

# Narrative Persuasion\*

Kai Barron

WZB Berlin

Tilman Fries

WZB Berlin

September 30, 2022

For the current version, click [here](#).

## Abstract

Financial advisors frequently propose narratives to explain the recent performance of the stock market and to forecast what it will do next. When advisors' incentives are not fully aligned with those of the individuals they are advising, there is scope for self-interested persuasion using narratives. In this paper, we use an experiment to learn about the underlying mechanisms that govern this form of persuasion. Our results reveal several insights. First, we characterize the strategies advisers use to construct persuasive narratives. Second, we show that advisors are able to shift investors' beliefs about the future performance of a company. This harms investors and benefits advisors. Third, we identify the types of narratives that investors adopt. Finally, we evaluate the efficacy of several potential policy interventions aimed at protecting investors. We find that narrative persuasion is difficult to protect against.

**JEL Codes:** D83, G40, G50, C90.

**Keywords:** Narratives, beliefs, finance, conflicts of interest, financial advice, disclosure.

---

\*We are greatly indebted to Jasmin Droege, who played an indispensable role in the initial stages of developing the project. We would also like to thank Chiara Aina, Peter Andre, Daniele Caliri, Felix Chopra, Nicola Gennaioli, Jeanne Hagenbach, Luca Henkel, Emeric Henry, Alessandro Ispano, Agne Kajackaite, Christine Laudenbach, Yiming Liu, George Loewenstein, Thomas Graeber, Katrin Gödker, Salvatore Nunnari, Davide Pace, Chris Roth, Paul Seabright, Heidi Thysen, and Florian Zimmermann for many interesting discussions and helpful suggestions. We thank the WZB for generously funding this project through means of its "seed money" programme and gratefully acknowledge financial support from the *Deutsche Forschungsgemeinschaft* through CRC TRR 190 (project number 280092119). The study was pre-registered in the AEA registry with the unique identifier: AEARCTR-0009103.

# 1 Introduction

Narratives are sense-making devices; they provide an explanation for a collection of events.<sup>1</sup> For example, when people discuss the reasons for the 2007 Financial Crisis, their explanations will typically draw causal links between different events—e.g., between the state of the housing market and stock prices. Learning about the causes of an event is not only useful for understanding the past but also for forming expectations about the future. An individual who believes that the causes of the 2007 Financial Crisis are still present in the financial system may be less willing to invest than an individual who believes that the causes are no longer present. Narratives are also conversational devices. Individuals share them through simple stories or jokes (Shiller, 2017). They may in turn be used by individuals with a vested interest to try to shape how others interpret events. Since information about events is often stored in data sets in modern life, a narrative can also be thought of as providing a causal explanation that organizes the information stored in a dataset.

The traditional approach in economics to studying communication and advice typically focuses on the transmission of factual information. For example, in the canonical cheap talk model (Crawford and Sobel, 1982), the sender sends information about a state of the world to the receiver. In this model, any attempt by the sender to explain the information she sends (e.g., “the firm’s profits are high because of advantageous market conditions”) is superfluous because (i) both sender and receiver agree on the prior distribution over states of the world and because (ii) strategic considerations make it difficult or impossible to achieve informative communication on non payoff-relevant domains. More recently, a nascent economic literature studies agents whose models of the world are misspecified and whose beliefs about the causal relations between different events are malleable (see, e.g., Eliaz and Spiegler, 2020; Schwartzstein and Sunderam, 2021). In this paper, we study empirically the provision of expert advice in a setting where the advisor can send information about a payoff-relevant state along with a narrative to endorse its veracity. This setting allows us to test several implications of recent theories on narratives and to relate them to the cheap talk approach.

We consider a setting in which financial advisors may try to influence investors’ beliefs by proposing subjective models (narratives) to explain the available objective data.<sup>2</sup> In everyday settings, narratives can take many forms and may draw on various sources of theory

---

<sup>1</sup>Currently, there is not a consensus on a single precise definition of the term “narrative” in the economics literature. In Appendix Section A, we provide a detailed discussion of the relationship between different conceptualizations of the concept in the literature and describe the working definition that we use in this paper.

