# Causal Narratives

Constantin Charles and Chad Kendall*

July 27, 2023

**Abstract**

We study causal narratives – narratives which describe a (potentially incorrect) causal relationship between variables. In a series of controlled experiments across a range of data-generating processes, we show that exogenously generated causal narratives manipulate the beliefs and actions of subjects in ways generally consistent with theory, but with important exceptions, including when subjects face multiple narratives that give different recommendations. To understand the generation and social transmission of narratives, we show that causal narratives arise endogenously when subjects who observe a dataset provide advice to future subjects. Homegrown causal narratives are perceived to be helpful, but 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 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 manipulate beliefs is critical for understanding how people come to form opinions on a vast range of topics.

*Constantin Charles: Department of Finance, London School of Economics (e-mail constantin.charles.phd@marshall.usc.edu). 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, Alexander Coutts, Kfir Eliaz, Cary Frydman, David Hirshleifer, Ryan Oprea, Ran Spiegler, and seminar participants at the BRIQ Beliefs Conference, ESA World Meetings, LSE Behavioral Political Economy Conference, UCI Finance Conference, Online Seminar in Economics and Data Science at ETH Zurich, Carnegie Mellon University, Texas A&M, University of Ottawa, University of Toronto, and USC for valuable input.

Although the potential importance of causal narratives was recognized at least three decades ago (Stone (1989)), only recently have economic theorists found a way to incorporate them into economic models (Spiegler (2016); Eliaz and Spiegler (2020); Eliaz, Galperti, and Spiegler (2022)). These theoretical contributions are important because, if the assumptions about how causal narratives work are correct, they provide a tractable means of incorporating narratives into economic models, allowing us to understand their impacts on economic issues from financial bubbles to political polarization (Shiller (2017, 2019)).

In this paper, we test these assumptions in a series of controlled experiments, but also use the experiments to more broadly understand how these types of narratives arise, and in what types of environments they are effective at manipulating peoples' beliefs and actions. Controlled experiments are ideal for answering these questions, because they allow us to control both the true data-generating process (DGP) as well as the narratives that subjects are exposed to.

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 though 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 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 does have a causal effect on the outcome. Thus, our baseline environment closely mirrors the social media and depression example, except that it is free of any context. In the example, and most other everyday situations, people will likely have preconceptions about how the variables are related even before hearing any narrative, but these preconceptions potentially confound an experiment because they are not controlled by the researcher. To avoid this problem, we use generic terms (i.e., "choice" and "payoff").

In both of our environments, subjects observe several compact datasets in sequence, each consisting of three binary variables: an action, an outcome, and an auxiliary variable. 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 treatments in which we construct the narratives, we generate *elaborate* narratives that suggest a causal story by leveraging correlations in the dataset to point out specific patterns. These elaborate narratives come in two distinct types. The first is what Eliaz and Spiegler (2020) call a *Lever* narrative: it implies a causal chain from action, to auxiliary variable, to outcome, as in the depression narrative. The second is what Eliaz and Spiegler (2020) call 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, so that we need to arm citizens to counteract this threat. The DAG in this case is $arm\ citizens \rightarrow public\ safety \leftarrow criminals$. In addition to the elaborate narratives, we also construct *simple* narratives, which simply recommend choosing one action or the other more often.

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. Thus, 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 two simple narratives, which only vary in terms of which action is recommended, the formula predicts no impact on beliefs. For the Lever narrative, it predicts beliefs will be

distorted so that one action is 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.

To understand how two narratives can lead to opposite effects for the same dataset, recall the gun example above. The Threat narrative supports arming citizens to counteract the threat of armed criminals. In contrast, a Lever narrative with the corresponding DAG, $arm\ citizens \rightarrow criminals \rightarrow public\,safety$, argues that if citizens are not armed, criminals won't arm themselves either, resulting in an overall improvement in public safety.

In our baseline environment, in which actions and outcomes are unrelated, we find that the Lever and Threat narratives cause changes in policy choices in the directions predicted by the Bayesian-network factorization formula, but of somewhat smaller magnitudes than predicted. More surprisingly, the simple narrative that recommends the same action as the Lever narrative has almost as big of an impact on policies as the Lever narrative itself, though the Bayesian-network factorization formula predicts no change. We show that the reason for this is that subjects pick up on the pattern described by the Lever narrative on their own, especially when nudged in this direction by the simple narrative.

In the second environment, in which the action does have a causal effect on the outcome, we find that the Lever narrative (and the simple narrative that recommends the same action) continue to distort policy choices, even though these narratives recommend the action that actually produces the good outcome *less* often. Thus, these narratives can be effective even when they oppose a true causal relationship. On the other hand, in this environment, the Threat narrative has only small effects. This finding shows that Lever narratives are more robust than Threat narratives, suggesting that some qualitative difference between the two, one that is not captured by the Bayesian-network factorization formula, is important for the efficacy of narratives. We discuss several possibilities in the concluding section of the paper.

We also test the robustness of the effects of narratives to possible inattention by subjects, leveraging variation in the auxiliary variable across datasets. Specifically, when the auxiliary variable is independent of the action and outcome, the Lever narrative is easily falsified by looking at the dataset, so that any subject that changes their policy choice in the direction of the narrative is likely not paying attention. In our robustness checks, we filter out these inattentive subjects. We show that the Lever narrative, as well as the simple narrative that gives the same recommendation as the Lever narrative, remain effective in both environments, ruling out inattention as the driver of the results.

In both environments, after subjects view a narrative on its own, we provide them with ei-

ther (i) a second, competing narrative that recommends the opposite policy or (ii) a summary that describes the true relationship between actions and outcomes and explicitly recommends the rational policy. When confronted with two competing narratives, subjects choose policies that lie between the policies they chose 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, rather than adopting 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 lies in contrast to the assumption made in theoretical models of competing narratives. In these theories, people are assumed to adopt one of the competing narratives according to some criterion (i.e., highest expected utility under subjective beliefs in Eliaz and Spiegler (2020) and best fit to the data in Schwartzstein and Sunderam (2021)).

Having shown that causal narratives are effective, we ask whether or not they can arise endogenously when people observe correlations in the data. To answer this question, in our second set of treatments we again provide subjects with datasets and ask them to choose policies, but instead of providing them with narratives, we incentivize them to construct their own. We ask them 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 some subjects produce elaborate narratives both in our baseline environment (in which actions and outcomes are unrelated) as well as 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 generated are Lever narratives, consistent with the previous evidence that subjects seem to pick up on the Lever narrative patterns 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 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), which argues compellingly that political actors deliberately associate events with *causal* stories in order to shape the political agenda and motivate partisan support for their

side. Within economics, narratives have recently begun to receive increased attention (Shiller (2017, 2019)). A growing literature has made important contributions in providing ways to think about narratives theoretically. For our purposes, Eliaz and Spiegler (2020), building on Spiegler (2016), is critical, as we test key assumptions of their innovative conceptual framework which represents narratives as causal graphs that weave in auxiliary variables. Our findings generally provide support for the use of the Bayesian-network factorization formula as a modeling device, but also point out subtleties in the types of narratives that are persuasive, and suggest that it may be fruitful to consider alternative ways of modeling competing narratives.

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 models explicitly, our experiment provides some of the first available evidence (together with Barron and Fries (2023)) that persuasion via models (as opposed to signals or Bayesian persuasion experiments (Kamenica and Gentzkow (2011)) can be effective.

On the empirical side, Andre et al. (2022) surveys people about the causes of recent inflation, maps their responses to DAGs, and tests the power of narratives to influence (self-reported) inflation expectations. The paper complements ours in that it demonstrates that people generate causal narratives and can be influenced by these narratives in real-world settings. On the other hand, our control over the DGP allows us to closely associate the narratives with the DGP in order to more tightly engage with theory.

A handful of recent experimental papers study narratives from different perspectives. Morag and Loewenstein (2021) shows 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 tests persuasion via narratives using the theoretical framework of Schwartzstein and Sunderam (2021). Graeber, Roth, and Zimmerman (2022) shows that stories are easier to recall than statistics, leading to larger impacts on beliefs. Our paper complements this work by focusing on narratives that convey causal stories.

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)). Our work complements this literature by demonstrating that narratives can have power not only because of emotional responses or because people relate to them, but because they create a lens through which people interpret data causally.

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

Cognitive scientists have tackled the important question of how causal and statistical processes differ formally (Pearl (2009); Sloman (2009)), and also conducted experiments to study how people perceive (and misperceive) causal relationships (see Waldmann and Hagmayer (2013) and Matute et al. (2015) for recent reviews). This literature has generally converged on causal Bayes networks as the best normative model of behavior (Sloman and Lagnado (2015)), but finds some departures in behavior from its predictions, as do we. 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). Most of this literature focuses on factors that allow people to identify causal relationships in two-variable environments. Steyvers et al. (2003) studies causal chains like we do, but focuses on how people determine the direction of causality whereas in our setting, the only possible causal relationship is from action to outcome.

The psychology literature on illusory correlation (Chapman (1967)) and apophenia / patternicity (Conrad (1958); Shermer (2008)), which identifies instances in which people believe correlational or causal relationships exist when they do not (generally from visual stimuli such as images), is related to our findings that people endogenously generate causal narratives. Perhaps the most closely related offshoot of this literature is work on the hot hand and gambler's fallacies, which are misperceptions of correlations in statistically independent events (Rabin (2002); Asparouhova, Hertzel, and Lemmon (2009)). Unlike in these settings, where the correlations are imagined, in our setting narratives leverage actual correlations to tell a causal story.

Finally, given that casual 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 (Kendall and Oprea (2022); Esponda, Vespa, and Yuksel (2021); Graeber (2023); Enke (2020)).

# 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 $y$. In the right 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 $y$.

($I^{NEU}$), $z$ 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 $p(y|a)$, and knows 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, $p(a, z, y)$.

Of course, the statistical formula on its own does not pin down the *causal* effect of $a$ on $y$, a fact often conveyed by the maxim, 'correlation does not imply causation'. However, if one also knows that $a$ is exogenous, the law of total probability *is* sufficient to determine the causal effect of $a$ on $y$.[2] For example, for either of the datasets in Table 1, a rational DM would calculate $p(y = 1|a) = p(y = 1) = \frac{1}{2}$. That is, a rational DM would realize that the auxiliary variable is irrelevant in both datasets.

## 2.2 Narrative Examples

Suppose now a DM is presented with the following *Lever* narrative when studying the $I^{+}$ dataset:

[2]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.

*"$z = 1$ 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 $a = 1$. That is, she might believe $p(y = 1|a = 1) > p(y = 1|a = 0)$.

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 and can be verified in the data. But, unlike the first one, this narrative implies a different causal relationship: $z$ and $a$ both influence $y$ directly. A DM believing it may form the opposite belief, $p(y = 1|a = 1) < p(y = 1|a = 0)$. So, 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 narratives that suggest a causal relationship. Thus, causal narratives critically depend on the ability to exploit correlations in the data. Note, though, that the Lever and Threat narratives exploit the same correlation differently, by pointing to different patterns to imply different causal relationships.

