Lever narrative for any subjective belief, $\gamma$.

## 2.5 Other Theoretical Considerations

The BNFF is the only theory of which we are aware that provides quantitative predictions in our environment.[8] However, research from economics, cognitive science, and psychology suggests other qualitative factors that might affect the ability of narratives to influence beliefs. We introduce these factors here and discuss in Section 3.1.1 how they influenced our experimental design.

**Consistency and Coverage**: The ideas of consistency and coverage come from Pennington and Hastie's (1993) study of juror decision-making. They develop a qualitative model that identifies the features of causal narratives that make them more convincing when jurors link events using the evidence presented at trial. A narrative is *consistent* if it is not directly contradicted by the evidence (or, in our setting, by the dataset). For instance, elaborate narratives are consistent with the $I^+$ dataset, but inconsistent with the $I^{NEU}$ dataset. A narrative provides *coverage* if, in the context of a trial, it explains all of the available evidence. For our purposes, we say that a narrative provides coverage if it explains how all of the variables in a dataset come about. Elaborate narratives therefore provide coverage, while simple narratives do not.

**Falsification**: Narratives might influence subjects' choices because they are difficult to falsify. A Bayesian who has priors over narratives/causal models would always be able to reject a Lever or Threat narrative in favor of the rational model if the rational model is in the support of her priors.[9] But, even a non-Bayesian who believes

[7] We also consider perturbed datasets where all combinations of $a$ and $y$ occur. See Section 2.6.

[8] The cognitive science literature has put forth other models of causal reasoning (Waldmann and Hagmayer (2013)), but we are not aware of any that apply to our setup.

[9] Formally, the probability distribution implied by the dataset is not compatible (Markov) with

$p_G(y = 1|a = 1) \lesseqgtr \frac{1}{2}$ for some narrative should ask themselves why they observe $p(y = 1|a = 1) = \frac{1}{2}$ in the dataset.

Falsification of a narrative may be easier for some narratives than others. Eliaz and Spiegler (2020) show that Threat narratives generically violate what they call non-status quo distortion. Under the status quo policy (the frequency of $a = 1$ in the dataset), even the *unconditional* distribution of $y$ implied by the Threat narrative should be different than it is in the dataset.[10] Beliefs under a Lever narrative instead always lead to the correct unconditional distribution. Thus, Threat narratives might be easier to falsify than Lever narratives if it is easier to recognize that the unconditional distribution in the dataset is incorrect than that the conditional distributions are incorrect.

**Complexity**: Simple narratives are arguably less complex than elaborate narratives and thus may be more readily believed.

**Inattention**: Although we present the joint distributions in a very parsimonious way, as a small number of rows in a dataset, it is arguably easier to process a narrative than the statistical information in a dataset. If so, narratives might work due to inattention (rational or otherwise).

**Illusion of control**: A narrative may be more appealing if it provides *illusion of control* (Langer (1975)). For the $I^+$ dataset, a rational DM recognizes that she cannot influence the outcome and therefore chooses the least costly policy. Lever and Threat narratives instead suggest that the DM can control the outcome which might make them more compelling (Stone (1989)).

**Anticipatory utility**: Eliaz and Spiegler (2020) make the assumption that, when two narratives compete, the more 'hopeful' narrative is adopted – the narrative that provides the highest expected utility given the subjective beliefs it induces. We calculate the anticipatory utilities for each dataset and narrative combination used in our experiment in Table A2 of Appendix A.

the DAGs corresponding to the Lever and Threat narratives.

[10] In $I^+$, for example, under a Threat narrative, $p(y = 1) = \frac{5}{16} + \frac{\gamma}{8} < \frac{1}{2}$, whereas in the dataset, $p(y = 1) = \frac{1}{2}$.

Table 2. Additional Datasets

| $a$ | $z$ | $y$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |
| 0 | 0 | 0 |
| 0 | **1** | 1 |
| 1 | **1** | 0 |
| 1 | **0** | 1 |
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |

| $a$ | $z$ | $y$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |

| $a$ | $z$ | $y$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 1 |
| 0 | 1 | 0 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 1 |

| $a$ | $z$ | $y$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |
| 0 | 0 | 0 |
| 0 | **1** | 1 |
| 0 | 0 | 1 |
| 1 | **1** | 0 |
| 1 | 0 | 0 |
| 1 | **0** | 1 |

Notes: The datasets, $I^{NOISE}$, $C^{+}$, $C^{NEU}$, and $C^{NOISE}$ from left to right, respectively. Bold values indicate perturbations from $I^{+}$ and $C^{+}$, respectively.

## 2.6 Additional Datasets and Theoretical Predictions

In addition to the $I^{+}$ and $I^{NEU}$ datasets, we utilize four more datasets in the experiment, each of which is presented in Table 2. We refer to the set of $I$ datasets as independent datasets and the set of $C$ datasets as causal datasets.

The $I^{NOISE}$ dataset weakens the correlations in the dataset by perturbing some of the $z$ values in the $I^{+}$ dataset. In this case, the patterns for the Lever and Threat narratives only hold statistically, rather than deterministically. Because these perturbations do not change the relationship between $a$ and $y$, they do not change the rational predictions (but they do change the BNFF predictions). The causal datasets map one–to-one to the independent datasets except that a rational subject *should* infer a causal relationship from $a$ to $y$ because $p(y = 1|a = 1) = \frac{1}{3}$ while $p(y = 1|a = 0) = \frac{2}{3}$.

We choose $d^*$ and $c$ for the experiment with two goals in mind: (i) to be able to observe deviations from the policy that would be chosen by a rational subject and (ii) to make deviations costly so that any deviation observed is not simply due to a lack of incentives. To satisfy the first goal, we set $d^* = \frac{1}{2}$ so that we can observe deviations in either direction for independent datasets.[11] We chose $c = \frac{2}{3}$ to strike a balance between goals (i) and (ii): a lower value for $c$ will make deviations from the rational prediction easier to detect, but also reduce the cost from deviating. $c = \frac{2}{3}$ is large enough to ensure that flat incentives are not responsible for the results: a subject that deviates to one of the most extreme policies (0 or 1) earns one third less (on average) than a subject that chooses rationally for an independent dataset. In Table A1 of Appendix A, we summarize the predictions for all narrative and dataset combinations for both the rational and behavioral (BNFF) theories.

# 3 Testing Narratives

The goals of our first treatment are: (i) to see which types of causal narratives are most effective and in which environments, and (ii) to see whether behavior is best described by the behavioral (BNFF) theory or standard, rational theory. The basic idea behind our design is straightforward and broadly consists of three main steps (with slight variations). In the first step, we provided subjects with only a dataset and asked them to choose an initial policy. In the second step, we provided subjects with a narrative alongside the same dataset and asked them to make a second policy choice. In the third step, we provided subjects with an additional narrative or summary of the dataset alongside the first narrative and the dataset, and asked them to make a third policy choice. By observing subjects' policies in each of these steps, we can identify the effects of narratives in isolation and competition. In the section below, we provide a more detailed description of our experimental design and its variations.

## 3.1 Experimental Design – CONSTRUCTED

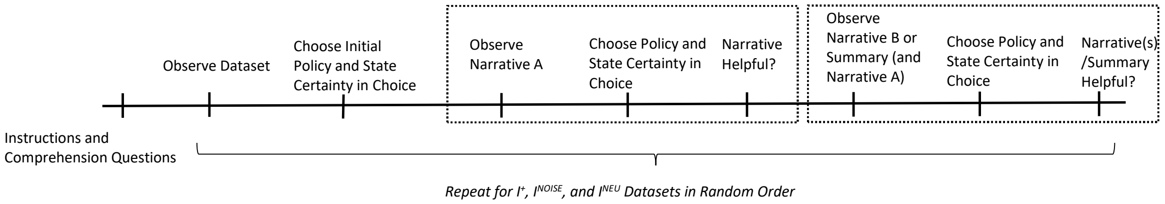
In the CONSTRUCTED treatment, subjects were placed into one of two arms. Subjects in the first arm observed the three independent datasets, $I^{+}$, $I^{NEU}$, and $I^{NOISE}$, while subjects in the second arm observed the three causal datasets $C^{+}$, $C^{NEU}$, and $C^{NOISE}$, in randomized order. The independent datasets consisted of 16

[11] The fraction of $y = 1$ in the dataset could also be chosen differently. We chose 50% because when the good outcome occurs very frequently, people are more likely to view a DGP as causal (Matute et al. (2015)).

rows, while the causal datasets consisted of 12 rows. We described the datasets with neutral variable values: $a \in \{BLUE(1), GREEN(0)\}$, $z \in \left\{\blacktriangle(1), \bigcirc(0)\right\}$, $y \in \{HIGH(1), LOW(0)\}$. We provide screenshots of all treatments in the Supplementary Material. The datasets were described as summarizing thousands of historical observations such that each row occurred an equal number of times. Subjects were also explicitly told 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. Figure 1 shows the timeline of the CONSTRUCTED treatment. In both CONSTRUCTED treatment arms, for each dataset, subjects completed the following tasks in order:

1. We presented the dataset, telling subjects that the variables may or may not be related, and asking 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 (2), in dollars. We gave subjects no feedback on the realization of $y$ or their payoff 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 (and the statistical summaries discussed below) 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 either a second narrative or a statistical summary of the data alongside the narrative from step 4 (randomizing which appears first), except in the case of the $I^{NEU}$ and $C^{NEU}$ datasets. For these datasets,

Figure 1. Timeline for CONSTRUCTED Treatments


Instructions and Comprehension Questions
Observe Dataset
Choose Initial Policy and State Certainty in Choice
Observe Narrative A
Choose Policy and State Certainty in Choice
Narrative Helpful?
Observe Narrative B or Summary (and Narrative A)
Choose Policy and State Certainty in Choice
Narrative(s) /Summary Helpful?
Repeat for $I^+$, $I^{NOISE}$, and $I^{NEU}$ Datasets in Random Order

we instead provided a second narrative on its own. In all cases, subjects could review the dataset and the piece(s) of advice together. They then made a third and final policy choice, rated how certain they were that their policy choice maximizes their earnings, and indicated whether they found the piece(s) of advice helpful.

As described in the task list above, subjects were presented with narratives and/or statistical summaries in steps 4 and 5. Here, we describe each of these objects in turn.

**Elaborate narratives:** We constructed both Lever and Threat narratives. The Lever narrative was *"X is a ▲ only when the choice is BLUE. Further, when X is a ▲, the payoff is always HIGH. So, choose BLUE more often."* The corresponding Threat narrative was *"If X is a ○, the payoff is always LOW when the choice is BLUE. To counteract this, choose GREEN more often so that the payoff can be HIGH even if X is a ○."*

**Noisy elaborate narratives:** For use with $I^{NOISE}$ and $C^{NOISE}$, we also constructed "noisy" versions of the Lever and Threat narratives that are identical to those above, except that we replace 'always' with 'more often', etc. We refer to the narratives in this case as *noisy* elaborate narratives.

**Simple narratives:** We constructed narratives that simply recommended an action. The Simple Up narrative was *"Choose the BLUE action more often"*, and the Simple Down narrative was *"Choose the GREEN action more often"*.

**Statistical summaries:** The statistical summary was a 2x2 table which summarized how often $y = 1$ and $y = 0$ occurred for each choice in the dataset. In addition to summarizing the data, the summary information explicitly told subjects to choose $d = 0.5$ in the case of independent ($I$) datasets and "*more green*" in the case of causal ($C$) datasets, thus providing an explicit recommendation (as with the narratives).

The summary tables can also be thought of as 'narratives' (e.g., in the case of the independent datasets, the table is a non-causal narrative, implying no causal relationship). We refer to the summary tables as summaries to avoid confusion, but consider them as narratives when we discuss competing narratives in Section 3.2.3.

Finally, we describe the randomization used to determine what was presented to subjects.

**$I^+$, $C^+$, $I^{NOISE}$, $C^{NOISE}$ datasets:** In step 4, subjects observed one of the two simple or two elaborate narratives, randomized across subjects.[12] In step 5, if subjects saw a simple narrative in step 4, they observed it again *together with* a statistical summary. If subjects saw an elaborate narrative in step 4, they observed it again with either the other elaborate narrative (i.e., they saw the Lever and Threat narratives together) or with a statistical summary, randomized across subjects.