## 2.3 Directed Acyclic Graphs and Beliefs

To discuss these narratives formally, we describe the causal relationships they imply as Bayesian networks using 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 beliefs of a DM faced with a joint probability distribution, and Eliaz and Spiegler (2020) used DAGs as a means of describing narratives.

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 causal relationships between variables precisely: the conditional probabilities can be calculated using the Bayesian-network factorization formula (BNFF). From the perspective of a DM, the BNFF provides a normative description of how the DM *should* form beliefs given knowledge of the joint distribution and belief in a causal model.

For example, consider the first narrative described in the previous section. This narrative

is an example of what we refer to as an *elaborate* narrative - it weaves the auxiliary variable into the narrative to imply a causal relationship. In this case, it leverages the auxiliary variable to imply the causal chain, $a \to z \to y$. For this reason, Eliaz and Spiegler (2020) refers to it as a *Lever* narrative. Here, the BNFF prescribes $p(y|a) = \sum_{z=0,1} p(y|z)p(z|a)$ (see Spiegler (2016) for the general form of the BNFF). When compared to the conditional law of total probability, the conditioning of $y$ on $a$ is dropped. This lack of conditioning can lead astray a DM that believes such a narrative.

For example, for the $I^+$ dataset in Table 1, the BNFF prediction for the Lever narrative is $p(y = 1|a = 1) = \frac{2}{3}$. A DM that adopts this narrative should believe that if she chooses $a = 1$, the desirable outcome, $y = 1$, will occur with probability two-thirds instead of the true probability of one-half. Because of the positive (upward) shift in beliefs for this dataset under the Lever narrative, we refer to $I^+$ as a 'positive' dataset.

The second example of a narrative above is another elaborate narrative, one which Eliaz and Spiegler (2020) refers to as a *Threat* narrative.[3] Here, $z$ 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 a DM astray. 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, $p(y|a) = \sum_{z=0,1} p(y|a)p(z)$. Compared to the conditional law of total probability in (1), the fact that $z$ is independent allows it to be 'factored out'. We will refer to such a narrative as a *simple* narrative, one that implies that the action is the sole determinant of the outcome. A DM that views either of the $I^+$ or $I^{NEU}$ datasets and believes the simple narrative should have rational beliefs, $p(y|a)$. In this case, belief in a causal relationship is of no consequence. 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 - a key prediction we test in the experiment.

As mentioned previously, 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.

[3]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 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.

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

$$\max_d d \cdot p_G(y = 1|a = 1) + (1 - d) \cdot p_G(y = 1|a = 0) - c(d - d^*)^2 \tag{3}$$

Note that 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 also has a more subtle indirect effect through learning - a change in $d$ will change the frequency of $a$ and therefore can affect the DM's beliefs through changes in the new data generated. We assume that the DM does not account for the indirect effect (an assumption we enforce in the experiment by not providing subjects with feedback about the realizations of the variables). Specifically, beliefs are a treated 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 (3) via the first-order condition, 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, $d^*$.

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 $p_G(y = 1|a = 0) = \frac{\gamma}{4} + \frac{3}{8}$.[4] 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.[5] 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 which features of causal narratives make them more convincing and more likely to be adopted 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 in the $I^+$ dataset, but inconsistent in 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.[6] But, even a non-Bayesian that believes $p_G(y = 1|a = 1) \lesseqgtr \frac{1}{2}$ for some narrative should ask themselves why they observe $p(y = 1|a = 1) = \frac{1}{2}$ in the dataset.

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

[5] 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.

[6] Formally, the probability distribution implied by the dataset is not compatible (Markov) with the DAGs corresponding to the Lever and Threat narratives.

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.[7] 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 induced by each narrative. This idea is related to illusion of control, but there is an important distinction: while illusion of control is binary (one can or cannot influence the outcome), anticipatory utility is a continuous measure. We calculate the anticipatory utilities for each dataset and narrative combination used in our experiment in Appendix A.

## 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 discuss the purposes for these datasets in Section 3.1. We refer to the $I$ datasets as independent datasets and the $C$ 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

[7] 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.

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 $p(y = 1|a = 0) = \frac{2}{3}$.

For the experiment, we must choose $d^*$ and $c$. We do so 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.[8] The choice of $c$ must 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. $c = \frac{2}{3}$ strikes a compromise, but ensures 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.

With these parameter choices, in the the $I^{+}$ dataset the optimal policies under the Lever and Threat narratives are $d = \frac{6}{8}$ and $d = \frac{13}{32} - \frac{3}{16}\gamma$, respectively. With $\gamma \in [0, 1]$, the

[8]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)).

Table 3. Policy Predictions under the BNFF

| | $I^+$ | $I^{NOISE}$ | $I^{NEU}$ | $C^+$ | $C^{NOISE}$ | $C^{NEU}$ |
|---|---|---|---|---|---|---|
| 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.

optimal policy under the Threat narrative is in the range, $d \in [\frac{7}{32}, \frac{13}{32}] \approx [0.22, 0.41]$, so that Lever and Threat narratives produce optimal policies on opposite sides of $d = \frac{1}{2}$. Table 3 summarizes the predictions for all narrative and dataset combinations. In addition to the two elaborate narratives (Lever and Threat), 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).

# 3 Testing Narratives

Our experiment is designed to implement the environment described in the previous section. 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 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 how subjects' policies change across 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 rows, while the causal datasets consisted of 12 rows. The datasets were described as summarizing thousands of historical observations

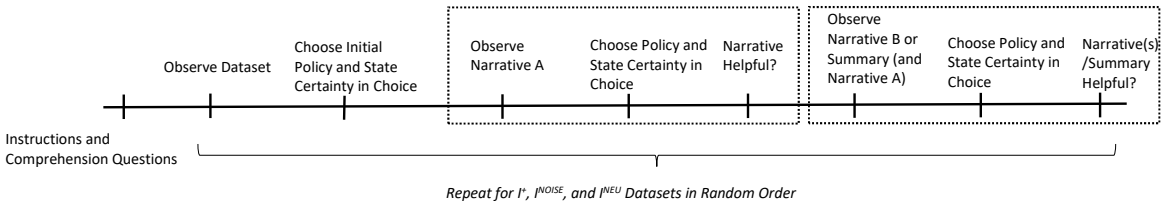
such that each row occurred an equal number of times. Subjects were also explicitly told that (i) the variables will maintain the same relationships (if any) they had in the past and (ii) no other 'hidden' variables influence the observed variables in any way. Figure 1 shows the timeline of the CONSTRUCTED treatment and we provide screenshots of all treatments in Appendix C. In both CONSTRUCTED treatment arms, for each dataset, subjects completed the following tasks in order:

1. We presented the dataset, telling subjects 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, we instead provided a second narrative on its own. In all cases, subjects could review the dataset and 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 numbered 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*

Figure 1. Timeline for CONSTRUCTED Treatments




*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 in $I^{NOISE}$ and $C^{NOISE}$, we also constructed "noisy" versions of the Lever and Threat narratives that are identical to the 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). This summary table can also be thought of as a 'narrative', one implying no causal relationship. We refer to the table as a summary to avoid confusion, but consider it to be a narrative when we discuss competing narratives in Section 3.2.5.

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.[9] 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

[9] 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.

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. Rather than referring to an 'action' and 'outcome' which could imply a causal relationship, we used 'choice' and 'payoff' which may be less suggestive. We labeled the auxiliary variable 'X' to avoid any preconceived relationship to the other variables. Finally, we labeled the variables neutrally: $a \in \{BLUE(1), GREEN(0)\}$, $z \in \left\{\blacktriangle(1), \bigcirc(0)\right\}$, $y \in \{HIGH(1), LOW(0)\}$.

The main goal of this neutral framing is, to the extent possible, reduce the chances that subjects import priors into the experiment. Of course, in reality, narratives are never context-free, but because we cannot control priors, using a non-neutral frame would likely introduce confounds.[10]

Second, to avoid deception, we constructed narratives that point out patterns in the data, rather than explicitly stating a causal model. The narratives also give policy recommendations, as many narratives do in practice. If the narratives do change beliefs, it implies that subjects both (i) form a causal model from the narrative and (ii) use that causal model to update their beliefs.

Third, we use both independent and causal datasets to compare situations in which a true causal relationship does and does not exist. Because the causal datasets imply a fairly strong causal relationship, we think of the two cases as testing two extremes. The causal datasets also allow us to test whether (and which types of) narratives work even when they oppose a true causal relationship in the data. Finally, the comparison of 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, policies do provide control over the outcome, so if narratives also work 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

[10] In a previous experiment, we used a less neutral frame, labeling the action, 'Manager Action', the outcome, 'Firm Profits', and the auxiliary variable, 'Employee Action'. We present the results of this experiment in Appendix B. The results are very similar.

datasets.

Fifth, we put narratives head-to-head to with other narratives and statistical summaries for two reasons. First, to see whether subjects adopt the narrative or summary with the highest anticipatory utility (Eliaz and Spiegler (2020)), highest coverage, the one that is more difficult to falsify, or the one that provides illusion of control. 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, observing that narratives work even when presented alongside summaries 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 subject 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 ($a \to z \to y$).

### 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.[11] We recruited a sample of the U.S.

[11]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.

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 Appendix C), 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).

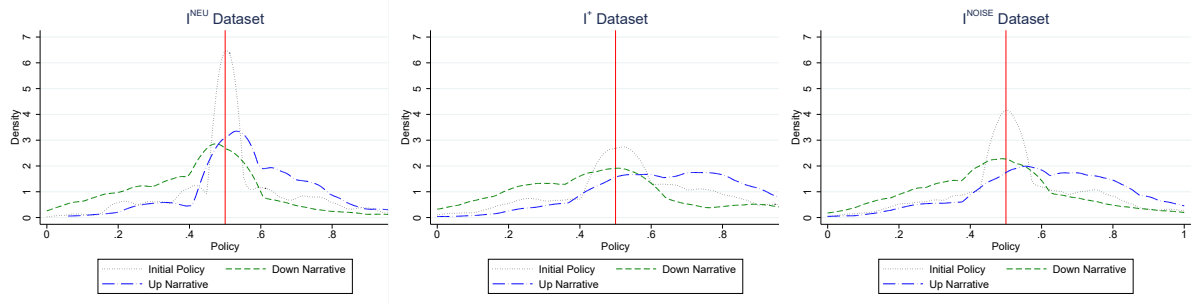
## 3.2 Results - CONSTRUCTED

We first show results for the CONSTRUCTED treatment with independent datasets, establishing that causal narratives, particularly Lever narratives, have robust effects on policy choices, including when 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 summarize the results with respect to the BNFF predictions, analyze the case of competing narratives, and run additional tests to rule out a pair of mechanisms that could have been driving some of our results.