**$I^{NEU}$ and $C^{NEU}$ datasets:** Subjects saw only simple narratives in step 4. In step 5, subjects saw the Lever narrative that was designed for the $I^+$ and $C^+$ datasets, which is clearly inconsistent with the data.

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

### 3.1.1 Understanding the Design

We designed the experiment to achieve several goals.

First, we purposefully framed the dataset as neutrally as possible to reduce the chances that subjects import prior beliefs about the DGP into the experiment.[13] In addition to labeling the values using neutral colors and symbols, we label the action, auxiliary, and outcome variables themselves as 'choice', 'X', and 'payoff', respectively, to try avoid any preconceived notions between the variables.

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

[12] 40% of subjects saw simple narratives and 60% saw elaborate narratives. We oversampled elaborate narratives to allow for more observations of these narratives in step 5.

[13] 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.

beliefs, it implies that subjects both (i) form a causal model from the narrative and (ii) use that causal model to update their beliefs.

Third, we use both independent and causal datasets to stress test the efficacy of narratives: because the causal datasets imply a fairly strong causal relationship, the two cases represent two extremes. The causal datasets also allow us to test whether (and which types of) narratives work when they oppose a true causal relationship in the data. Finally, the comparison across independent and causal datasets allows us to test for illusion of control: narratives may work in independent datasets because they give subjects false hope that they can control the outcome, even though policies actually have no effect on the outcome. In causal datasets, the rational policy also provides control over the outcome, so if policies deviate from the rational policy here, it cannot be because of illusion of control.

Fourth, we use datasets with noise, $I^{NOISE}$ and $C^{NOISE}$, to test whether narratives are robust to weaker correlations (more noise) in the data. Importantly, the $I^{NOISE}$ dataset also allows us to better compare the Lever and Threat narratives because the BNFF predicts slightly larger effects for the Threat narrative than the Lever narrative, unlike in the other datasets.

Fifth, we put narratives head-to-head to with other narratives and statistical summaries for two reasons. First, to see which factors drive subjects to adopt one or the other (e.g., anticipatory utility (Eliaz and Spiegler (2020)), coverage, etc.). Second, these tests provide a strong test of the inattention hypothesis. Narratives may work because subjects do not bother to process even the small number of rows in the dataset. To the extent that subjects fail to do so, summaries provide an extremely succinct description of the dataset and even go so far as to recommend the rational policy. Thus, if narratives work even when presented alongside summaries, it provides strong evidence that they work for reasons other than inattention to the dataset.

Sixth, we compare simple and elaborate narratives to test for coverage: both provide the same recommendation, but elaborate narratives provide coverage while simple narratives do not.

Seventh, we took seriously the possibility that subjects may respond to narratives regardless of whether or not they are consistent with the observed dataset. Though such behavior almost certainly occurs in reality (i.e., as in ‘fake news’), in an experimental environment it could reflect subjects simply not paying attention or trying to do what the experimentalist desires (a demand effect). To be able to identify and exclude such subjects in robustness tests, we presented each subject with an inconsistent narrative: a Lever narrative in $I^{NEU}$ or $C^{NEU}$. Because the pattern highlighted by the narrative is inconsistent with the dataset, such a narrative is easily falsified and will only be followed by inattentive subjects or those responding to demand effects.

Lastly, we randomized the order of the rows in a dataset across subjects to prevent any idiosyncrasy of the dataset from driving the results. We also randomized the order of presentation of the 'X' and 'Payoff' columns across subjects to test whether, for example, the Lever narrative is more likely to be adopted when the data is presented in the same order as the implied causal chain ($a \to z \to y$). In Appendix B, we show that column ordering has little effect, so we pool our data in what follows.

### 3.1.2 Implementation

We ran both arms of the CONSTRUCTED treatment online in April and May of 2023 using Qualtrics with custom Javascript coded by the authors.[14] We recruited a sample of the U.S. population, balanced between men and women, using Prolific (average age of 41.9). All sessions began with detailed instructions (replicated along with the decision screens in the Supplementary Material), after which subjects had to successfully answer several comprehension questions before continuing. We recruited 502 subjects in the CONSTRUCTED treatment with independent datasets and 500 in the CONSTRUCTED treatment with causal datasets. Subjects earned an average of $3.32 for an average of 13.5 minutes of their time ($14.76 per hour), a wage rate that is almost twice the minimum that Prolific requires ($8 per hour).

## 3.2 Results – CONSTRUCTED

We first show results for the independent datasets, establishing that causal narratives, particularly Lever narratives, have robust effects on policy choices, including when

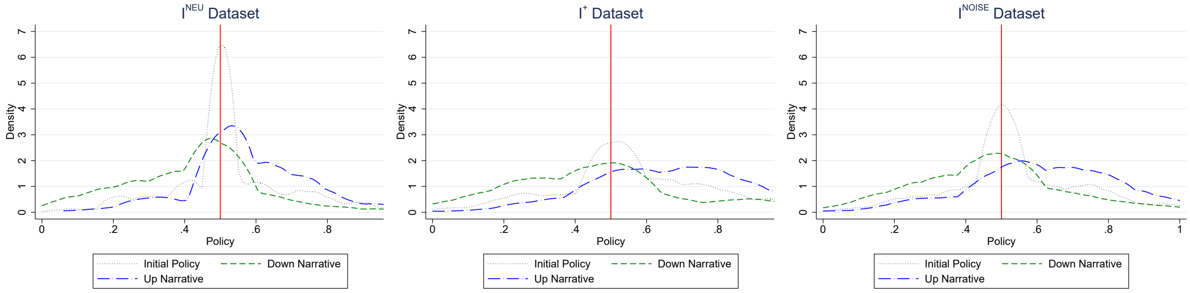
[14] To view the experiment directly, visit https://usc.qualtrics.com/jfe/form/SV_cYHxUUaMAhcypdI. A software bug resulted in incorrect initial bonuses. When we discovered the bug, we immediately corrected the issue by paying additional bonus payments (average of $0.05) in June of 2023. Importantly, the bug did not affect the data collected because the bonus was only reported to subjects at the end of the experiment.

narratives are noisy. We then present the results for causal datasets, showing that Lever narratives continue to work even when they recommend an action that opposes the true causal relationship. Lastly, we analyze the case of competing narratives.

### 3.2.1 Independent Datasets

We begin with an overview of policy choices in the three datasets. Figure 2 plots kernel density estimates of initial policy choices, as well as policy choices after seeing "Up Narratives" that recommend higher policy choices (Lever and Simple Up narratives) and after seeing "Down Narratives" that recommend lower policy choices (Threat and Simple Down narratives).

Figure 2. Policies in CONSTRUCTED – Independent Datasets



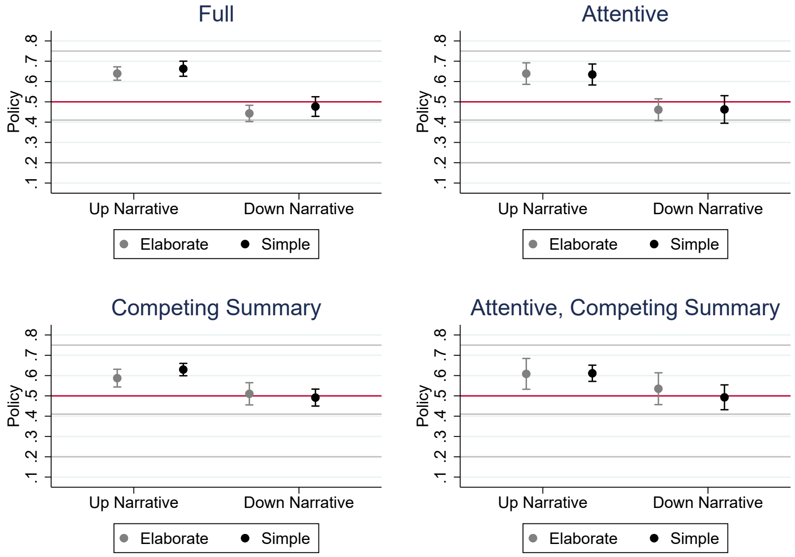
Notes: Kernel density estimates of policy choices in the three independent datasets of the CONSTRUCTED treatment. We show initial choices as well as choices after Up and Down narratives. Up narratives combine Simple Up and Lever narratives. Down narratives combine Simple Down and Threat narratives.

There are three key takeaways from Figure 2 that foreshadow the main results of the CONSTRUCTED treatment. First, initial policies are tightly concentrated around the rational policy choice of 0.5 in the $I^{NEU}$ dataset. Second, in the $I^{+}$ and $I^{NOISE}$ datasets, initial policy choices are more spread out, with a considerable mass of policies higher than the rational policy of 0.5. This result suggests that subjects may pick up on the correlations in the data associated with the Lever narrative on their own. Third, policy choices move in the direction of the narratives, particularly so in the $I^{+}$ and $I^{NOISE}$ datasets where the narratives point out true patterns in the data. Therefore, for the *same* dataset, different narratives can be constructed to move beliefs and actions in different directions.[15]

[15] In Appendix C, our prior experimental results also establish that the same narrative (e.g., Lever)

To perform statistical tests of the effects of narratives, we plot averages across subjects, beginning with the $I^+$ dataset in Figure 3. The upper left panel plots the averages of policy choices across all subjects, regardless of when they observed the $I^+$ dataset (first, second, or third).[16]

Figure 3. $I^+$ Dataset Average Policies



Notes: Average policy choices and 95 percent confidence intervals. The upper left panel is for all subjects. The upper right panel restricts to attentive subjects that do not respond to inconsistent narratives. The lower left panel is for policy choices when the narrative competed directly with a summary. The lower right panel is for policy choices of attentive subjects when the narrative competed directly with a summary. The red line indicates the rational prediction while the gray lines indicate the predictions of the BNFF for the Lever narrative (upper) and the Threat narrative (lower two lines indicate the predicted range).

Focusing first on the elaborate narratives, we confirm that Lever and Threat narratives result in different policy choices: the difference between the average policy choices under each narrative is highly significant (0.20, $p < 0.001$, two-sample t-test).[17] Furthermore, neither confidence interval contains 0.5, so that we can reject

can move beliefs and actions in different directions across datasets with different auxiliary variables. Thus, to the extent that someone constructing a narrative can choose the auxiliary variable they weave into the narrative, they can manipulate beliefs in the direction they prefer.

[16] The effects of narratives are slightly larger, albeit with larger standard errors due to reduced power, if we look only at those subjects that saw the $I^+$ dataset first (see Figure A1 of Appendix A).

[17] Leveraging the fact that subjects make initial policy choices, we can also look at changes in

the rational theory prediction. When we compare to the BNFF predictions, indicated by the gray horizontal lines, we see that average policies undershoot the prediction. We explore the reasons for this undershooting in Appendix B, showing that it is consistent with cognitive uncertainty (Enke and Graeber (2023)).

**Result 1**: *Lever and Threat narratives result in different policy choices for the same dataset. Both narratives distort policies away from the rational policy.*

For simple narratives, both the rational prediction and that of the BNFF is 0.5, but we can reject the null that the Simple Up narrative produces the rational policy of 0.5. In fact, this narrative produces a slightly larger response than the Lever narrative, though not significantly so ($p = 0.347$, two-sample t-test). The fact that the point estimates of the corresponding simple and elaborate narratives are very similar rules out coverage as being essential for narratives to work, at least in our setting.

One possible reason for this result is that the mere suggestion to choose a higher policy causes subjects to infer the causal relationship suggested by the Lever narrative themselves (as initial policy choices also seem to suggest). To test this hypothesis, we regress average policy choices when subjects observed a Simple Up narrative on a dummy indicating an $I^+$ dataset with the $I^{NEU}$ dataset as the omitted category, clustering standard errors at the subject level. Although no difference is predicted by either the behavioral or rational theories, we find that policies are significantly higher in the $I^+$ dataset (0.09, $p < 0.001$). Because the only difference between the two datasets is that the auxiliary variable is only correlated with the action and outcome variables in the $I^+$ dataset, it must drive the difference in policy choices.[18] When we look at Simple Down narratives that suggest to choose a lower policy, we also find a significant difference (0.05, $p = 0.043$), but policies are again higher in the $I^+$ dataset, contrary to the narrative. These results suggest that the causal model associated with

policy choices (relative to initial policy choices) as the result of observing a narrative. We find that Lever and Threat narratives result in highly significant changes in policies ($-0.099$, $p < 0.001$ via a t-test for the Threat narrative, 0.098, $p < 0.001$ for the Lever narrative). We also looked at heterogeneity in movement relative to the initial policy. While subjects do differ in how much they move away from their initial policies in response to a narrative, the bulk of the heterogeneity in movement is driven by the fact that subjects whose initial policies are further away from the BNFF predictions have more room to move towards these predictions.

[18] One may worry that this result is due to the fact that subjects previously saw a Lever narrative in one of the other datasets, but the difference is actually larger when we restrict to subjects that saw the $I^+$ dataset first (0.13, $p < 0.001$).

the Threat narrative is more difficult to pick up than that for the Lever, a finding we will confirm when we elicit advice from subjects (Section 4.2).

**Result 2**: *Subjects significantly respond more to Simple Up narratives when the auxiliary variable is correlated with the action and outcome variables, indicating that they infer the causal relationship implied by the Lever narrative on their own. As a result, coverage is not necessary for a narrative to affect choices.*

The above results could be driven by subjects blindly following narratives. In real-world settings, such as political debates, it is easy to imagine that many people do not pay close attention to the data backing up the narrative (indeed, they may not even have access to it), so these estimates may themselves be of interest. But in an experimental setting, blindly following narratives could reflect an artificial experimenter demand effect. To address this concern, the upper right panel of Figure 3 restricts the sample of subjects to *attentive* subjects – those who do not follow inconsistent narratives in the neutral dataset, where the pattern specified by the narrative does not exist. Specifically, if a subject changes from their initial policy in the direction of the inconsistent narrative by any amount, we exclude them. We adopt this very strict criterion to remove any possibility of demand effects, but the results are virtually identical if we adopt a weaker criterion, allowing changes of up to 0.05 in the direction of the narrative. Among attentive subjects (51% of all subjects), we see very similar effects to the full sample, demonstrating that Results 1 and 2 are not driven by subjects simply following whatever they are told.[19]

In the lower left panel of Figure 3, we look at policy decisions that subjects made when observing both a narrative *and* a summary that recommends choosing 0.5 explicitly. Testing for effects when narratives compete side-by-side with a summary of the true relationship serves two purposes. First, it helps to further address the concern that subjects might be blindly following narratives: given two recommendations, it is not clear why subjects who are inattentive or who want to please the experimentalist would follow one or the other, especially since we randomize the order in which the

[19] Rather than filtering out subjects that respond to the inconsistent narrative, we can include all subjects and compare average policy choices under the Lever narrative for $I^+$ datasets to those for $I^{NEU}$ datasets. To do so, we regress average policy choices when subjects observed a Lever narrative on a dummy indicating an $I^+$ dataset with the $I^{NEU}$ dataset as the omitted category, clustering standard errors at the subject level. The difference is highly significant (0.08, $p < 0.001$), indicating that Lever narratives are more effective when they can leverage correlations in the data.

two recommendations are displayed. Second, one may be concerned about a form of confirmation bias – subjects look only for information that supports the narrative. But, when given two pieces of advice, confirmation bias could work equally well for either. We continue to see significant positive effects for Lever and Simple Up narratives, and a significant difference between Lever and Threat narratives.

Finally, in the lower right panel of Figure 3, we look at the choice of only attentive subjects when they see both the narrative and the summary. This test is a fairly extreme robustness check in that it rules out inattention and demand effects via two methods simultaneously. Despite the resulting loss of statistical power, we still see a significant positive effect for both Simple Up and Lever narratives, indicating that these types of narratives are more robust than Threat and Simple Down narratives.

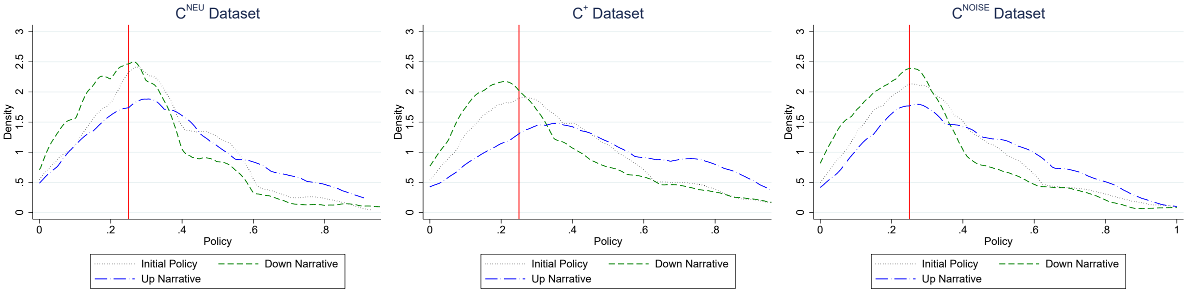
**Result 3**: *Lever narratives are more robust than Threat narratives.*

One may be concerned that this result is driven by the fact that the behavioral theory predicts a larger effect for the Lever narrative than for the Threat narrative, but we confirm this result with the $I^{NOISE}$ dataset where the theory predicts the opposite. In fact, broadly speaking, the results for noisy narratives are very similar to those for deterministic narratives (see Figure A2 in Appendix A), showing that they are just as effective. In particular, the Lever and Threat narratives significantly move choices in different directions (difference is 0.17, $p < 0.001$, two-sample t-test) and the Lever narrative always produces a statistically significant difference from the rational policy.

There are two subtle differences from the $I^{+}$ dataset, however. First, policies deviate somewhat less from the rational policy when the Lever is noisy, consistent with the BNFF prediction being lower. Second, the Threat narrative does not produce significant effects among attentive subjects or when competing with summaries, confirming Result 3 in a setting where the Threat narrative is predicted to have a slightly *larger* effect than the Lever narrative. These results suggest that there is something qualitatively different about Threat narratives that the behavioral theory does not capture. We consider several possibilities in Section 5.

**Result 4**: *Noisy Lever and noisy Threat narratives move subjects' policy choices in opposite directions. Therefore, deterministic patterns in the data are not necessary for narratives to be effective. Noisy Lever narratives are more robust than noisy Threat*

Figure 4. Policies in CONSTRUCTED – Causal Datasets



Notes: Kernel density estimates of policy choices in the three causal datasets of the CONSTRUCTED treatment. We show initial choices as well as choices after Up and Down narratives. Up narratives combine Simple Up and Lever narratives. Down narratives combine Simple Down and Threat narratives.

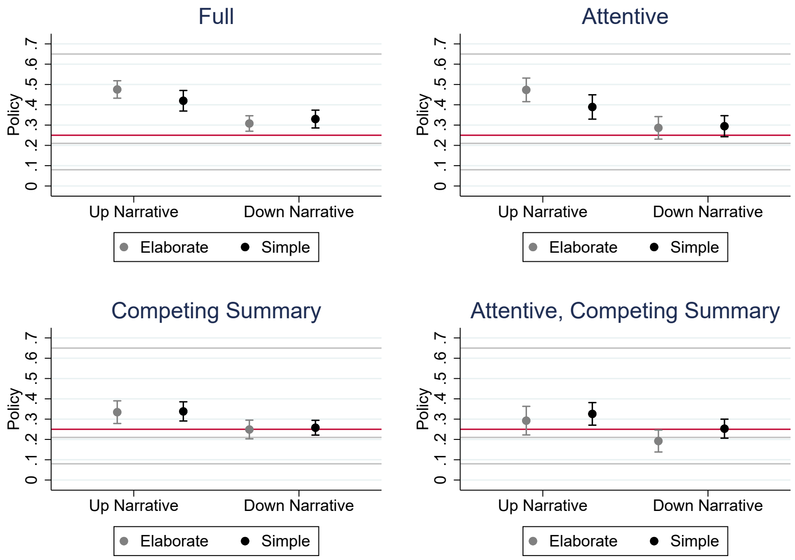
*narratives.*

### 3.2.2 Causal Datasets

Here, we test whether a causal narrative can be effective when it directly opposes a true causal relationship. As a first step, Figure 4 plots kernel density estimates of policy choices for each causal dataset in the CONSTRUCTED treatment. The modal initial policy choice in all three datasets is very close to the rational choice of 0.25. But, as we observed previously, initial densities shift towards higher policies in the $C^+$ and $C^{NOISE}$ datasets, again suggesting that subjects infer the causal relationship associated with the Lever narrative. We also observe notable upward shifts in policy choices after Up narratives and slight downward shifts after Down narratives. A particularly striking finding is that, after seeing an Up narrative, a considerable mass of subjects not only choose policies that are higher than the rational policy of 0.25, but many choose policies above 0.5 (i.e., in the opposite direction of the true causal relationship).

To formally test for the statistical significance of these patterns, we replicate Figure 3 for the $C^+$ dataset in Figure 5 (the corresponding figure for the $C^{NOISE}$ dataset is Figure A3 in Appendix A). In Figure 5, we see that the Lever narrative can be effective even when it contradicts the true causal relationship. The average policies chosen under Lever and Threat narratives are always significantly different (0.17, $p < 0.001$, two-sample t-test), and except for the lower right panel, which shows attentive subjects (56% of sample) that also saw a summary of the causal relationship, we can reject the null of no deviation from the rational policy for the Lever narrative.[20]

Figure 5. $C^+$ Dataset Average Policies

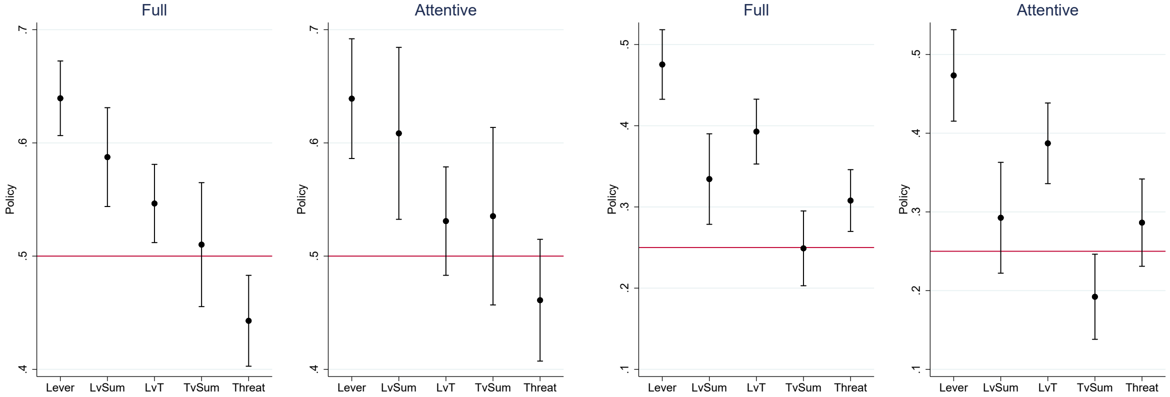


Notes: Average policy choices and 95 percent confidence intervals. The upper left panel is for all subjects. The upper right panel restricts to attentive subjects that do not respond to inconsistent narratives. The lower left panel is for policy choices when the narrative competed directly with a summary. The lower right panel is for policy choices of attentive subjects when the narrative competed directly with a summary. The red line indicates the rational prediction while the gray lines indicate the predictions of the Bayesian-network factorization formula for the Lever narrative (upper) and the Threat narrative (lower two lines indicate the predicted range).

Looking at simple narratives, subjects again appear to identify the correlations associated with the Lever narrative on their own: the Simple Up narrative produces similar policies to the Lever narrative. To formally test this hypothesis, we regress average policies under the Simple Up narrative on a dummy indicating the $C^+$ dataset with the $C^{NEU}$ dataset as the omitted category, clustering standard errors at the subject level. We find that the coefficient is positive but not significant in the full sample, but it is significant among attentive subjects (0.08, $p = 0.017$). In Figure A3, we see very similar effects when noise is added to the dataset, demonstrating again

[20] As we did for the independent datasets, we also compare policy choices under the Lever narrative in the $C^+$ and $C^{NEU}$ datasets, finding that the difference is highly significant (0.11; $p < 0.001$).