### 3.2.1 Independent Datasets

We begin with an overview of the 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 pattern 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.[12]

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).[13]

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).[14] Furthermore, neither confidence interval contains 0.5, so both averages are significantly different from the rational policy. When we compare to the BNFF predictions, indicated by the gray horizontal lines, we see that average policies undershoot the prediction. The Threat narrative is the only elaborate narrative for which we cannot reject the null that the formula predicts the average policy choice correctly.

**Result 1**: *Lever and Threat narratives result in different policy choices for the same dataset.*

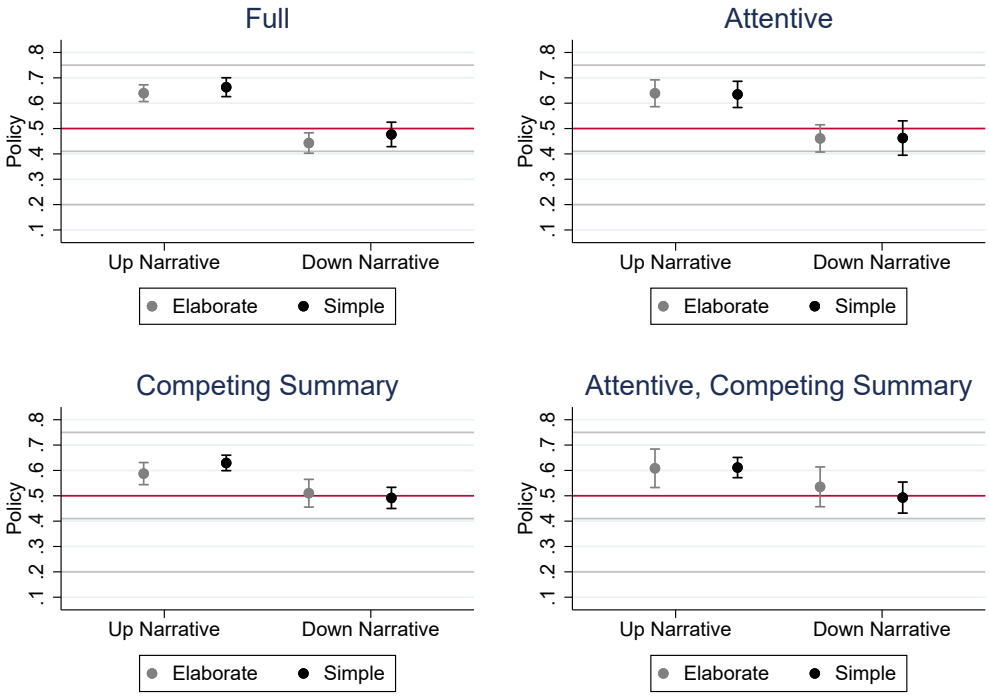
For simple narratives, both the rational prediction and that of the BNFF is 0.5, but we can reject the null that the Simple Up narrative produces an average policy of 0.5. In fact, this narrative produces a slightly larger response than the Lever narrative, though not significantly so ($p = 0.347$, two-sample t-test). The fact that the point estimates of the

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

[13]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).

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

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).

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 find the pattern corresponding to 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 theory, we find a highly significant difference (0.09, $p < 0.001$). Because the only difference between the two datasets is the auxiliary variable, it must drive the difference in policy choices.[15] 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

[15]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 point estimate goes the opposite way (lower policies in $I^{NEU}$ than in $I^{+}$). This suggests that the patterns associated with the Threat narrative are more difficult to pick up, a finding we will confirm in Section 4.2.

**Result 2**: *Subjects significantly respond to Simple Up narratives only when the auxiliary variable is correlated with the action and outcome variables, suggesting that they pick up on the pattern 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 (up) 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 the attentive subjects (51% of all subjects), we see very similar effects to the full sample except that we can no longer reject rational policy choices for the Threat narrative. This finding demonstrates that Results 1 and 2 are not driven by subjects simply following whatever they are told.[16]

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 two recommendations are displayed. Second, one may be concerned about a form of confirmation bias - subjects look only for information

[16]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.

that supports the narrative. But, when given these two pieces of advice, confirmation bias could work equally well for either. Even though this test rules out many possible reasons that subjects may follow narratives, 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 of 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 Lever narrative is predicted to have a larger effect than the Threat narrative, but we confirm the finding in the next section with the $I^{NOISE}$ dataset, where the Threat narrative is predicted to have the larger effect.

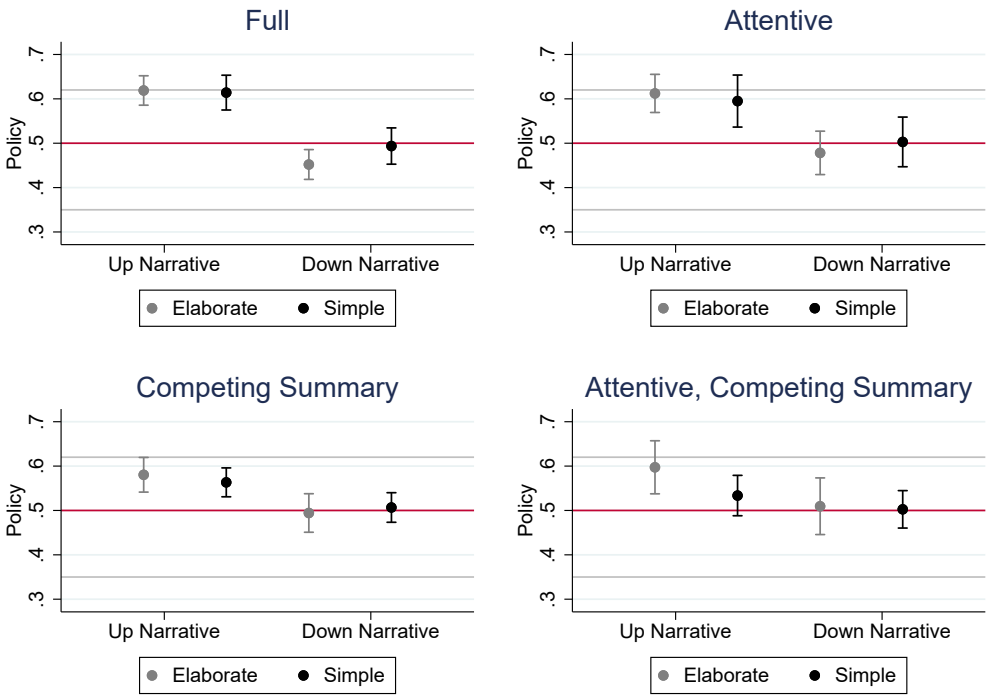
### 3.2.2 Noise Datasets

Here, we test whether noisy narratives that leverage weaker correlations in the data can also have effects. We do so by plotting average choices for the $I^{NOISE}$ dataset in Figure 4.

Broadly speaking, the results for noisy narratives are very similar to those for deterministic narratives, showing that noisy narratives work as well as deterministic narratives. 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 three subtle differences, however. First, policies deviate somewhat less from the rational policy when the Lever is noisy, consistent with the BNFF prediction being lower. In fact, here we cannot reject the hypothesis that the BNFF prediction for the Lever narrative is correct. Second, the Simple Up narrative also produces slightly smaller effects, which may mean that subjects have a harder time picking up a noisy pattern in the dataset compared to a deterministic one. Third, 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 makes them less robust. We consider several possibilities in Section 5.

**Result 4**: *Noisy Lever and noisy Threat narratives move subjects' policy choices in oppo-*

Figure 4. $I^{NOISE}$ Dataset Average Policies



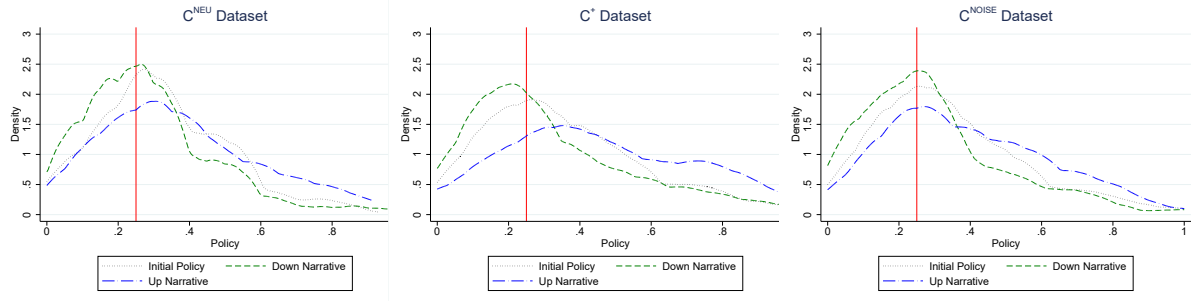

Notes: Average policy choices and 95 percent confidence intervals. The upper left panel is for all subjects. The upper right panel restricts to attentive subjects that do not respond to inconsistent narratives. The lower left panel is for policy choices when the narrative competed directly with a summary. 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).

*site directions. Therefore, deterministic patterns in the data are not necessary for narratives to be effective. Noisy Lever narratives are more robust than noisy Threat narratives.*

### 3.2.3 Causal Datasets

Here, we test whether a causal narrative can be effective when it directly opposes a strong causal relationship. As a first step, Figure 5 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, the initial densities shift towards higher policies in the $C^{+}$ and $C^{NOISE}$ datasets, again suggesting that subjects pick up on the pattern 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

Figure 5. 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.

policy of 0.25, but many choose policies above 0.5 (i.e., in the opposite direction of the true causal relationship).