Figure 6. Competing Narratives



Notes: Average policy choices and 95 percent confidence intervals. LvSum indicates the average policy choice when subjects observed both a Lever narrative and a summary. TvSum indicates a Threat narrative and a summary. LvT indicates both Lever and Threat narratives. The left pair of graphs is for the $I^+$ datasets and the right pair of graphs is for the $C^+$ 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.

that deterministic patterns in the data are not necessary for narratives to be effective.

**Result 5**: *Lever, noisy Lever, and Simple Up narratives have significant effects even when they oppose a true causal relationship.*

### 3.2.3 Competing Narratives

In this section, we investigate how subjects behave when they face competing narratives. To do so, we analyze policy choices when subjects faced Lever and Threat narratives simultaneously and revisit the choices that subjects made when jointly facing an elaborate narrative (either a Lever or a Threat) and a summary (which is itself a form of narrative).

Figure 6 plots the average policy choices in $I^+$ and $C^+$ for the Lever and Threat narratives alone, each of these competing with a summary (LvSum and TvSum), and the two competing with each other (LvT).

The striking finding in Figure 6 is that subjects do not appear to adopt one narrative over the other, contrary to all of the theories put forth in Section 2.5. Instead, when subjects see competing narratives, they tend to choose policies that lie between the policies that they choose when they see each narrative on their own. For both datasets and both samples of subjects, we can statistically reject that average policy choices when jointly facing Lever and Threat narratives are the same as average policy choices for either narrative alone. Similarly, choices made when facing both a Lever narrative and a summary also always lie between those made when facing a Lever narrative alone and the rational prediction (though among attentive subjects we can't reject the null that they overlap with one or the other).[21] For the independent dataset, this finding can be interpreted as some evidence in favor of causal narratives over non-causal narratives, because the summary can be considered a non-causal narrative: although the summary reduces the effect of the Lever narrative, it does not eliminate it even though it has the advantage of being true.

The only case in which the policies under competing narratives do not lie between the two competing predictions is when the Threat narrative competes with a summary, which is likely a consequence of summaries killing off the effect of the Threat narrative. Figure A4 in Appendix A shows that the patterns for the $I^{NOISE}$ and $C^{NOISE}$ datasets are quite similar.

**Result** 6: *When subjects are confronted with competing narratives, they appear to mentally combine the two, producing policies that lie between the policies that they choose when they evaluate either narrative on its own.*

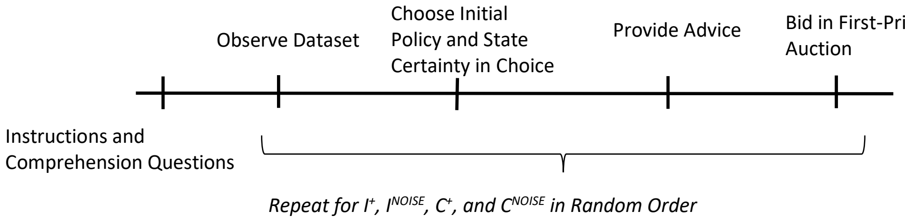
In Appendix B, we provide additional results on competing narratives. We show that choosing policies between those implied by each narrative alone is not driven by confusion: subjects confronted with two narratives are just as certain in their policies choices as when given only a single narrative. We also show that our results are consistent with a simple model in which subjects first form policy choices according to each narrative and then use the simple average when making their choice.

# 4 Creating and Transmitting Narratives

The pair of treatments we describe here serve two purposes. First, to make explicit the previous finding that subjects form their own causal models of the data-generating process. Second, to show that subjects construct narratives to communicate these

[21] In Appendix B, we provide evidence that this result is not driven by subjects anchoring on the Lever narrative because they see it first.

Figure 7. Timeline for Elicit Treatment


Observe Dataset
Choose Initial Policy and State Certainty in Choice
Provide Advice
Bid in First-Price Auction
Instructions and Comprehension Questions
Repeat for I+, INOISE, C+, and CNOISE in Random Order

models to other subjects, and that these narratives manipulate beliefs in ways very similar to the narratives we constructed.

## 4.1 Experimental Design – ELICIT and NATURAL

The first three steps of the ELICIT treatment are identical to those of CONSTRUCTED (observe the dataset, choose a policy, and state certainty). After completing step 3, subjects were given a free-form text box and asked to provide specific advice to future subjects. We endowed each subject with \$1.00 which they could use to bid in a first-price auction along with a group of nineteen other subjects. The winner's advice (we broke ties randomly) was provided to 40 future subjects (on average) and the winner was paid \$0.025 for each future subject that rated the advice as helpful (versus unhelpful). We told subjects that if their advice was not specific (didn't explicitly or implicitly imply a policy choice), it would be excluded from the auction.

Subjects completed these tasks for four datasets in random order: $I^+$, $I^{NEU}$, $C^+$, and $C^{NEU}$. The main comparison of interest is across datasets for which only the $z$ variable differs ($I^+$ vs. $I^{NEU}$ and $C^+$ vs. $C^{NEU}$), as we hypothesized that we would observe more causal narratives in the datasets in which $z$ is correlated with $a$ and $y$. We chose one of the four policy choices and one of the four auctions randomly for payment. Figure 7 summarizes the timeline for the ELICIT Treatment.

The NATURAL treatment is virtually identical to the CONSTRUCTED treatment except for the source of the narratives. In NATURAL, the narratives came from subjects that had observed the corresponding dataset and had bid the most in ELICIT for the right to share their advice (and subjects in NATURAL were made aware of this fact). Also, unlike in CONSTRUCTED, for the $I^+$ and $C^+$ datasets, subjects always saw the narrative paired with a statistical summary (i.e., no simultaneous narratives). For $I^{NEU}$ and $C^{NEU}$, instead of seeing a narrative paired with a summary, subjects saw the Lever narrative we constructed on its own.

### 4.1.1 Understanding the Design

We designed the experiment to achieve several goals.

First, we wanted to see whether causal narratives arise naturally when people have access to data in which correlations are present. Note that subjects in ELICIT were paid for their advice based on its perceived helpfulness (akin to receiving 'likes' on social media) instead of for the policy choices future subjects make. As such, subjects providing advice had no explicit incentive to try to manipulate the beliefs of future subjects. Future research should consider the effectiveness of causal narratives when a conflict of interest is present.

Second, we wanted to see whether causal narratives have to be deliberately constructed (for example, by a politician or marketing professional) to be effective, or whether narratives that arise naturally can have similar effects. Comparisons between CONSTRUCTED and NATURAL achieve this goal.

Third, we wanted to see whether or not elaborate narratives are more likely to be generated when no causal relationship between action and outcome exists in the data, relative to the case in which such a relationship does exist. Perhaps when a simpler, direct causal narrative exists, subjects are less likely to generate elaborate causal narratives.

### 4.1.2 Implementation

We ran the ELICIT and NATURAL treatments online in May of 2023 using Qualtrics with custom Javascript coded by the authors.[22] We recruited a sample of the U.S. population, balanced between men and women, using Prolific (average age of 41.1). All sessions began with detailed instructions (replicated with decision screens in the Supplementary Material), after which subjects had to successfully answer several comprehension questions to continue. We recruited 201 subjects in the ELICIT treatment, who earned an average of $4.31 for an average of 17.3 minutes of their time ($14.94

[22] To view the ELICIT experiment directly, visit https://usc.qualtrics.com/jfe/form/SV_1NdPWwQlZuaiFHE. For the NATURAL experiment, see https://usc.qualtrics.com/jfe/form/SV_aY1eM2y7jCKsyWO.

per hour). In NATURAL, 401 subjects earned an average of $3.25 for an average of 16.0 minutes of their time ($12.16 per hour).

## 4.2 Results – ELICIT

Subjects in the ELICIT treatment, for the most part, followed our instructions by providing advice that explicitly or implicitly recommended a policy choice: 608 of the 800 pieces of advice (76%) fulfilled this requirement. The remainder is generic advice such as "*the study needs to be read carefully*". As we told subjects we would do, we excluded such advice from the auction because it indicates a lack of attention to the instructions. To do so, each co-author independently decided whether each piece of advice provided an explicit or implicit recommendation and classified the narrative into one of several categories. We conservatively excluded only advice that both of us decided should be rejected. Our initial classifications agreed in just over 90% of cases, and, when not, we discussed until agreement was reached. In Appendix D, we provide the details of our classification procedure and a link to each piece of advice we elicited, together with how we classified it.

We classified the 608 pieces of advice into detailed categories in Table A3 of Appendix A. In Table 3, we combine the original categorizations into broader categories (as described in the notes for Table A3). The first column of Table 3 shows the advice that subjects produced for the $I^+$ dataset. The most common advice was rational advice that argued 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%"*). But, when combining all forms of causal advice, almost as much advice suggested a causal relationship, with the overwhelming majority of that arguing for a higher policy (as in a Simple Up or Lever narrative). Thus, almost half of subjects that provided explicit advice pointed to a causal relationship rather than discovering the true, non-causal relationship in the data.

Most strikingly, much of the causal advice explicitly points out the Lever 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 Threat narrative on their own. The fact that subjects

Table 3. Elicited Narratives

| Classification | $I^+$ | $I^{NEU}$ | $C^+$ | $C^{NEU}$ |
|---|---|---|---|---|
| Simple Up | 11.5 | 18 | 5.5 | 3.5 |
| Lever | 14.5 | 1.5 | 7 | 0.5 |
| Simple Down | 8.5 | 5 | 2.5 | 6 |
| Threat | 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 Lever narrative and Threat narrative or a Lever narrative and rational advice. 'Other' consists mainly of advice that says the process is random or to choose 0.5 in the causal datasets (very few are Lever narratives that point towards low, rather than high, policies).

identify the Lever narrative much more often than the Threat narrative makes explicit the finding that subjects in CONSTRUCTED implicitly infer an erroneous causal model in choosing their initial policies.[23]

The second column of Table 3 provides a breakdown of elicited narratives in the $I^{NEU}$ dataset. In this dataset, where all three variables are statistically independent, rational advice is even more prevalent. Furthermore, the number of elaborate narratives (Lever or Threat) almost disappears entirely: three Lever narratives are produced and for two of these, the subject had already seen one of the $I^+$ or $C^+$ datasets where the Lever pattern is actually present. The absence of elaborate narratives in this dataset indicates that correlations of $z$ with $a$ and $y$ are necessary for the emergence of elaborate narratives.

In the third and fourth columns of Table 3, we show the breakdown for the $C^+$ and $C^{NEU}$ datasets, respectively. As with the independent datasets, the most common advice is rational but we still observe Lever narratives when the auxiliary variable is correlated with the action and outcome variables (in $C^+$). This finding is particularly striking, because the Lever narrative implies choosing a higher policy, directly contradicting the strong causal relationship in the data that implies the opposite.[24]

[23] In a previous experiment, documented in Appendix C, we implemented a similar version of the ELICIT treatment and also found that subjects are much more likely to identify Lever narratives than Threat narratives.

[24] Lever narratives are not only discovered after first discovering them in the $I^+$ dataset: the fraction of Lever narratives produced is about the same when the $C^+$ dataset is observed before the

**Result** 7: *Subjects produce elaborate causal narratives after simply observing a dataset containing auxiliary variables, but almost exclusively when correlations in the dataset are consistent with such narratives. Subjects are much more likely to produce the Lever narrative than the Threat narrative, and do so even when the Lever narrative directly contradicts a strong causal relationship in the data.*

The fact that subjects construct causal stories from correlations in the data is reminiscent of apophenia or patternicity (Conrad (1958); Shermer (2008)), in which people see patterns that don't necessarily exist in the data. This psychology literature typically shows that people find patterns in visual data, such as images in ink blots. 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, Hertzel, 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.[25]

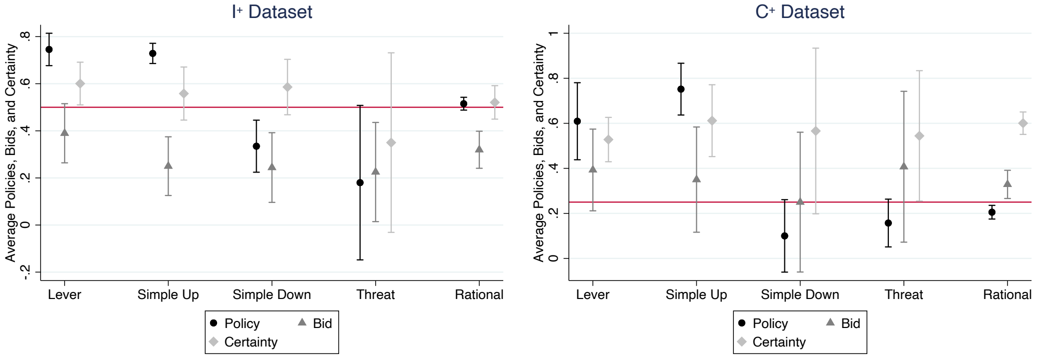
We designed the ELICIT treatment such that it is in subjects' best interests to provide advice that appears helpful (as opposed to necessarily being helpful). But, because we also elicit policy choices for these subjects, we can evaluate whether subjects follow their own advice. On the one hand, subjects may truly believe their own advice and act in accordance with it. On the other hand, they may provide some type of advice (e.g., a Lever narrative) but not act on it, either because they think that their advice will be perceived as helpful though realizing it is not, or because they are uncertain whether it is actually good advice. In Figure 8, we show average policies for the $I^+$ and $C^+$ datasets for each type of narrative (we exclude the categories Multiple and Other to simplify the figure). We see that, on the whole, subjects follow their own advice: policies are very close to the rational policy when subjects provide rational advice, and deviate in the indicated direction of the advice for the other types of advice.

Figure 8 also shows subjects' average bids and average certainty in their policy

$I^+$ dataset.

[25] Subjects that construct an elaborate narrative in either of the $I^+$ or $C^+$ 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 8. 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 $I^+$ dataset, and the right for the $C^+$ dataset. Red lines indicate rational policies.

choices. In the $I^+$ dataset, we find that those who produce a Lever narrative are more certain on average, and bid slightly more than those who produce rational advice, providing additional evidence that they believe their own advice. In the $C^+$ 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.

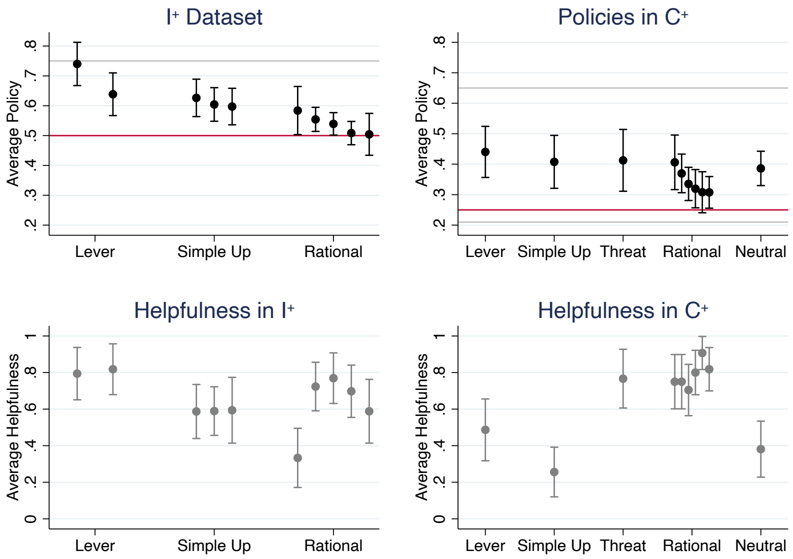
In terms of bid levels, we observe substantial underbidding on average: across datasets, average bids (among those whose advice we did not reject) are \$0.31-\$0.34 while 65-69 percent of subjects rate advice as helpful (equating to an expected value of \$0.65-\$0.69). There is a sizable winner's curse in most cases, however, with winning bids averaging \$0.70-\$0.97.

## 4.3 Results – NATURAL

In the NATURAL treatment, subjects received the narratives constructed by the subjects in ELICIT. Figure 9 illustrates the effect of each individual narrative for the $I^+$ and $C^+$ datasets. The upper two panels show the average policy that subjects chose after seeing one of the narratives, and the lower two panels show how subjects perceived its helpfulness, on average.

Focusing on the the $I^+$ dataset, we find that the average policies chosen by subjects

Figure 9. 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 $I^+$ dataset, and the right for the $C^+$ dataset.

who saw a Lever or Simple Up narrative are consistently higher than those of subjects who saw a rational narrative (though not all pairwise comparisons are statistically significant). The Lever and rational narratives are also considered somewhat more helpful than simple narratives.

In the $C^+$ dataset, subjects who saw a rational narrative chose the lowest policies, in some cases approaching the rational policy of 0.25. The single Lever and Simple Up narratives produce policies away from the rational policy but, although these narratives are followed, they are not rated as being as helpful as rational advice.

Overall, these results suggest that even naturally-generated causal narratives, particularly Lever narratives, can have strong effects, in some cases as large as the effects of the narratives we constructed.

**Result** 8: *Endogenously grown Lever narratives have strong effects. They are perceived as helpful, though less so when they contradict a strong causal relationship in the data.*

# 5 Discussion

## 5.1 Mistaking Correlation for Causation

The key mechansim driving our results is that subjects are apt to infer misleading causal relationships from correlations in the data. First, causal narratives are generally only effective when the auxiliary variable generates correlations in the data. Second, even simple narratives work when correlations are present, because subjects find the patterns associated with a causal model themselves, a result made explicit through the advice subjects provide to others. Third, correlations through the auxiliary variable allow causal narratives to be effective even when they suggest a causal relationship that directly opposes the true causal relationship. Although work in cognitive science has shown that people tend to overinfer causality (Waldmann and Hagmayer (2013); Matute et al. (2015)), we believe that we are the first to show that the simple presence of an auxiliary variable induces mistaken beliefs in a causal model. This result is important: in reality, actions and outcomes will inevitably be correlated with some irrelevant variable, either through reverse causality or omitted variables.

## 5.2 Implications for Theory

We found that causal narratives produce costly deviations in the directions predicted by the Bayesian-network factorization formula. Although we could reject the exact point predictions of the Bayesian-network factorization formula in most cases, we would encourage researchers to continue to use the formula to model causal narratives because it does better than the rational model: it gets the directions right and is a very tractable, parameter-free way of incorporating narratives into theoretical models.

On the other hand, our finding that subjects do not adopt one narrative over another when confronted with competing narratives suggests it may be fruitful to model competing narratives differently than is currently assumed. As we show in Appendix B, one can, for example, easily model the averaging behavior we observed by calculating beliefs according to the BNFF for each narrative and then performing some weighted average of the two.

## 5.3 Not all Narratives are Created Equally

Our results point to two key findings about the types of narratives that are effective. First, Lever narratives are more effective than Threat narratives. Second, Simple Up narratives are almost as effective as Lever narratives. Both findings depart from the BNFF predictions, suggesting that narratives come with properties not captured by the formula.

The fact that Simple Up narratives work just as well as Lever narratives can be explained by subjects picking up on the correlations in the data after being given a simple 'nudge' in this direction. Importantly, this result means that narratives always have to implicitly compete with the stories people tell themselves, which may be one reason Threat narratives are not as effective as Lever and Simple Up narratives.

There are additional reasons that Lever narratives may be more effective. Causal chains may come more naturally to people, or people may have a demand for simple, 'mechanistic' causal chains.[26] Alternatively, it may be that Lever narratives are less complex by some measure of complexity (Oprea (2020), Kendall and Oprea (2022)), such as the number of exogenous variables involved (one for Lever, but two for Threat narratives). Some evidence consistent with this possibility comes from studies (e.g., Vrantsidis and Lombrozo (2022)) showing that people tend to value simplicity in explanations. It is also possible that Threat narratives are easier to falsify because they violate non-status quo distortion: if the Threat narrative were true, the unconditional distribution of outcomes should be different than that observed in the data. Pinning down exactly why Lever narratives do better is non-trivial because each implied DAG is a discrete object: there does not appear to be any straightforward way to modify some feature of a Lever narrative to make it 'closer' to a Threat narrative, for example. Extending our methods to settings with more than three variables or unobserved variables (i.e., omitted variable bias) may be a promising direction for generating a metric that ranks the appeal of narratives.

[26] On the other hand, superstitions such as 'knock on wood' are quintessential examples of Threat narratives.

## 5.4 False Narratives

The results of the ELICIT and NATURAL treatments demonstrate how misspecified models of the world can arise, be transmitted as narratives, and mislead both the sender and receiver, all with no malicious intent. In light of these results, it is perhaps not surprising that false narratives and conspiracy theories are so pervasive. In fact, there is a sense in which we may underestimate the problem. In our experiment, narratives and statistical information are exogenously assigned. If, as Bursztyn et al. (2022) find, people prefer opinion programs to straight news, people may select into hearing misleading narratives over statistics, further exacerbating the problem.

This finding obviously has troubling implications, raising the question as to what can be done to counteract the effects of false narratives. On the receiving side, we considered several possibilities, but showed that Lever narratives are very robust, working even when they imply only a noisy relationship, when they point in a direction opposite to that of the true causal relationship in the data, and when competing against overwhelming statistical information that should invalidate the narrative.

The problem may be that subjects have difficulty falsifying the narrative because they do not realize that if the subjective belief the narrative gives them were true, the existing data should be different (i.e., they do not think through the counterfactual). If so, one possible means of killing off the effects of narratives may be to point out this counterfactual explicitly. It may also be interesting to allow for learning. Even though we made the joint distribution available to subjects, so that there is technically nothing to be learned, a literature in cognitive psychology has found some evidence that people better learn causal relationships when they make actual choices instead of simply observing data (see Waldmann and Hagmayer (2013) for a survey).

## 5.5 Limitations: Real-World Narratives

Using a setting in which we control, and make explicit, the DGP was necessary to isolate one mechanism through which causal narratives operate – by providing a causal interpretation of correlations in the data. It was also necessary to cleanly identify differences in the effectiveness of different types of narratives, holding everything else constant. But, we readily acknowledge that in the real world other factors likely come into play. In particular, we found that subjects strike a compromise between two competing narratives. While this seems plausible in real-world contexts, there the narrative that is adopted may also depend on the trustworthiness of its source, how well it accords with the receiver's 'knowledge of the world' (Pennington and Hastie (1993)), and what (potentially partial) correlations in the data the narrative activates in the receiver's memory (when full access to the data is not readily available). Similarly, though we found Threat narratives were not very effective, they may be more effective when the DGP is not explicit and the narrative can point to specific partial correlations (e.g., the Threat narrative about guns we gave in the introduction may be effective if it highlights cases in which a gun was useful in preventing crime). Finally, we showed that simple narratives are as effective as Lever narratives when correlations in the data are explicit. When they are not, Lever narratives might be more effective because they provide coverage (Pennington and Hastie (1993)) or justification (Bursztyn et al. (2023)).

We believe that studying how causal narratives work when the joint correlations in the data are known – as in our experiments – is critical for establishing a theoretical understanding of how they work, but future work should consider their efficacy in real-world settings where the correlations in the data may be either unknown or only partially known. Some recent empirical work has taken the first steps in this direction (Andre et al. (2022); Angrisani, Samek, and Serrano-Padial (2023); Espín-Sánchez, Gil-Guirado, and Ryan (2022); Goetzmann, Kim, and Shiller (2022)), and has developed methods to identify narratives using textual analysis (Ash, Gauthier and Widmer (2021); Lange et. al. (2022); Flynn and Sastry (2022); Hüning, Mechtenberg, and Wang (2022)).

## 5.6 Conclusion

Causal narratives are abundant and have potential impacts in politics, financial markets, macroeconomics, health, etc. We have provided some first evidence on what types of causal narratives are most impactful and under what conditions, but we strongly believe economics as a field would benefit from further research, both theoretical and empirical.

# References

[1] Aina, Chiara. 2022. “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. 2022. “Narratives About the Macroeconomy.” working paper.