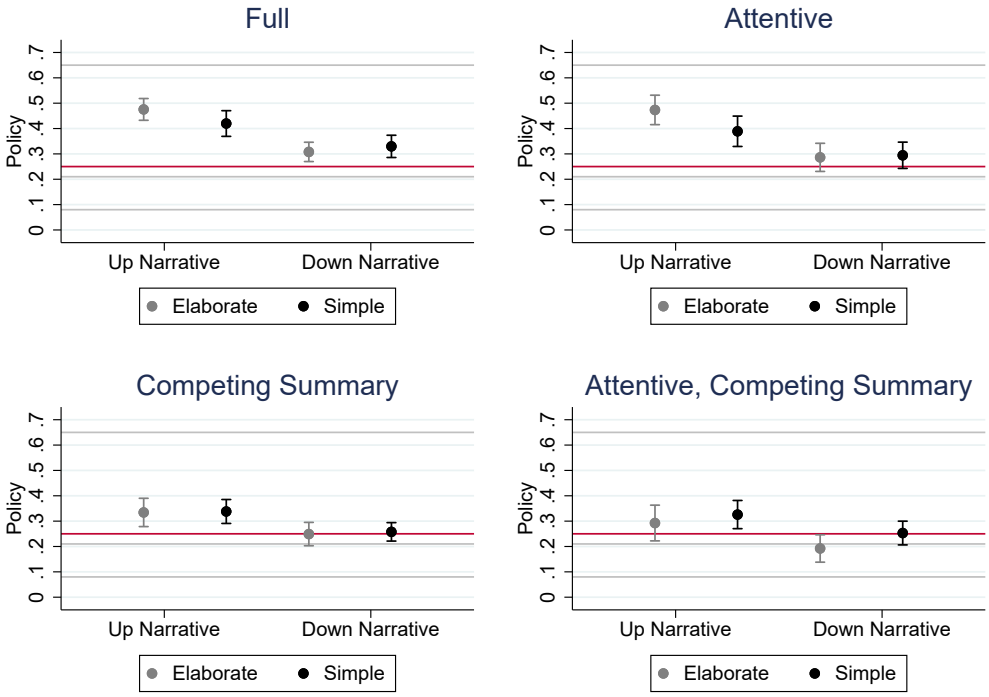
To formally test for the statistical significance of these patterns, we replicate Figures 3 and 4 for the $C^{+}$ and $C^{NOISE}$ datasets in Figures 6 and 7, respectively. In Figure 6, we see that the Lever narrative can be effective even when it contradicts a causal relationship. The average policies 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.[17]

Looking at simple narratives, subjects again appear to find the pattern 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 7, we see very similar effects when noise is added to the dataset, demonstrating again 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 causal relationship.*

[17]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. $C^+$ Dataset Average Policies



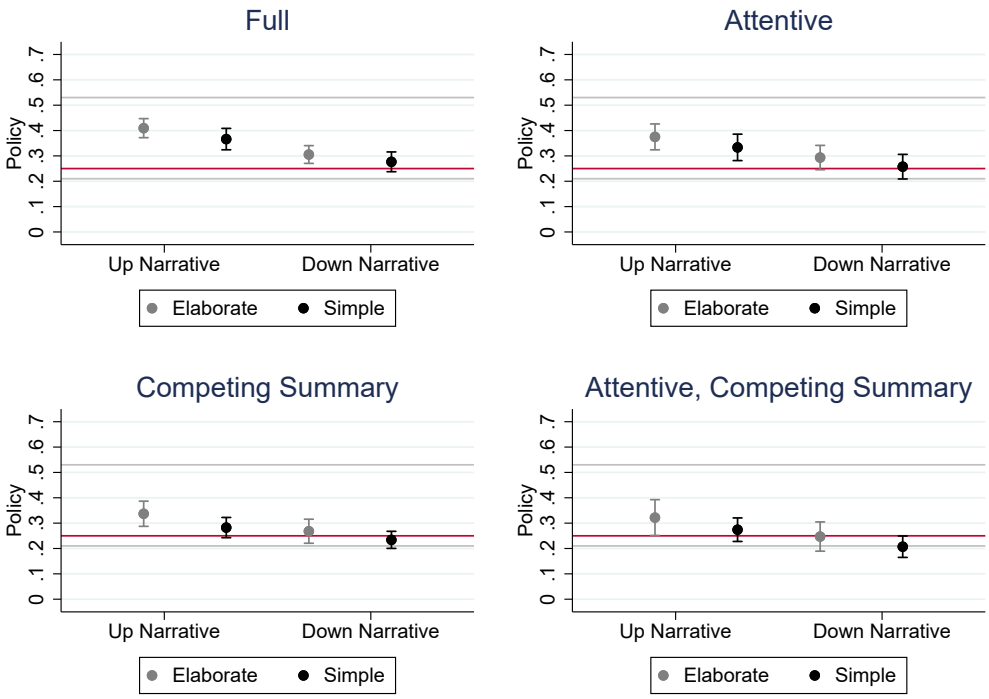

Notes: Average policy choices and 95 percent confidence intervals. The upper left panel is for all subjects. The upper right panel restricts to attentive subjects that do not respond to inconsistent narratives. The lower left panel is for policy choices when the narrative competed directly with a summary. 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).

### 3.2.4 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

Figure 7. $C^{NOISE}$ 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).

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 initial policy choices in $C^{NEU}$. The idea is that the 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 A4, we show that Lever and Simple Up narratives continue to have positive effects by this measure, except when subjects see a summary simultaneously.[18]

A possible reason for the observed compression to the middle is cognitive uncertainty (Enke and Graeber (2023)).[19] Subjects might treat the least costly policy of 0.5 as a cognitive

[18]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.

[19]Risk aversion doesn't straightforwardly result in compression because choosing an extreme policy reduces

default, on which they lean when they are uncertain about the optimal policy. To investigate this possibility, we split the sample at the median 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 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.

**Result 6**: *The Bayesian-network factorization formula tends to overpredict the amount by which narratives affect policy choices. This overprediction is driven by subjects compressing their policies towards the least costly policy of 0.5.*

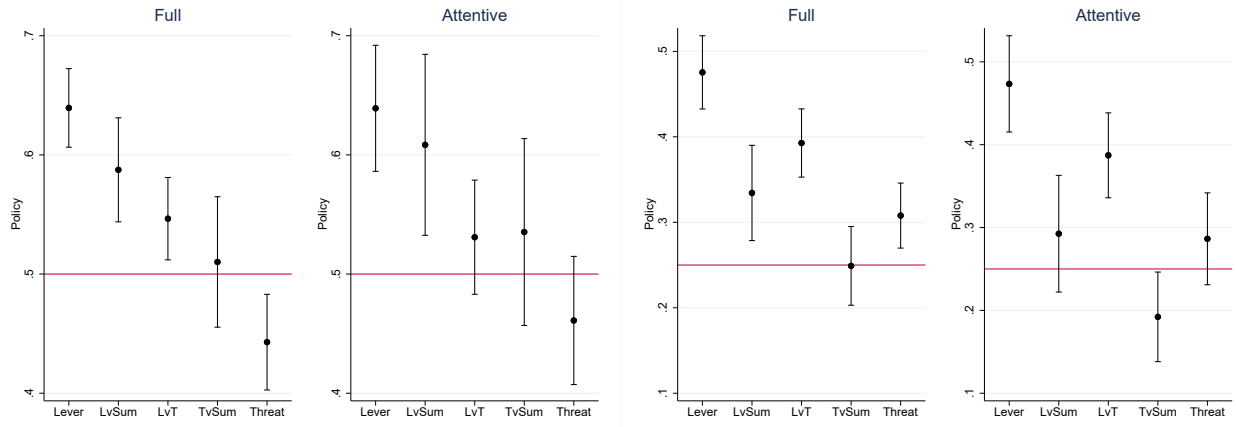
#### 3.2.5 Competing Narratives

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).

First, consider the predictions of the theories put forth in Section 2.5. Each of these theories points to exactly one of the narratives being adopted. Coverage would favor the Lever and Threat narratives because the statistical summary does not explain where the auxiliary variable comes from. Falsification favors the Lever narrative over the Threat narrative because the Threat narrative implies a counterfactual unconditional distribution, but favors the summary over both because the summary describes the truth. Illusion of control would favor Lever and Threat narratives over the summary, but only in the independent datasets. Finally, anticipatory utility makes predictions according to Table A1 of Appendix A, ranking the two elaborate narratives and the summary differently across datasets.

variability in the outcome.

Figure 8. 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.

Figure 8 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). Figure A2 in Appendix A shows that the patterns for the $I^{NOISE}$ and $C^{NOISE}$ datasets are quite similar.

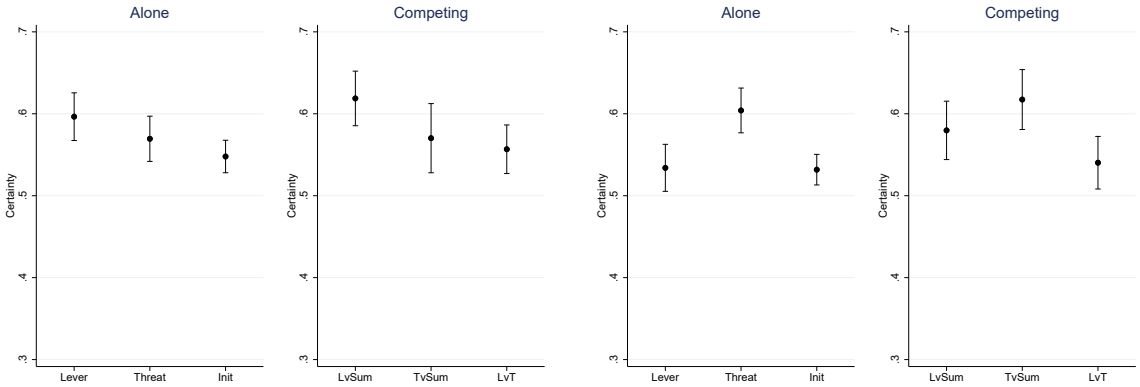
The striking finding in Figure 8 is that subjects do not appear to adopt one narrative over the other as all of the above theories predict. Instead, when subjects see competing narratives, they choose policies that lie in between the policies that they choose when they see each narrative on their own. In all cases, 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. 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). The only exception is when the Threat narrative competes with a summary. Here, the policies do not lie between the Threat narrative alone and the rational prediction. This result is likely a consequence of summaries killing off the effect of the Threat narrative, as shown in the previous sections.

Because Figure 8 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 A3 of Appendix A 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 9 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.

Finally, another piece of evidence that subjects consider both narratives comes from how they rate their helpfulness. When elaborate narratives compete with summaries in the independent datasets, the elaborate narratives are rated as helpful by 65% of subjects while the summaries are rated as helpful by 70% of subjects, a difference that is statistically insignificant ($p = 0.181$; two-sample t-test). In causal datasets, there is a difference: 62% of subjects rate elaborate narratives as helpful (consistent with the fact that they affect choices), but 87% rate summaries as helpful ($p < 0.001$). Importantly though, in both cases, a majority of subjects consider both pieces of advice to be helpful, suggesting that subjects are weighting both narratives.[20]

Figure 9. Confidence




Notes: Average subject confidence in their policy choices, pooled across $I^+$ and $I^{NOISE}$ (left panels) and $C^+$ and $C^{NOISE}$ (right 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).

We are left with the possibility that subjects combine the two models provided by the narratives in some sophisticated way.[21] Although intuitive, such behavior is markedly differ-

[20]The fraction of subjects rating each narrative as helpful also exceeds what we might consider to be a lower bound for helpfulness: that of inconsistent narratives (55% and 39% in independent and causal datasets, respectively).

[21]Vespa and Wilson (2016) similarly find that subjects average the recommendations of two senders in a

ent 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 (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 – our results show that 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.

**Result 7**: *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.*

#### 3.2.6 Additional Tests

In this section, we leverage additional randomization in the design to test for two possible mechanisms that could be driving our results.

The first mechanism is column ordering: 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

communication game even when it is not optimal.

at the 10% level. Overall, column ordering does not seem to have large effects.

The second mechanism is anchoring: 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.

# 4 Creating and Transmitting Narratives

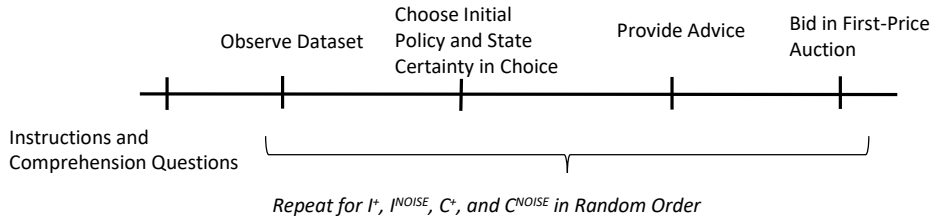
The pair of treatments we describe here serve two purposes. First, to make explicit the previous finding that subjects appear to pick up on patterns in the dataset, forming their own implicit models of the data-generating process. Second, to show that subjects construct narratives to communicate these 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

Figure 10. Timeline for Elicit Treatment




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 10 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 the narrative paired with a summary, they saw the Lever narrative 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 2022 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 Appendix C), 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 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 A2 of Appendix A. In Table 4, we combine the original categorizations into broader categories (as described in the notes for Table A2). The first column of Table 4 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).

Most strikingly, over half 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*

[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.

Table 4. 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).

*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 identify the Lever narrative much more often than the Threat narrative makes explicit the finding that subjects in CONSTRUCTED appear to implicitly pick up on the Lever narrative pattern in choosing their initial policies.[23]

The second column of Table 4 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 4, we show the breakdown for the $C^+$ and $C^{NEU}$ datasets, respectively. In these datasets, there is a causal relationship that favors choosing a lower policy. As with the independent datasets, the most common advice is rational. However, 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 supports a lower policy.[24]

**Result 8**: *Subjects produce elaborate causal narratives after simply observing a dataset*

[23]In a previous experiment, documented in Appendix B, 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 $I^+$ 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 while realizing themselves that it is not actually helpful, or because they are uncertain whether it is actually good advice. In Figure 11, 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).

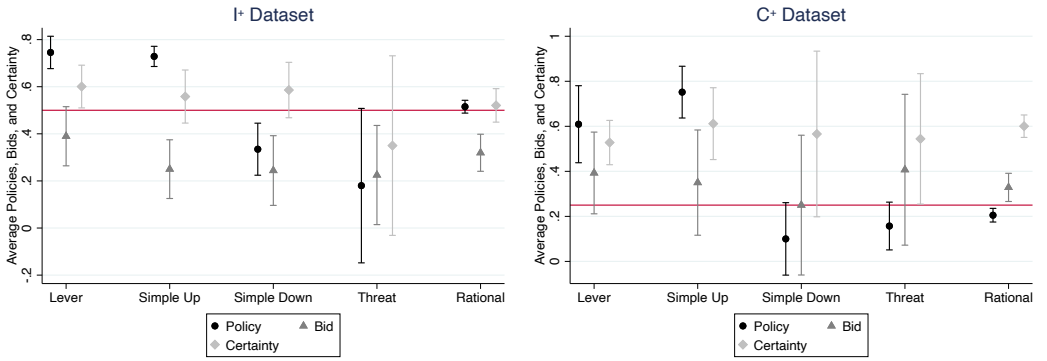
We see that, on the whole, subjects follow their own advice: policies are very close to the rational policy when subjects provide rational advice, and deviate in the indicated direction of the advice for the other types of advice. However, this figure masks important heterogeneity in behavior: in the $I^+$ dataset, there are five subjects who provide the Lever narrative while choosing a policy of exactly 0.5, suggesting that they did not believe, or were not certain about, the advice they produced.[26] Similarly, in the $C^+$ dataset, there are three subjects who provide the Lever narrative while choosing a policy of $<= 0.25$.

Figure 11 also shows subjects' average bids and average certainty in their policy choices.

[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.

[26]Evaluating the tone of these subjects' advice, we believe it is more likely that they were uncertain about their advice and not trying to purposely mislead other subjects.

Figure 11. 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.

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. 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 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.

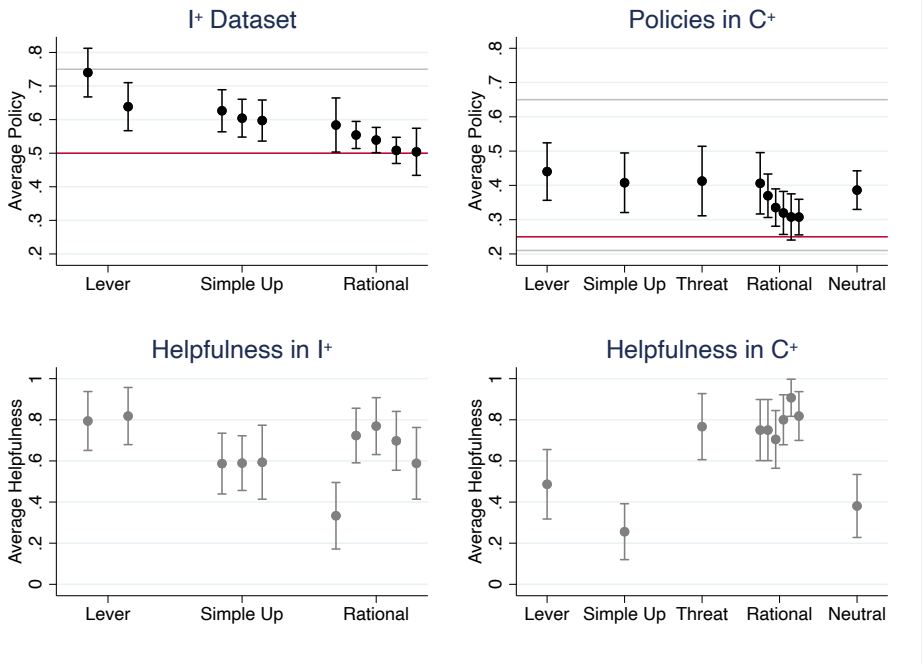
## 4.3 Results - NATURAL

In the NATURAL treatment, subjects received the narratives constructed by the subjects in ELICIT. Figure 12 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 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

Figure 12. 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.

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 particularly helpful. Instead, subjects consider rational advice, along with the Threat narrative, as most helpful.

Overall, these results suggest that even naturally-generated causal narratives, particularly Lever narratives, can have strong effects. In some cases the effects are as large as the effects of the narratives we constructed, as presented in Section 3.2.

**Result 9**: *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 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 theorists to continue to use the formula to model causal narratives for several reasons. First, and perhaps

most importantly, the formula 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. Second, most other theories we rely on in our standard models are also not 100 percent accurate when confronted with experimental data (e.g., Bayes' theorem or expected utility). Third, we presented evidence that the undershooting we observe may be due to compression to a mental default (Danz, Vesterlund, and Wilson (2022); Enke and Graeber (2023)), the least costly policy in our case.

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. One could, for example, easily model the averaging behavior we observed by calculating beliefs according to the BNFF for each narrative and then calculating some weighted average of the two. As we have shown, such a model would be consistent with our results (setting aside compression).

However, we also acknowledge that more evidence is needed about how narratives are combined. First, in our setting, the anticipatory utilities of the competing narratives were often very similar in magnitude - with wider separation, it could be that the narrative with higher anticipatory utility is adopted outright. Second, in real-world settings, people may adopt the narrative that comes from a trusted source or that they hear most frequently. Alternatively, they may adopt the narrative which they consider more plausible or more consistent with their knowledge of the world (Pennington and Hastie (1993)). We deliberately chose neutral language to avoid the confounds of unknown priors, but causal narratives should also be studied in settings in which they come with naturalistic connotations. 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 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)).

## 5.2 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, a hypothesis confirmed by the finding that subjects produce narratives that point out these correlations on their own when asked to explicitly give advice. 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 other reasons Lever narratives may be more effective, however. Causal chains may simply come more naturally to people.[27] 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 a literature (e.g., Vrantsidis and Lombrozo (2022)) showing that people tend to value simplicity in explanations. However, complexity does not appear to be the whole story because when a true causal relationship exists, summarizing this relationship is arguably simpler than a Lever narrative – yet we find that Lever narratives are powerful even in these settings. 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 a Lever narrative to be 'closer' to a Threat narrative, for example. Perhaps extending to settings with more than three variables or unobserved variables (i.e., omitted variable bias) will help to provide a metric for ranking the appeal of narratives.

## 5.3 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. It is perhaps not surprising in light of these results 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

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

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).

### 5.4 Conclusion

Causal narratives are abundant with 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 think economics as a field would benefit from further research, both theoretical and empirical.

## References


[1] Aina, Chiara. 2022. "Tailored Stories." working paper.

[2] Andre, Peter, Ingar Haaland, Christopher Roth, and Johannes Wohlfart. 2022. "Narratives About the Macroeconomy." working paper.

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

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

[5] 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).

[6] Barron, Kai and Tilman Fries. 2022. "Narrative Persuasion." working paper.

[7] Benabou, Roland, Armand Falk, and Jean Tirole. 2018. "Narrative, Imperatives, and Moral Reasoning." working paper.

[8] Bursztyn, Leonardo, Aakaash Rao, Christoper Rao, and David Yanagizawa-Drott. 2022. "Opinions as Facts." *The Review of Economic Studies*, forthcoming.

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

[10] 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.

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

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

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

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

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

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

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

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

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

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

[21] Graeber, Thomas, Christopher Roth, and Florian Zimmerman. 2022. "Stories, Statistics, and Memory." working paper.

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

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

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

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

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

[27] Kendall, Chad and Ryan Oprea. 2022. "On the Complexity of Forming Mental Models." working paper.

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

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

[30] 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.

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

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

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

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

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

[36] 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.

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

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

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

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

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

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

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

[44] Sloman, Steven and David Lagnado. 2015. "Causality in Thought." *Annual Review of Psychology*, 66:223-247.

[45] 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).

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

[47] Steyvers, Mark, Joshua B. Tenenbaum, Eric-Jan Wagenmakers, and Ben Blem. 2003. "Inferring Causal Networks from Observations and Interventions." *Cognitive Science*, 27, 453-489.