[4] Angrasani, Marco, Anya Samek, and Ricardo Serrano-Padial. 2023. “Competing Narrative in Action: An Empirical Analysis of Model Adoption Dynamics.” working paper.

[5] Ash, Elliott, Germain Gauthier, and Philine Widmer. 2022. “RELATIO: Text Semantics Capture Political and Economic Narratives.” working paper.

[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).

[7] Barron, Kai and Tilman Fries. 2022. “Narrative Persuasion.” working paper.

[8] Benabou, Roland, Armin Falk, and Jean Tirole. 2018. “Narrative, Imperatives, and Moral Reasoning.” working paper.

[9] Bursztyn, Leonardo, Aakaash Rao, Christopher Roth, and David Yanagizawa-Drott. 2022. “Opinions as Facts.” *The Review of Economic Studies*, forthcoming.

[10] Bursztyn, Leonardo, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth. 2023. “Justifying Dissent.” *Quarterly Journal of Economics*, 138 (3): 1403-1451.

[11] Chapman, Loren. 1967. “Illusory Correlation in Observational Report.” *Journal of Verbal Learning and Verbal Behavior*, 6 (1): 151-155.

[12] 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.

[13] Danz, Daniel, Lise Vesterlund, and Alistair Wilson. 2022. “Belief Elicitation and Behavioral Incentive Compatibility.” *American Economic Review*, 112 (9): 2851-2883.

[14] Eliaz, Kfir and Ran Spiegler. 2020. “A Model of Competing Narratives.” *American Economic Review*, 110 (12): 3786-3816.

[15] Eliaz, Kfir, Simone Galperti, and Ran Spiegler. 2022. "False Narratives and Political Mobilization." working paper.

[16] Enke, Benjamin. 2020. "What You See is All There is." *Quarterly Journal of Economics*, 135 (3): 1363-1398.

[17] Enke, Benjamin and Thomas Graeber. 2023. "Cognitive Uncertainty." *Quarterly Journal of Economics*, forthcoming.

[18] Espín-Sánchez, José-Antonio, Salvador Gil-Guirado, and Nicholas Ryan. 2022. "Praying for Rain: On the Instrumentality of Religious Belief." working paper.

[19] Esponda, Ignaçio, Emanuel Vespa, and Sevgi Yuksel. 2021. "Mental Models and Learning: the Case of Base-Rate Neglect." working paper.

[20] Flynn, Joel P. and Karthik A. Sastry. 2022. "The Macroeconomics of Narratives." working paper.

[21] Frechette, Guillaume, Emanuel Vespa, and Sevgi Yuksel. 2023. "Extracting Models From Data Sets: An Experiment Using Notes-to-Self". working paper.

[22] Fryer, Bronwyn. 2003. "Storytelling That Moves People." *Harvard Business Review.*

[23] Goetzmann, William N., Dasol Kim, and Robert Shiller. 2022. "Crash narratives." working paper.

[24] Glazer, Jacob and Ariel Rubinstein. 2021. "Story Builders." *Journal of Economic Theory*, 193.

[25] Graeber, Thomas. 2023. "Inattentive Inference." *Journal of the European Economic Association*, 21(2):560-592.

[26] Graeber, Thomas, Christopher Roth, and Florian Zimmerman. 2024. "Stories, Statistics, and Memory." working paper.

[27] Hüning, Hendrik, Lydia Mechtenberg, and Stephanie Wang. 2022. "Using Arguments to Persuade: Experimental Evidence." working paper.

[28] Izzo, Federica, Gregory J. Martin and Steven Callander. 2021. "Ideological Competition." working paper.

[29] Jenni, Karen E. and George Loewenstein. 1997. "Explaining the 'Identifiable Victim Effect'." *Journal of Risk and Uncertainty*, 14, 235-257.

[30] Kamenica, Emir and Matthew Gentzkow. 2011. "Bayesian persuasion." *American Economic Review*, 101 (6): 2590-2615.

[31] Kendall, Chad and Ryan Oprea. 2022. "On the Complexity of Forming Mental Models." *Quantitative Economics*, 15(1): 175-211.

[32] Lange, Kai-Robin, Matthis Reccius, Tobias Schmidt, Henrik Müller, Michael Roos, and Carsten Jentsch. 2022. "Towards Extracting Collective Economic Narratives from Texts." working paper.

[33] Langer, Ellen J. 1975. "The Illusion of Control." *Journal of Personality and Social Psychology*, 32(2), 311–328.

[34] 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.

[35] Morag, Dor and George Loewenstein. 2021. "Narratives and Valuations." working paper.

[36] Oprea, Ryan. 2020. "What Makes a Rule Complex?" *American Economic Review*, 110 (12):3913-3951.

[37] Pearl, Judea. 1985. "Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning." In *Proceedings, Cognitive Science Society*, 329-334.

[38] Pearl, Judea. 2009. "Causality: Models, Reasoning, and Inference." Cambridge University Press.

[39] Pennington, Nancy and Reid Hastie. 1993. "Reasoning in Explanation-Based Decision Making." *Cognition*, 123-163.

[40] Quesenberry, Keith and Michael Coolsen. 2014. "What Makes a Super Bowl Ad Super? Five-Act Dramatic Form Affects Consumer Super Bowl Advertising Ratings." *The Journal of Marketing Theory and Practice*, 22(4):437-454.

[41] Rabin, Matthew. 2002. "Inference by the Believers in the Law of Small Numbers." *The Quarterly Journal of Economics*, 117 (3), 775-816.

[42] Schotter, Andrew. 2023. "Advice, Social Learning, and the Evolution of Conventions." Cambridge University Press.

[43] Schwartzstein, Joshua and Adi Sunderam. 2021. "Using Models to Persuade." *American Economic Review*, 111 (1): 276-323.

[44] Shermer, Michael. 2008. "Patternicity: Finding Meaningful Patterns in Meaningless Noise". *Scientific American*, 299 (6): 48.

[45] Shiller, Robert. 2017. "Narrative Economics." *American Economic Review*, 107 (4): 967-1004.

[46] Shiller, Robert. 2019. "Narrative Economics: How Stories Go Viral and Drive Major Economic Events." Princeton University Press.

[47] Sloman, Steven. 2009. "Causal Models: How People Think About the World." Oxford University Press.

[48] Song, Hayoung, Emily S. Finn, and Monica D. Rosenberg. 2021, "Neural Signatures of Attentional Engagement During Narratives and its Consequences for Memory." *PNAS*, 118 (33).

[49] Spiegler, Ran. 2016. "Bayesian Networks and Boundedly Rational Expectations." *Quarterly Journal of Economics*, 131 (3): 1243-1290.

[50] Stone, Deborah A. 1989. "Causal Stories and the Formation of Policy Agendas", *Political Science Quarterly*, 104 (2): 281-300.

[51] Vespa, Emanuel and Alistair J. Wilson. 2016. "Communication with Multiple Senders: An Experiment", *Quantitative Economics*, 7: 1-36.

[52] Vrantsidis, Thalia H. and Tania Lombrozo. 2022. "Simplicity as a Cue to Probability: Multiple Roles for Simplicity in Evaluating Explanations." *Cognitive Science*. 46 (7).

[53] Waldmann, Michael and York Hagmayer. 2013. "Causal Reasoning." *The Oxford Handbook of Cognitive Psychology*. Oxford University Press.

[54] Wallentin, Mikkel, Andreas H$\phi$jlund Nielsen, Peter Vuust, Anders Dohn, Andreas Roepstorff, Torben Ellegaard Lund. 2011. "Amygdala and Heart Rate Variability Responses from Listening to Emotionally Intense Parts of a Story." *NeuroImage*, 58 (3):963-973

# Online Appendix

# A Additional Figures and Tables

Table A1: Policy Predictions

| | $I^+$ | $I^{NOISE}$ | $I^{NEU}$ | $C^+$ | $C^{NOISE}$ | $C^{NEU}$ |
|---|---|---|---|---|---|---|
| Rational | 0.5 | 0.5 | 0.5 | 0.25 | 0.25 | 0.25 |
| Lever | 0.75 | 0.62 | 0.5 | 0.65 | 0.53 | 0.5 |
| Threat | [0.22,0.41] | 0.35 | 0.5 | [0.08,0.21] | 0.21 | 0.25 |
| Simple Up | 0.5 | 0.5 | 0.5 | 0.25 | 0.25 | 0.25 |
| Simple Down | 0.5 | 0.5 | 0.5 | 0.25 | 0.25 | 0.25 |

Notes: Predicted policy for each dataset (column) and narrative (row). For the Threat 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

| | $I^+$ | $I^{NOISE}$ | $I^{NEU}$ | $C^+$ | $C^{NOISE}$ | $C^{NEU}$ |
|---|---|---|---|---|---|---|
| Rational | 0.5 | 0.5 | 0.5 | 0.54 | 0.54 | 0.54 |
| Lever | 0.54 | 0.51 | 0.5 | 0.52 | 0.5 | 0.5 |
| Threat | [0.32,0.49] | 0.52 | 0.5 | [0.42,0.56] | 0.56 | 0.54 |

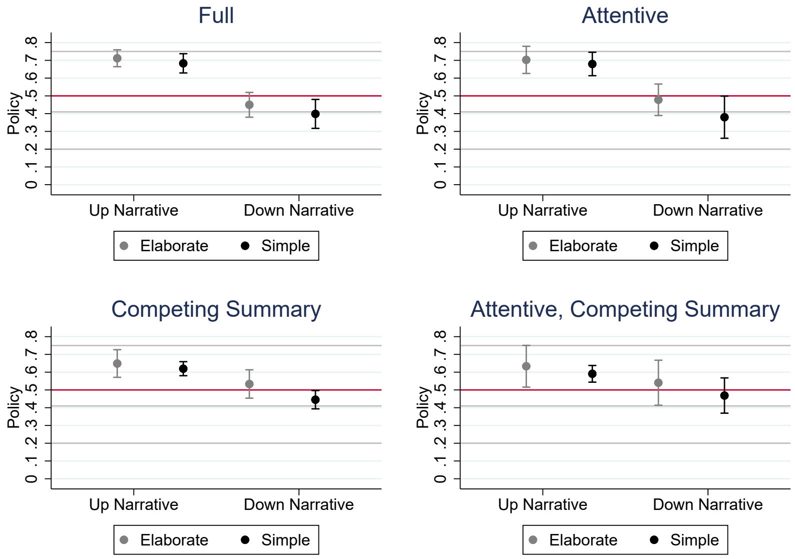
Notes: Anticipatory utility for each dataset (column) and narrative (row), calculated using the expected utility equation from the main text and the subjective beliefs under each narrative. For Threat narratives in the absence of noise, a range of utilities is predicted because beliefs are not completely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

Table A3: Elicited Narratives – Detailed

| Classification | $I^+$ | $I^{NEU}$ | $C^+$ | $C^{NEU}$ |
|---|---|---|---|---|
| Blue High | 10 | 11 | 2.5 | 2.5 |
| Blue Lever | 14.5 | 1.5 | 7 | 0.5 |
| Blue Threat | 0 | 0 | 0 | 0 |
| Blue Other | 1.5 | 7 | 3 | 1 |
| Green High | 5.5 | 2 | 46.5 | 47 |
| Green Lever | 0 | 0.5 | 0.5 | 0.5 |
| Green Threat | 2 | 0 | 3.5 | 0 |
| Green Other | 3 | 3 | 2.5 | 6 |
| Neutral | 31 | 40 | 5 | 10 |
| Rational | 6 | 8 | 8 | 8.5 |
| Blue Lever / Green High | 0 | 0 | 1.5 | 0 |
| Blue Lever / Green Threat | 1 | 0 | 0.5 | 0 |
| Reject | 25.5 | 27 | 19.5 | 24 |

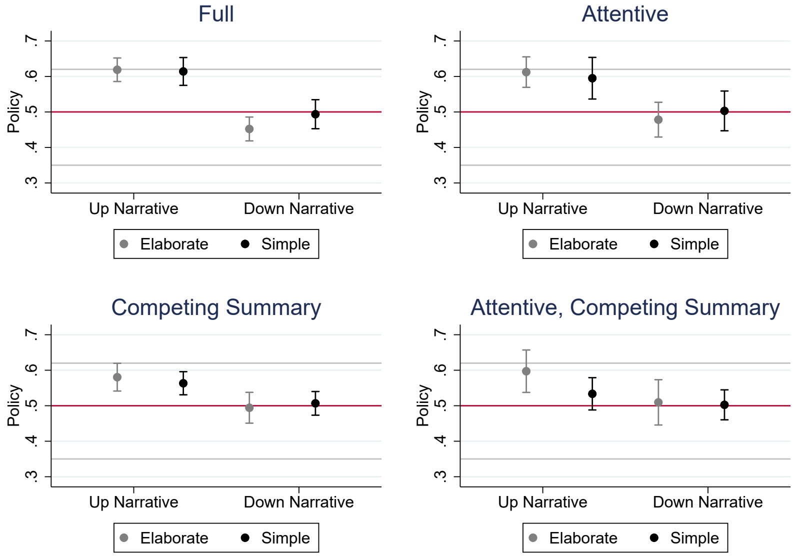
Notes: Classification of elicited narratives (percentages) in each dataset. Blue High and Green High suggest that the corresponding color leads to HIGH ($y = 1$) outcomes more often. Blue Other and Green Other recommend the corresponding color, but do not provide a particular reason. Rational recommends 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 3, we combined the advice into broader categories as follows. For all datasets, we combined Blue High and Blue Other into Simple 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 Simple Down, and relabeled Green Lever as Other. For the causal datasets, we combined Green High and Rational into Rational, relabeled Green Other as Simple Down, and combined Neutral and Green Lever into Other.

Figure A1: $I^{+}$ Dataset Average Policies (First Dataset Only)



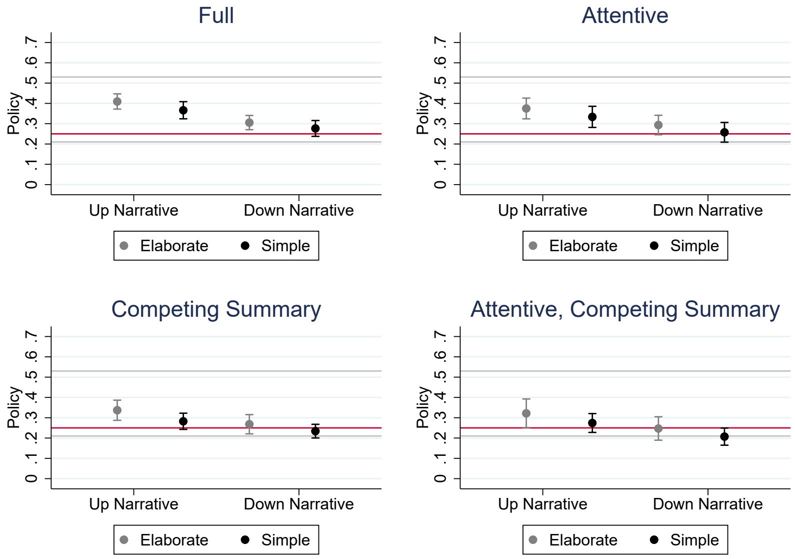
Notes: Average policy choices and 95 percent confidence intervals. We restrict the data to subjects who saw the $I^{+}$ dataset first. The upper left panel is for all subjects. The upper right panel restricts to attentive subjects that do not respond to inconsistent narratives. The lower left panel is for policy choices when the narrative competed directly with a summary. The lower right panel is for policy choices of attentive subjects when the narrative competed directly with a summary. The red line indicates the rational prediction while the gray lines indicate the predictions of the BNFF for the Lever narrative (upper) and the Threat narrative (lower two lines indicate the predicted range).

Figure A2: $I^{NOISE}$ Dataset Average Policies



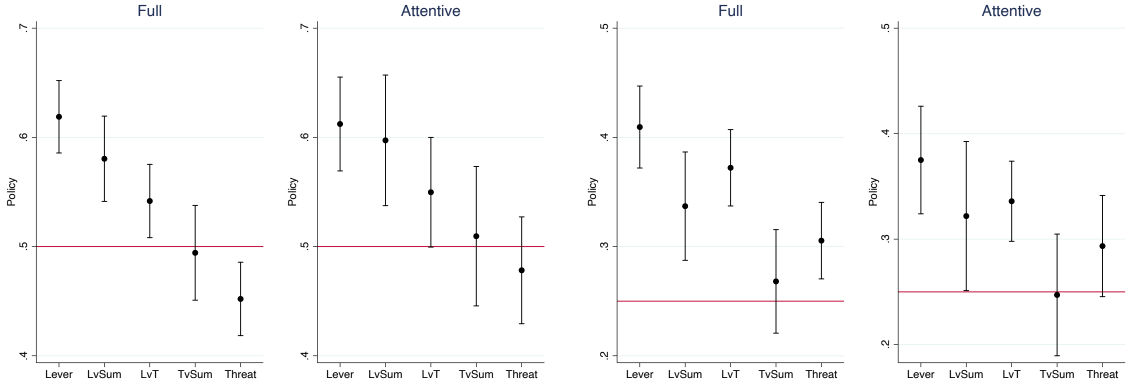
Notes: Average policy choices and 95 percent confidence intervals. The upper left panel is for all subjects. The upper right panel restricts to attentive subjects that do not respond to inconsistent narratives. The lower left panel is for policy choices when the narrative competed directly with a summary. The lower right panel is for policy choices of attentive subjects when the narrative competed directly with a summary. The red line indicates the rational prediction while the gray lines indicate the predictions of the Bayesian-network factorization formula for the Lever narrative (upper) and the Threat narrative (lower).

Figure A3: $C^{NOISE}$ Dataset Average Policies



Notes: Average policy choices and 95 percent confidence intervals. The upper left panel is for all subjects. The upper right panel restricts to attentive subjects that do not respond to inconsistent narratives. The lower left panel is for policy choices when the narrative competed directly with a summary. The lower right panel is for policy choices of attentive subjects when the narrative competed directly with a summary. The red line indicates the rational prediction while the gray lines indicate the predictions of the Bayesian-network factorization formula for the Lever narrative (upper) and the Threat narrative (lower).

Figure A4: Competing Narratives in $I^{NOISE}$ and $C^{NOISE}$



Notes: Average policy choices and 95 percent confidence intervals. LvSum indicates the average policy choice when subjects observed both the Lever narrative and the summary. TvSum indicates the Threat narrative and the summary. LvT indicates both Lever and Threat narratives. The left pair of graphs is for the $I^{NOISE}$ dataset and the right pair of graphs is for the $C^{NOISE}$ 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.

# B Additional Results

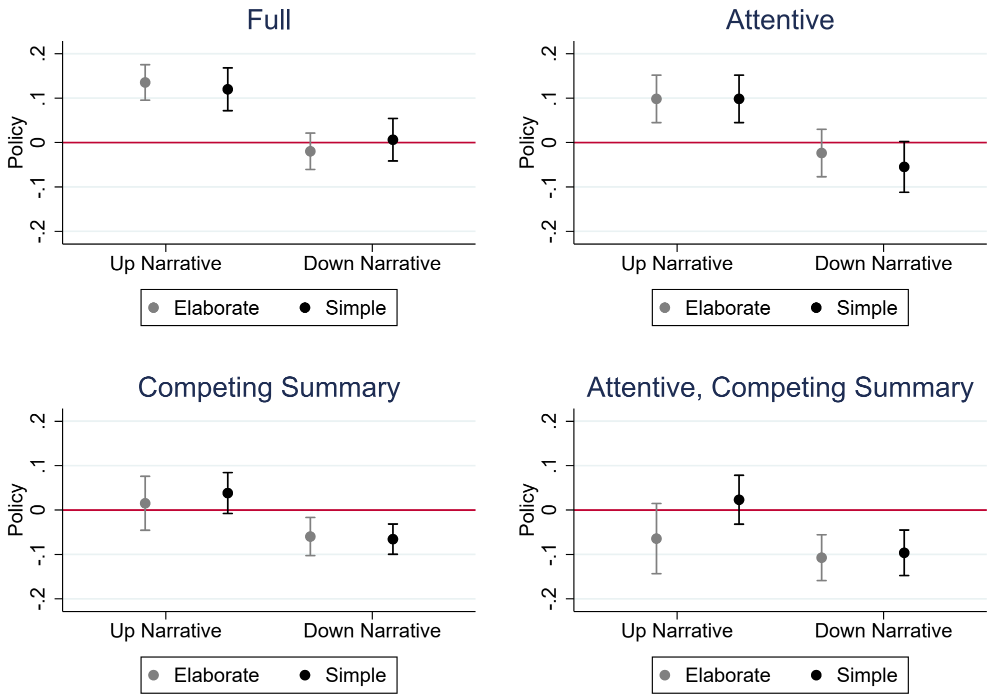
## B.1 Undershooting of BNFF Predictions

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 $C^{NEU}$ 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 in causal datasets: when we compare policies to the rational policy of 0.25, we might overstate the effects of Lever and Simple 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 $C^+$ when they observe either a Lever or a Simple Up narrative and their initial policy choices in $C^{NEU}$. The idea is that initial policies in $C^{NEU}$ already capture the compression towards 0.5; any remaining movement towards 0.5 in $C^+$ must be in response to the narrative. In Figure B1, we show that Lever and Simple Up narratives continue to have positive effects by this measure, except when subjects see a summary simultaneously.[27]

[27] The tests when comparing to a summary are particularly strict because they ignore the fact that had a subject seen a statistical summary when choosing their initial policy, the subject would likely have chosen a policy closer to the rational policy of 0.25 in response to the summary.

Figure B1: Policy Differences in CONSTRUCTED – Causal Datasets



Notes: Estimates of average differences in policy choices in $C^+$ after seeing a narrative and initial policy choices in $C^{NEU}$. Error bars indicate 95 percent confidence intervals with standard errors clustered at the subject level. The upper left panel is for all subjects. The upper right panel restricts to attentive subjects that do not respond to inconsistent narratives. The lower left panel is for policy choices when the narrative competed directly with a summary. The lower right panel is for policy choices of attentive subjects when the narrative competed directly with a summary.

A possible reason for the observed compression to the middle is cognitive uncertainty (Enke and Graeber (2023)).[28] 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 at the median reported certainty in policy choices in the $I^+$ and $I^{NOISE}$ datasets. We find that subjects who are more certain deviate more from 0.5 for Lever and Simple Up narratives. These differences are significant ($p < 0.05$) except in the case of the Lever narrative in $I^{NOISE}$ ($p = 0.547$). On the other hand, we find no robustly significant differences for the Threat and Simple Down narratives and the point predictions often go in the opposite direction, with more certain subjects being closer to 0.5. In the causal

[28] Risk aversion doesn't straightforwardly result in compression because choosing an extreme policy reduces variability in the outcome.

datasets, subjects who are more certain are closer to the rational policy of 0.25 in their initial choices in $C^{NEU}$ ($p = 0.078$).

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 Threat and Simple Down narratives, 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 Lever and Simple 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 Lever narrative might be more robust than the Threat 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 Lever 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 four datasets ($I^+$, $I^{NOISE}$, $C^+$, and $C^{NOISE}$) for all four types of narratives, with one exception. In the $I^+$ dataset with the Threat narrative, average policies are lower by 0.08 (i.e., the Threat narrative is more effective) when the column ordering is $a, y, z$, but the effect is only significant at the 10% level. Overall, column ordering does not seem to have large effects.