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

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

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

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

[52] 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: 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 |

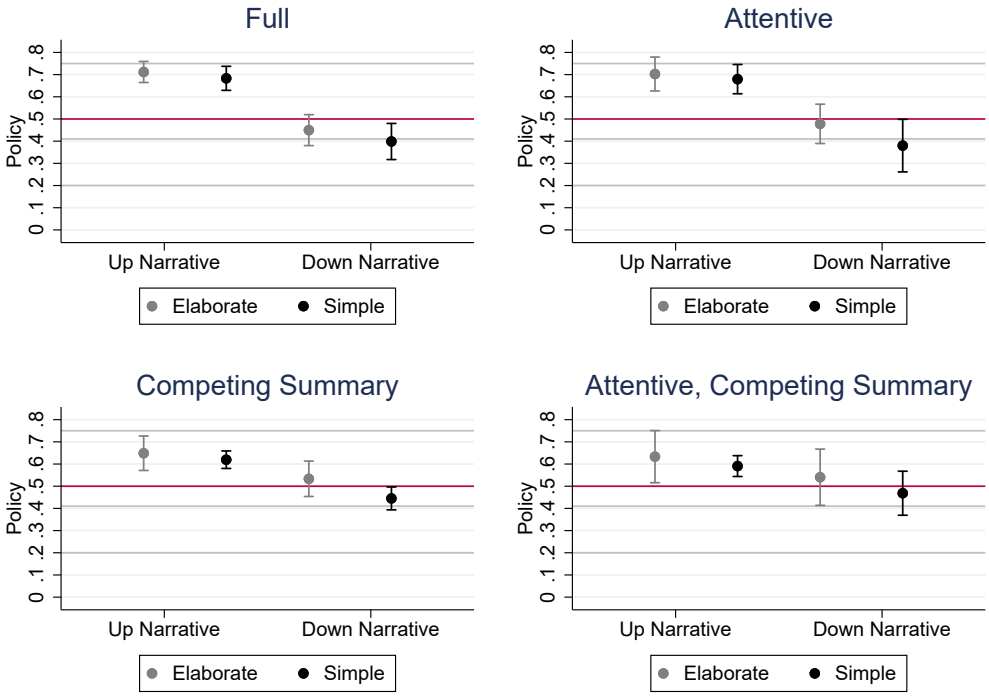
Notes: Anticipatory utility for each dataset (column) and narrative (row), calculated using Equation (3) from the main text and the subjective beliefs under each narrative. For Threat narratives in the absence of noise, a range of utilities is predicted because beliefs are not completely pinned down by the dataset.

Table A2: 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 |

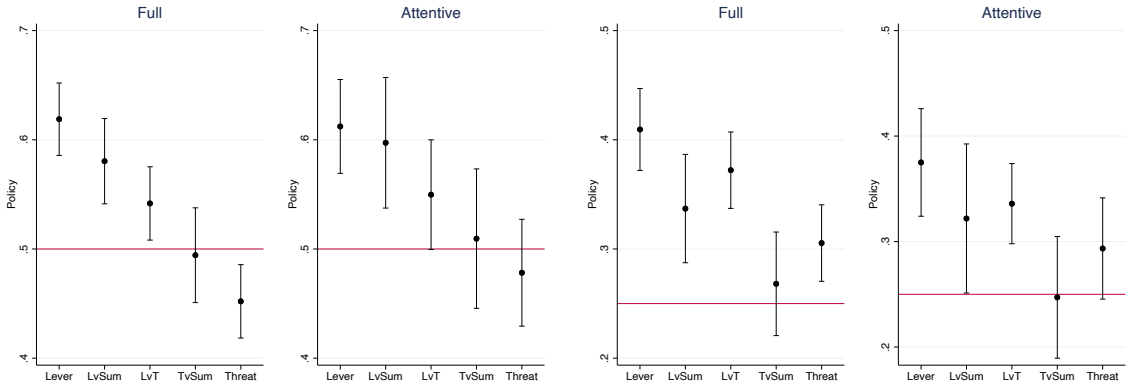
Notes: Classification of elicited narratives (percentages) in each dataset. Blue High and Green High suggest that the corresponding color leads to HIGH ($y = 1$) outcomes more often. Blue Other and Green Other recommend the corresponding color, but do not provide a particular reason. Rational recommends 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 4, 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)



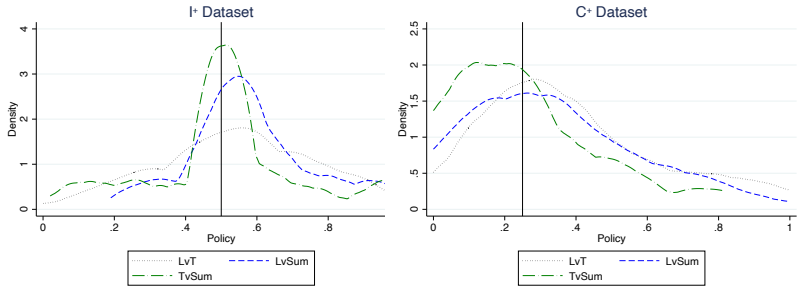

Notes: Average policy choices and 95 percent confidence intervals. We restrict the data to subjects who saw the $I^{+}$ dataset first. The upper left panel is for all subjects. The upper right panel restricts to attentive subjects that do not respond to inconsistent narratives. The lower left panel is 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: Competing Narratives in $I^{NOISE}$ and $C^{NOISE}$



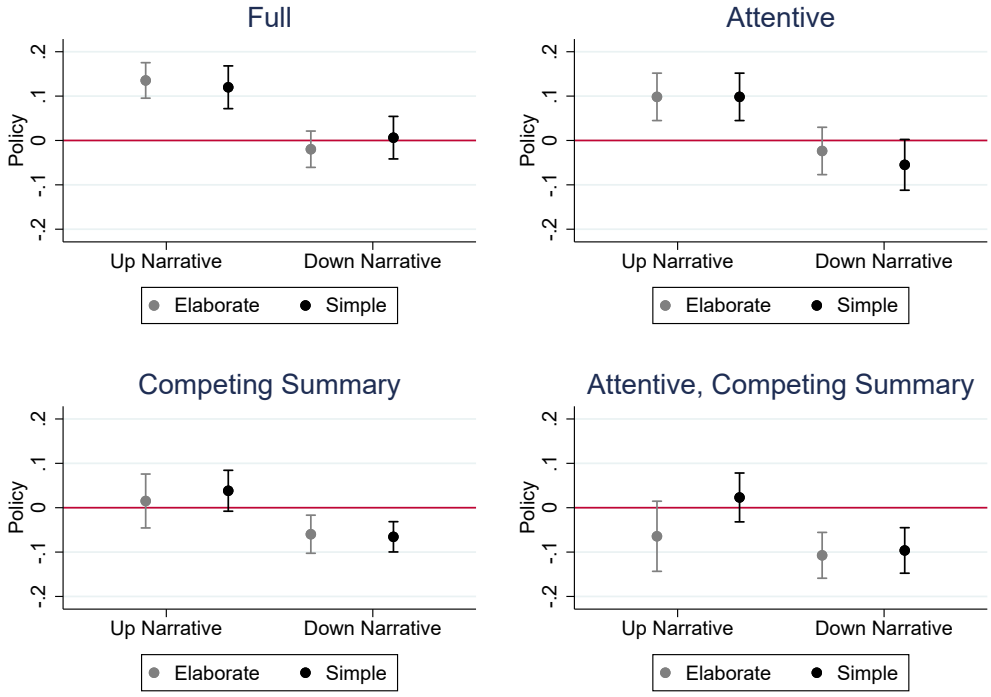

Notes: Average policy choices and 95 percent confidence intervals. LvSum indicates the average policy choice when subjects observed both the Lever narrative and the summary. TvSum indicates the Threat narrative and the summary. LvT indicates both Lever and Threat narratives. The left pair of graphs is 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.

Figure A3: 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.

Figure A4: 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.

# B Prior Experimental Results

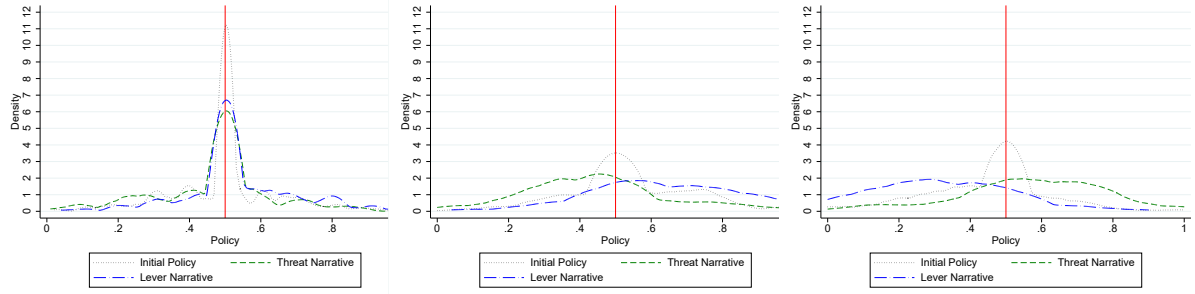
We circulated an older version of this paper in 2022 (Charles and Kendall, 2022). This "2022 version" contains the results of three experiments, which we have now removed from the current version of the paper. The previous working paper (available here) describes the prior experimental design and results 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.

## B.1 CONSTRUCTED

Figure B1 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.

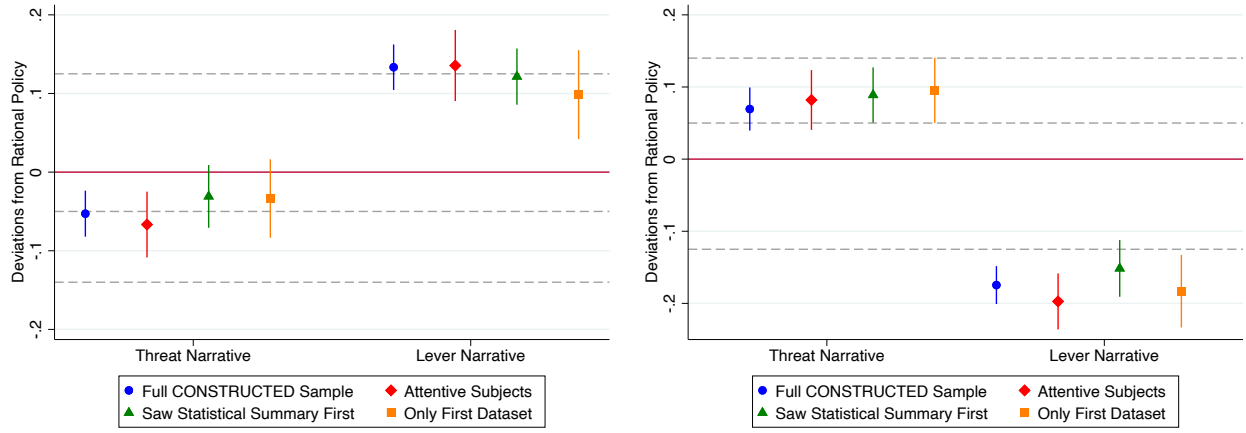
Figure B1: Policy Densities in CONSTRUCTED




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 B2 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 B2: 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 B1, 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 B2, 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.

## B.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.[28] 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.

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

# C Instructions

The instructions and decision screens for the CONSTRUCTED treatment follow. Only the decision screens for the first dataset ($I^{NEU}$ in this case) are shown.

USC University of Southern California

# Instructions (Overview)

In this experiment, you will observe a dataset and make decisions based on it. You will do this in three separate trials (in sequence). The datasets in each trial are **different** and not related to those in any other trial in any way, so nothing that you learn in one trial applies to any of the other trials. In each trial, you should only use the information that is provided to you in that trial.

In each of the three trials, you will make policy decisions. The policy you select determines the probability (percentage chance) you implement one of two choices: the BLUE choice or the GREEN choice.

To help you select your policy, you will receive access to historical data that summarizes many thousands of choices in a series of rows. **Each row in the summary table was historically observed an equal number of times (and so is equally likely).** The data shows the following information:

- Which choice was implemented: BLUE or GREEN.
- A variable simply labeled 'X'. The historical data uses a code to record the 'X' column, but does not indicate what the code means. All you know is that the same code always means the same thing. 'X' is coded as ▲ or ◯.
- Whether the payoff was **HIGH** or LOW.

Your goal is to obtain a **HIGH** payoff. The choices, 'X' values, and payoffs may or may not be related. When you choose your policy, the resulting choice will have the same relationships (if any) with the 'X' variable and payoff as it did in the past. The historical data contains all of the information you need (there are no other 'hidden' variables that matter) so you should use it to determine any relationships.

After each policy decision, you will report your confidence in your policy decision. For the second and third policy decisions in each period, you will receive advice before you make your decision. This advice may or may not help you in making your policy decisions - you should always assess for yourself whether or not the advice makes sense! You will be asked whether or not you found each piece of advice helpful or not.

You will be paid the bonus that results from one of your policy decisions (chosen randomly with equal chance). We describe in more detail how your bonus will be determined on the following page.

When you have read these instructions, please continue to the next page.

USC University of Southern California

# Instructions (Bonus)

Your bonus depends on the policy you select and the payoff. When you select a policy, you will determine the percentage chance that the BLUE choice is implemented. Call this P. The GREEN choice will be implemented with the remaining chance, 100-P. The least costly policy is P=50: the policy in which you implement the BLUE action 50% of the time. If you choose a higher or lower chance of BLUE, the cost is higher.

Specifically, you will get a bonus from two parts (added together):

1. You will receive $1.00 if the payoff is **HIGH** and $0 if the payoff is LOW.
2. You will receive 0.167-0.667x(P/100-0.5)^2 where P is the policy you choose. For example, you will receive $0.167 if you choose P=50 but $0 if you choose P=0 or P=100.

When choosing the policy, think carefully! A more expensive policy might be worth it if it increases the chance that the payoff is **HIGH**. When you have finished reading these instructions, please proceed to the next page.

USC University of Southern California

Please answer the following quiz questions:

1. Your bonus depends on:

- ◯ the cost of the policy you choose
- ◯ the payoff
- ◯ both of the above

2. True or false: your bonus will be higher if the payoff is **HIGH**:

- ◯ True
- ◯ False

3. True or false: a policy that implements the BLUE choice always is less costly than one which implements the BLUE choice and the GREEN choice each with a 50% chance.

- ◯ True
- ◯ False

USC University of Southern California

# Historical Data for Trial 1

The table below shows you the historical information for this trial. Recall that the table summarizes many thousands of choices and each row is equally likely. You should carefully study the table to look for clues as to which policy will maximize your bonus. Then, proceed to the next page to make your policy decision.

| Choice | X | Payoff |
|---|---|---|
| GREEN | ▲ | LOW |
| GREEN | ◯ | **HIGH** |
| GREEN | ◯ | LOW |
| GREEN | ◯ | LOW |
| BLUE | ◯ | LOW |
| BLUE | ▲ | LOW |
| GREEN | ▲ | **HIGH** |
| BLUE | ▲ | **HIGH** |
| GREEN | ▲ | **HIGH** |
| BLUE | ◯ | **HIGH** |
| GREEN | ◯ | **HIGH** |
| GREEN | ▲ | LOW |
| BLUE | ◯ | LOW |
| BLUE | ▲ | **HIGH** |
| BLUE | ▲ | LOW |
| BLUE | ◯ | **HIGH** |

USC University of Southern California

# Policy Decision for Trial 1

You will now choose a policy. Please choose it using the slider below. The value of the slider corresponds to the percentage chance that you make the BLUE choice. To help you make the best decision, below is the historical information that you just reviewed.

| Choice | X | Payoff |
|---|---|---|
| GREEN | ▲ | LOW |
| GREEN | ◯ | **HIGH** |
| GREEN | ◯ | LOW |
| GREEN | ◯ | LOW |
| BLUE | ◯ | LOW |
| BLUE | ▲ | LOW |
| GREEN | ▲ | **HIGH** |
| BLUE | ▲ | **HIGH** |
| GREEN | ▲ | **HIGH** |
| BLUE | ◯ | **HIGH** |
| GREEN | ◯ | **HIGH** |
| GREEN | ▲ | LOW |
| BLUE | ◯ | LOW |
| BLUE | ▲ | **HIGH** |
| BLUE | ▲ | LOW |
| BLUE | ◯ | **HIGH** |

Less BLUE — More BLUE

0 10 20 30 40 50 60 70 80 90 100

Percent chance of BLUE choice:


How certain are you that your chosen policy maximizes your bonus? Please use the slider to indicate your level of certainty.

Very Uncertain — Very Certain

0 10 20 30 40 50 60 70 80 90 100

Certainty

USC University of Southern California

# Policy Decision for Trial 1 (first advice)

You have now received advice from someone who has analyzed the historical data that you just reviewed. The advice is as follows:

*Choose BLUE more often.*

Recall that this advice may or may not be useful in helping you make your policy decision. Please think about it carefully to see if it makes sense given the historical information you reviewed (repeated again below).

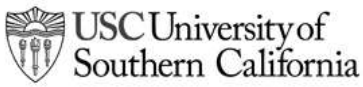
| Choice | X | Payoff |
|---|---|---|
| GREEN | ▲ | LOW |
| GREEN | ◯ | **HIGH** |
| GREEN | ◯ | LOW |
| GREEN | ◯ | LOW |
| BLUE | ◯ | LOW |
| BLUE | ▲ | LOW |
| GREEN | ▲ | **HIGH** |
| BLUE | ▲ | **HIGH** |
| GREEN | ▲ | **HIGH** |
| BLUE | ◯ | **HIGH** |
| GREEN | ◯ | **HIGH** |
| GREEN | ▲ | LOW |
| BLUE | ◯ | LOW |
| BLUE | ▲ | **HIGH** |
| BLUE | ▲ | LOW |
| BLUE | ◯ | **HIGH** |

Please choose the policy that you want to implement in this period. Recall that the policy you choose determines the percentage chance that you take the BLUE choice.

Less BLUE — More BLUE

0 10 20 30 40 50 60 70 80 90 100

Percent chance of BLUE choice:




Consider the advice you just received:

*Choose BLUE more often.*

Please indicate whether you found this advice helpful in making your policy decision:

- ◯ Helpful
- ◯ Not helpful

How certain are you that your chosen policy maximizes your bonus? Please use the slider to indicate your level of certainty.

Very Uncertain — Very Certain

0 10 20 30 40 50 60 70 80 90 100

Certainty

USC University of Southern California

# Policy Decision for Trial 1 (second advice)

You have now received a second piece of advice from someone who has analyzed the historical data that you just reviewed. The advice is as follows:

*X is a ▲ only when the choice is BLUE . Further, when X is a ▲, the payoff is always **HIGH**. So, choose BLUE more often.*

Recall that this advice may or may not be useful in helping you make your policy decision. Please think about it carefully to see if it makes sense given the historical information you reviewed (repeated again below).

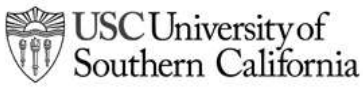
| Choice | X | Payoff |
|---|---|---|
| GREEN | ▲ | LOW |
| GREEN | ◯ | **HIGH** |
| GREEN | ◯ | LOW |
| GREEN | ◯ | LOW |
| BLUE | ◯ | LOW |
| BLUE | ▲ | LOW |
| GREEN | ▲ | **HIGH** |
| BLUE | ▲ | **HIGH** |
| GREEN | ▲ | **HIGH** |
| BLUE | ◯ | **HIGH** |
| GREEN | ◯ | **HIGH** |
| GREEN | ▲ | LOW |
| BLUE | ◯ | LOW |
| BLUE | ▲ | **HIGH** |
| BLUE | ▲ | LOW |
| BLUE | ◯ | **HIGH** |

Please choose the policy that you want to implement in this period. Recall that the policy you choose determines the percentage chance that you take the BLUE choice.

Less BLUE — More BLUE

0 10 20 30 40 50 60 70 80 90 100

Percent chance of BLUE choice:




Consider the advice you just received:

*X is a ▲ only when the choice is BLUE . Further, when X is a ▲, the payoff is always **HIGH**. So, choose BLUE more often.*

Please indicate whether you found this advice helpful in making your policy decision:

- ◯ Helpful
- ◯ Not helpful

How certain are you that your chosen policy maximizes your bonus? Please use the slider to indicate your level of certainty.

Very Uncertain — Very Certain

0 10 20 30 40 50 60 70 80 90 100

Certainty

The instructions and decision screens for the ELICIT treatment follow. Only the decision screens for the first dataset ($C^+$ in this case) are shown. For brevity, we omit the comprehension questions and the screens that present the dataset, ask for the initial policy choice, and ask subjects to state their certainty. They are identical to those in CONSTRUCTED.

USC University of Southern California

# Instructions (Overview)

In this experiment, you will observe a dataset and make decisions based on it. You will do this in four separate trials (in sequence). The datasets in each trial are **different** and not related to those in any other trial in any way, so nothing that you learn in one trial applies to any of the other trials. In each trial, you should only use the information that is provided to you in that trial.

In each of the four trials, you will (i) make a policy decision, (ii) give advice to future participants in our study, and (iii) report your confidence in your policy decision.

The policy you select determines the probability (percentage chance) you implement one of two choices: the BLUE choice or the GREEN choice.

To help you select your policy, you will receive access to historical data that summarizes many thousands of choices in a series of rows. **Each row in the summary table was historically observed an equal number of times (and so is equally likely).** The data shows the following information:

- Which choice was implemented: BLUE or GREEN.
- A variable simply labeled 'X'. The historical data uses a code to record the 'X' column, but does not indicate what the code means. All you know is that the same code always means the same thing. 'X' is coded as ▲ or ◯.
- Whether the payoff was **HIGH** or LOW.

Your goal is to obtain a **HIGH** payoff. The choices, 'X' values, and payoffs may or may not be related. When you choose your policy, the resulting choice will have the same relationships (if any) with the 'X' variable and payoff as it did in the past. The historical data contains all of the information you need (there are no other 'hidden' variables that matter) so you should use it to determine any relationships.

You will be paid two bonuses. The first comes from one of your four policy decisions (chosen randomly with equal chance). The second comes from one of the four pieces advice you give (also chosen randomly with equal chance). We describe in more detail how these bonuses will be determined on the following two pages.

When you have understood these instructions, please continue to the next page.

USC University of Southern California

# Instructions (Policy Choice)

Your first bonus depends on the policy you select and the payoff. When you select a policy, you will determine the percentage chance that the BLUE choice is implemented. Call this P. The GREEN choice will be implemented with the remaining chance, 100-P. The least costly policy is P=50: the policy in which you implement the BLUE action 50% of the time. If you choose a higher or lower chance of BLUE, the cost is higher.