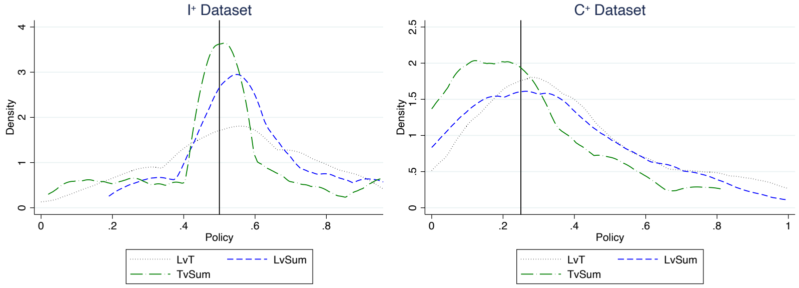
One plausible reason narratives may be effective when they compete with summaries is that subjects may anchor their choices to the first narrative they see. To test for this possibility, we make use of the fact that some subjects see the Lever narrative and then both the Lever and Threat narratives, while others see the Threat narrative first. We regress choices when subjects see both elaborate narratives on a dummy that indicates that they saw the Lever 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: $I^+ : 0.06$ ($p = 0.103$),

$I^{NOISE}$: 0.02 ($p = 0.621$) $C^+$ : 0.04 ($p = 0.291$), and $C^{NOISE}$: 0.05 ($p = 0.165$). Thus, although each of the point estimates is consistent with anchoring, none of the results are significant at the five percent level. The lack of significance is unlikely due to a lack of power because we have about 150 observations in each dataset, so, while we can't rule out some amount of anchoring, we conclude that it is at most of second-order importance.

## B.3 Further Results on Competing Narratives

Because Figure 6 plots average policies across subjects, the averages could reflect some subjects following one narrative and others following the other. However, the distributions of policy choices shown in Figure B2 largely rule out this hypothesis. Each of these distributions is fairly unimodal, suggesting that many individual subjects appear to be performing some kind of averaging of competing narratives. We can also rule out simple confusion. Subjects do not simply choose the least costly policy when confronted with conflicting information: policy choices when subjects face both Lever and Threat narratives are statistically different from 0.5 using the full sample of subjects in both of the $I^+$ and $C^+$ datasets. Furthermore, in Figure B3 we plot subjects' confidence in their policies. We find that narratives weakly increase average confidence, both when subjects see one narrative on its own as well as when they see competing narratives.

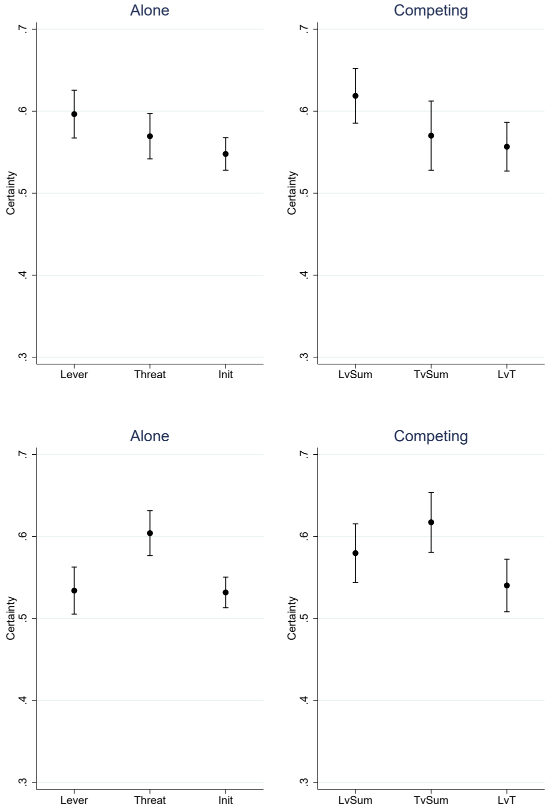
Figure B2: Policy Densities with Competing Narratives



Notes: Kernel densities of policy choices. LvSum indicates the average policy choice when subjects observed both the Lever narrative and the summary. TvSum indicates the Threat narrative and the summary. LvT indicates both Lever and Threat narratives. The left graph is for the $I^+$ dataset and the right is for the $C^+$ dataset.

We conclude that subjects combine the two models provided by the narratives

Figure B3: Confidence



Notes: Average subject confidence in their policy choices, pooled across $I^+$ and $I^{NOISE}$ (upper panels) and $C^+$ and $C^{NOISE}$ (lower panels). The error bars indicate 95 percent confidence intervals. The left panel in each pair is for initial choices (Init) and after observing a single narrative. The right panel in each pair is for narratives that compete with a summary (LvSum and TvSum) or for competing Lever and Threat narratives (LvT).

in some sophisticated way.[29] Although intuitive, such behavior is markedly different from the assumption made in recent theoretical work that people adopt one narrative or the other (Eliaz and Spiegler (2020), Schwartzstein and Sunderam (2021)).

One possibility is that subjects form beliefs separately for each narrative and then average them before making their choices. To see how this might work, suppose one is willing to assume all subjects use the same weight, so that average policy choices when facing both the Lever and the Threat narrative reflect a weighted average of the average policies under each narrative on its own. Then, for both the $I^+$ and $C^+$ datasets, the required weight on the Lever narrative is almost exactly one-half

[29] Vespa and Wilson (2016) similarly find that subjects average the recommendations of two senders in a communication game even when it is not optimal.

(0.53 and 0.51, respectively). Of course, assuming homogeneous weights is a strong assumption, but, unfortunately, we can't calculate weights at the individual level because each subject only sees either the Lever or the Threat narrative on its own.

While such 'averaging' has a Bayesian feel – subjects assign a uniform prior to the models implied by the two narratives and combine the two models using this prior – the behavior cannot be truly Bayesian. Consider the case in which the Lever narrative competes with the summary. A Bayesian would form a posterior about the likelihood that each model is correct using the data available in the dataset and thus reject the Lever narrative in favor of the summary. Instead, our results suggest that subjects adopt a somewhat sophisticated, albeit imperfect, approach to combining narratives.

# 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 experiments, 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 experiment in the main text differs from the experiment 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 experiment reported in the main text. The treatments in the 2022 version were very similar to the CONSTRUCTED, 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 labeled slightly differently compared to the current experiments. Specifically, the "positive" dataset corresponds to the $I^+$ dataset, the "neutral" dataset corresponds to the $I^{NEU}$ dataset, and the "negative" dataset corresponds to a dataset that is symmetric to $I^+$, except that it swaps $a = 0$ and $a = 1$. This effectively yields

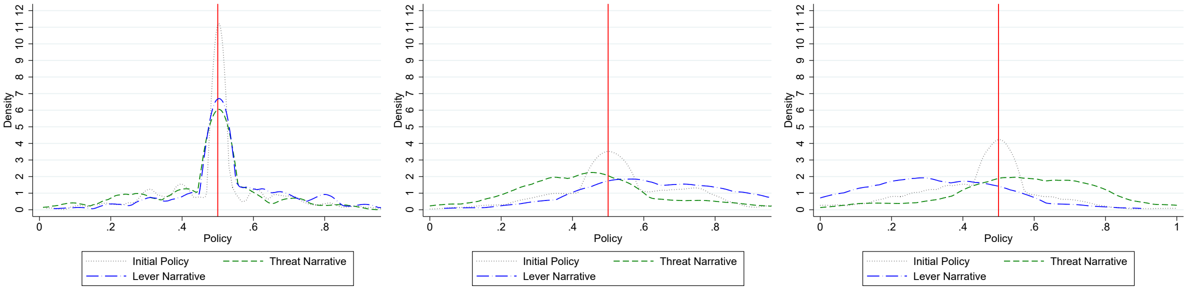
an $I^-$ dataset, one in which the Lever narrative supports a downward deviation from the rational policy of 0.5, and the Threat narrative supports an upward deviation.

4. Subjects only saw elaborate narratives and statistical summaries (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 a statistical summary (in randomized order). These 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 CONSTRUCTED

Figure C1 shows kernel density estimates of the policies subjects chose after seeing either the Lever or the Threat narrative along with their initial policy choices, for each dataset. The distributions of policies in the neutral dataset are quite tight, regardless of the type of narrative, indicating that subjects do not respond strongly to inconsistent narratives. In contrast, we see much larger movements for narratives in the positive and negative datasets.

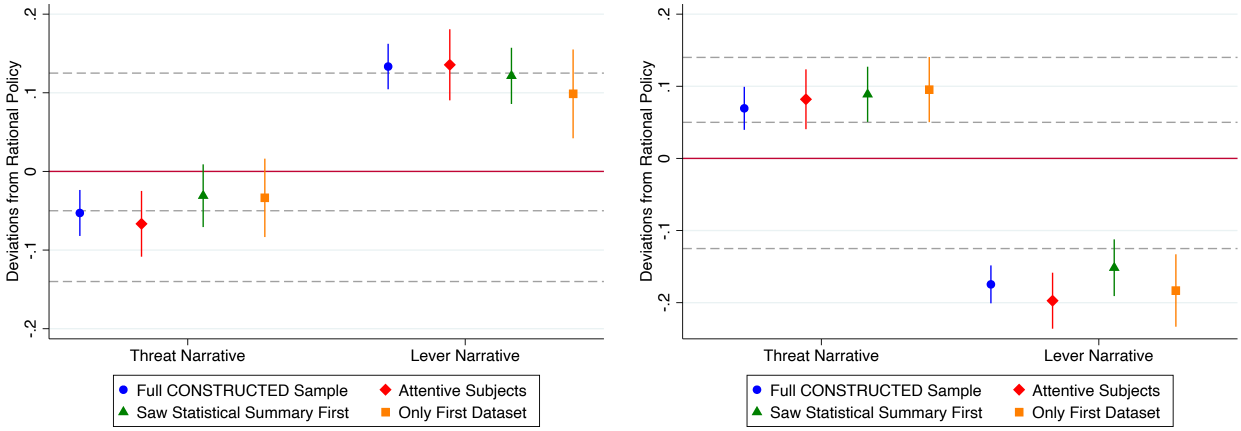
Figure C1: Policy Densities in CONSTRUCTED



Notes: Kernel densities of policy choices. The left graph is for the $I^{NEU}$ dataset, the middle is for the $I^+$ dataset, and the right is for the $I^-$ dataset.

Figure C2 plots deviations from the rational policy for several subsets of the data. The dashed gray lines in the graphs indicate the predictions of the Bayesian-network factorization formula (BNFF) for each type of narrative. Recall that the BNFF gives a point prediction for Lever narratives, while it only gives a range of predictions for Threat narratives. We find that, for the most part, subjects' policies are remarkably close to the predictions of the BNFF.

Figure C2: Deviations from Rational in CONSTRUCTED



Notes: Average policy deviations from the rational policy (0.5). Error bars indicate 95 percent confidence intervals. The left graph is for the $I^+$ dataset, and the right is for the $I^-$ dataset.

We would like to highlight that across the positive and negative datasets, the movements in response to narratives are mirror images of each other. Specifically, in Figure C1, the upward shift in mass in response to the Lever narrative in the positive dataset is mirrored by a downward shift in mass in the negative dataset (and vice versa for Threat narratives). Similarly, in Figure C2, 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 $I^+$ dataset and omit the $I^-$ dataset in the experiment reported in the main text.

In contrast to the experiment reported in the main text, subjects in our prior experiment saw statistical summaries for each dataset in isolation (i.e., without competing narratives). This allows us to check which policies subjects choose when they see only a statistical summary of the data. We find that statistical information moves policy choices very close to the rational policy of 0.5: average policy choices after observing the statistical summary are 0.53, 0.47, and 0.51, in the positive, negative, and neutral datasets, respectively (not shown in a figure). The finding that subjects choose rational policies when provided with only a statistical summary motivated us to omit isolated responses to statistical summaries from our current experiment, and to instead focus on responses to narratives. 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 datasets with noise.

## C.2 ELICIT

Similar to the treatment reported in the main text, subjects in our prior ELICIT treatment observed the positive, negative, and neutral datasets in randomized order. For each dataset, they gave free-from 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 positive or negative datasets, we classified 18% as elaborate narratives, 51% as simple narratives, and 31% as neutral narratives.[30] Of the 18% elaborate narratives that subjects identified, the vast majority (89%) are Lever narratives, providing further support for the result in the main text that subjects find it easier to identify Lever 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 neutral 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 negative datasets, 25% are elaborate narratives (all Lever narratives), 55% are simple narratives, and 20% are neutral 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.

[30] In the 2022 version, we labeled these categories slightly differently. Specifically, elaborate narratives were labeled as "causal" narratives and simple narratives as "other" narratives.

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 Lever | GL | Suggests green (policy $< 0.5$) and describes pattern for Lever narrative |
| Green Threat | 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 Lever | BL | Suggests blue (policy $> 0.5$) and describes pattern for Lever narrative |
| Blue Threat | 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 |

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).