Specifically, you will get a bonus from two parts (added together):

1. You will receive $1.00 if the payoff is **HIGH** and $0 if the payoff is LOW.
2. You will receive 0.167-0.667x(P/100-0.5)^2 where P is the policy you choose. For example, you will receive $0.167 if you choose P=50 but $0 if you choose P=0 or P=100.

When choosing the policy, think carefully! A more expensive policy might be worth it if it increases the chance that the payoff is **HIGH**. After choosing the policy, you will be asked to give advice to future participants in our study about which policy to choose and why. When you have finished reading these instructions, please proceed to the next page for instructions on how this advice can increase your bonus.

USC University of Southern California

# Instructions (Advice)

After making your policy choice, we will give you a text box to enter advice to give to future participants in our study. Giving good advice can increase your second bonus, as we explain here.

After you enter your advice, you will participate in an auction to decide whether or not your advice will be given to future participants in our experiment. To participate in the auction, you will receive $1.00 that you can use to make a bid. You can bid any amount up to this amount and will get to keep the rest.

In the auction, you will be matched with 19 other participants based on the order in which we receive responses. The advice from the participant that bids the most in the auction will be given to an average of 40 future participants. The future participants will know that the advice came from the past participant that bid the most for their advice to be used.

If your advice is given to future participants, you will receive $0.025 for each future participant that says your advice was helpful (versus unhelpful), minus the cost of your bid.

So, for example, if you bid $0.50, and if your bid is accepted, and 30 future participants say it was helpful, you will earn $0.75 because your advice was helpful. But, you will have to pay the cost of your bid, $0.50. With the $1.00 you received initially, your bonus is then $1.00+$0.75-$0.50=$1.25. By contrast, if you bid $1.00, and your bid is accepted but no future participant says your advice was helpful, you will receive no bonus. If your bid is not accepted, you will simply keep the initial $1.00 you received. We will make these additional bonus payments after the future study is completed, which we expect will be within a week.

You will assist us in our research if your advice is specific - it tells future participants which policy you think they should choose either explicitly or implicitly (by telling them what to look for in the dataset). **We will exclude your bid from the auction if it doesn't meet this requirement.**

When you have understood these instructions, please proceed to the next page to answer a few quiz questions before the experiment starts. You will have to answer the questions correctly to proceed.

USC University of Southern California

# Advice for Firm 1

Please give advice to a future participant in our study. Try to be specific about what you think the participant should do. And, importantly, explain why you think this is the case (so that the advice is more likely to be helpful).

How much would you like to bid in the auction to have the advice you gave above be used in our future study (maximum $1.00)? Recall that if your bid is successful, your advice will be given to an average of 40 future participants and you will earn $0.025 for each participant that says your advice is helpful, but you will have to pay the cost of the bid from the initial $1.00 that you received.

The instructions and decision screens for the NATURAL treatment follow. Only the decision screens for the first dataset ($I^+$ in this case) are shown. For brevity, we omit the comprehension questions and the screens that present the dataset, ask for the initial policy choice, ask subjects to state their certainty, and ask subjects to rate the helpfulness/certainty of the narrative on its own. They are identical to those in CONSTRUCTED.

USC University of Southern California

# Instructions (Overview)

In this experiment, you will observe a dataset and make decisions based on it. You will do this in four separate trials (in sequence). The datasets in each trial are **different** and not related to those in any other trial in any way, so nothing that you learn in one trial applies to any of the other trials. In each trial, you should only use the information that is provided to you in that trial.

In each of the four trials, you will make policy decisions. The policy you select determines the probability (percentage chance) you implement one of two choices: the BLUE choice or the GREEN choice.

To help you select your policy, you will receive access to historical data that summarizes many thousands of choices in a series of rows. **Each row in the summary table was historically observed an equal number of times (and so is equally likely).** The data shows the following information:

- Which choice was implemented: BLUE or GREEN.
- A variable simply labeled 'X'. The historical data uses a code to record the 'X' column, but does not indicate what the code means. All you know is that the same code always means the same thing. 'X' is coded as ▲ or ◯.
- Whether the payoff was **HIGH** or LOW.

Your goal is to obtain a **HIGH** payoff. The choices, 'X' values, and payoffs may or may not be related. When you choose your policy, the resulting choice will have the same relationships (if any) with the 'X' variable and payoff as it did in the past. The historical data contains all of the information you need (there are no other 'hidden' variables that matter) so you should use it to determine any relationships.

After each policy decision, you will report your confidence in your policy decision. For the second and third policy decisions in each period, you will receive advice before you make your decision. You will be paid the bonus that results from one of your policy decisions (chosen randomly with equal chance). We describe in more detail where the advice comes from and how your bonus will be determined on the following pages.

When you have read these instructions, please continue to the next page.

USC University of Southern California

# Instructions (Bonus)

Your bonus depends on the policy you select and the payoff. When you select a policy, you will determine the percentage chance that the BLUE choice is implemented. Call this P. The GREEN choice will be implemented with the remaining chance, 100-P. The least costly policy is P=50: the policy in which you implement the BLUE action 50% of the time. If you choose a higher or lower chance of BLUE, the cost is higher.

Specifically, you will get a bonus from two parts (added together):

1. You will receive $1.00 if the payoff is **HIGH** and $0 if the payoff is LOW.
2. You will receive 0.167-0.667x(P/100-0.5)^2 where P is the policy you choose. For example, you will receive $0.167 if you choose P=50 but $0 if you choose P=0 or P=100.

When choosing the policy, think carefully! A more expensive policy might be worth it if it increases the chance that the payoff is **HIGH**. When you have finished reading these instructions, please proceed to the next page.


# Instructions (Advice)

The advice you receive before your second and third policy decisions in each trial will either be generated by us (the people running the experiment) or will come from a previous participant in the experiment. We will always tell you where the advice comes from, but no matter where it comes from, you should always assess the advice for yourself. It may or may not be useful!

If the advice was generated by a previous participant, they saw the same historical data as you (though potentially ordered differently) and were then asked to choose a policy and to give advice to a future participant (yourself!) about how to choose the policy. The previous participant then participated in an auction with other previous participants. The advice you're seeing comes from one of the 5% of previous participants that paid the most to give you their advice. That previous participant will be paid $0.025 if you say their advice was helpful and nothing otherwise. (In some trials, we will give you a second piece of advice and ask you which is helpful. In this case, past participants will be paid based only on your initial indication of helpfulness.) The participants that provide advice in each trial could be the same or different participants.

When you have finished reading these instructions, please proceed to the next page.

USC University of Southern California

# Policy Decision for Trial 1 (first advice)

You have now received advice from **a past participant** who has analyzed the historical data that you just reviewed. The advice is as follows:

*Historical data has shown blue to be more producing in higher payoffs than green.*

Recall that this advice may or may not be useful in helping you make your policy decision. Please think about it carefully to see if it makes sense given the historical information you reviewed (repeated again below).

| Choice | Payoff | X |
|---|---|---|
| BLUE | LOW | ○ |
| GREEN | **HIGH** | ○ |
| GREEN | **HIGH** | ○ |
| BLUE | **HIGH** | ▲ |
| BLUE | **HIGH** | ▲ |
| BLUE | LOW | ○ |
| BLUE | LOW | ○ |
| GREEN | LOW | ○ |
| BLUE | **HIGH** | ▲ |
| GREEN | **HIGH** | ○ |
| BLUE | LOW | ○ |
| GREEN | LOW | ○ |
| BLUE | **HIGH** | ▲ |
| GREEN | **HIGH** | ○ |
| GREEN | LOW | ○ |
| GREEN | LOW | ○ |

Please choose the policy that you want to implement in this period. Recall that the policy you choose determines the percentage chance that you take the BLUE choice.

Less BLUE — More BLUE

0 10 20 30 40 50 60 70 80 90 100

Percent chance of BLUE choice:

USC University of Southern California

# Policy Decision for Trial 1 (second advice)

You have now received a second piece of advice **generated by us (the experimenters)** . The new advice is given below, along with the previous advice (in no particular order).

Advice 1 (from past participant):
*Historical data has shown blue to be more producing in higher payoffs than green.*

Advice 2 (from us):
*Look at this table I created using the historical data. The table shows the number of times the payoff was LOW or HIGH when the choice was BLUE and when it was GREEN:*

| | LOW Payoff | HIGH Payoff |
|---|---|---|
| *BLUE* | *4* | *4* |
| *GREEN* | *4* | *4* |

*You can see that the payoff is HIGH the same number of times under each choice, so that the policy does not matter for the payoff. You should choose the least costly policy (BLUE 50% of the time).*

Recall that each piece of advice may or may not be useful in helping you make your policy decision. Please think about it carefully to see if either piece of advice makes sense given the historical information you reviewed (repeated again below).

| Choice | Payoff | X |
|---|---|---|
| BLUE | LOW | ○ |
| GREEN | **HIGH** | ○ |
| GREEN | **HIGH** | ○ |
| BLUE | **HIGH** | ▲ |
| BLUE | **HIGH** | ▲ |
| BLUE | LOW | ○ |
| BLUE | LOW | ○ |
| GREEN | LOW | ○ |
| BLUE | **HIGH** | ▲ |
| GREEN | **HIGH** | ○ |
| BLUE | LOW | ○ |
| GREEN | LOW | ○ |
| BLUE | **HIGH** | ▲ |
| GREEN | **HIGH** | ○ |
| GREEN | LOW | ○ |
| GREEN | LOW | ○ |

Please choose the policy that you want to implement in this period. Recall that the policy you choose determines the percentage chance that you take the BLUE choice.

Less BLUE — More BLUE

0 10 20 30 40 50 60 70 80 90 100

Percent chance of BLUE choice:

USC University of Southern California

Consider the advice you just received:

*Historical data has shown blue to be more producing in higher payoffs than green.*

Please indicate whether you found this advice helpful in making your policy decision:

- ◯ Helpful
- ◯ Not helpful

Consider the other advice you just received:

*Look at this table I created using the historical data. The table shows the number of times the payoff was LOW or HIGH when the choice was BLUE and when it was GREEN:*

| | **LOW Payoff** | **HIGH Payoff** |
|---|---|---|
| *BLUE* | *4* | *4* |
| *GREEN* | *4* | *4* |

Please indicate whether you found this advice helpful in making your policy decision:

- ◯ Helpful
- ◯ Not helpful

How certain are you that your chosen policy maximizes your bonus? Please use the slider to indicate your level of certainty.

Very Uncertain — Very Certain

0 10 20 30 40 50 60 70 80 90 100

Certainty

# D Narrative Classification

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

Table D1: Classification Descriptions

| Classification | Code | Description |
|---|---|---|
| Reject | REJ | Does not contain an explicit or implicit (describes pattern) policy recommendation |
| Green Other | GO | Suggests green (policy $< 0.5$) but does not describe causal pattern |
| Green 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 |

You can view all 804 elicited narratives and their classifications 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).