<sup>2</sup>Examples of other domains where narratives may play a key role in shaping the inference drawn from objective data include the following. *Climate change*: lobbyists and politicians propose alternative interpretations of weather data; *Academia*: different academics propose models to organize the available empirical evidence; *Immigration policy*: people circulate stories about the impact of immigrants on crime rates and unemployment levels. The importance of narratives also extends to the *law*, where both sides generally build their case around the same body of evidence (see Pennington and Hastie, 1986, 1988, 1992, for a discussion of the ‘story-model’ of juror decision-making). Finally, during the *COVID-19* pandemic, there was vigorous debate over the correct interpretation of public health data and this appears to have generated polarization in the way that the general population formed beliefs about the health risks of different behaviors (e.g., regarding the efficacy of mask wearing for preventing the spread of COVID-19; see Allcott et al., 2020).

and evidence. This makes it difficult to identify the mechanisms that determine their selection and effectiveness in naturally occurring data. We therefore study narrative construction and adoption in a financial advice experiment. This setting provides us with full knowledge of the information set of both advisors and investors, as well as precise knowledge about the characteristics of the narratives that advisors send to investors. Moreover, it allows us to introduce exogenous variation in the content of these information sets and to elicit the beliefs of all participants. This knowledge is essential for fully understanding how advisors construct the narratives they propose in relation to their beliefs, and also for identifying the causal effects of these narratives on investors' beliefs.

Drawing inspiration from the theoretical work of Schwartzstein and Sunderam (2021) (henceforth S&S), in the experiment we consider a setting with a financial advisor (narrative-sender) and an investor (narrative-recipient). Both individuals observe the historical data from a hypothetical company. The investor wishes to use this information to form an accurate belief about the likelihood that the company is going to be profitable in the coming year. The advisor is more intimately acquainted with the company and therefore knows more than the investor about the true underlying process that generated the public data. Taking into account her superior information, the advisor provides advice to the investor in the form of a narrative—i.e., she proposes an explanation of the underlying process that produced the public historical company performance data. This narrative may guide how the investor interprets the data and influence the beliefs he forms. Importantly, advisors might face a conflict of interest—i.e., the advisor might hold incentives that are misaligned with those of the investor. Such an advisor might use the narrative she sends to try to induce a biased belief in the investor.

The historical company data that the investor and advisor observe consists of a time series which spans a period of ten years. For each year, both individuals see the outcome of a binary variable (success or failure, which reflects whether the company had a profitable or unprofitable year). They know that a simple model generated this data. Specifically, they know that the company had exactly one CEO change during this period and that, prior to the CEO change, the probability that the company was successful in each year was constant ( $\theta_{pre}$ ). When the CEO changed, the company shifted to a new probability of being successful in each year ( $\theta_{post}$ ). The investor does not know exactly in which year  $c$  the CEO changed nor does he know the company's probability of success under the old CEO or the new CEO. In contrast to the investor, the advisor knows the true underlying model. The advisor's task is to send a message to the investor which consists of three parameters,  $(c^A, \theta_{pre}^A, \theta_{post}^A)$ . This is the advisor's narrative. After receiving the message and the historical data, the investor reports his own assessment of the likelihood of the company being successful in the next year—i.e., his assessment of the probability of success under the new CEO. The advisor's message may, therefore, influence how the investor interprets the data, helping the advisor to achieve her own objectives. Advisors in the experiment are one of three types: up-advisors, who are incentivized to persuade investors that the company is likely to be profitable, down-advisors, who are incentivized to persuade investors that the company is not likely to be profitable, and aligned advisors, who are per-

fectly aligned with investors and are incentivized to induce accurate beliefs in their matched investors.

Figure 1: An example of historical company data and a possible narrative.

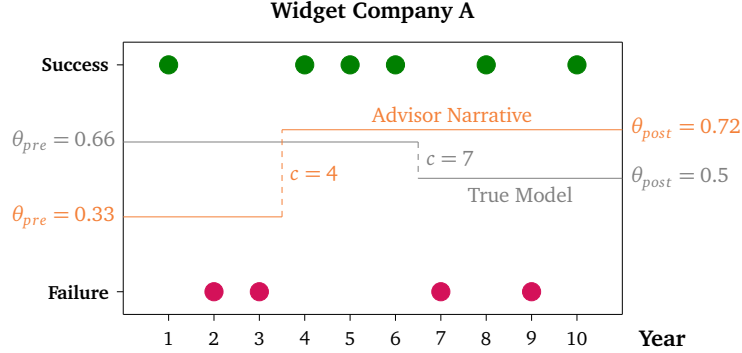


Figure 1 illustrates the basic intuition of narrative persuasion in our experiment. The green and red dots show an example of the historical company data. The grey line indicates an example of a possible true underlying model observed by the advisor, and the orange line illustrates a potential narrative that an up-advisor might use to try to persuade an investor to hold an upward biased belief about  $\theta_{post}$ . Importantly, this example highlights a central feature of narrative persuasion. While the up-advisor only cares about moving the investor’s belief about the probability of success under the new CEO,  $\theta_{post}$ , she can choose the remaining parameter values in a way that improves the fit of the narrative to the data. In the example, she adjusts the year in which the CEO changed from year 7 to year 4 to make it appear as if the new CEO had more successful years. Consequently, according to her narrative, there are fewer successful years during the tenure of the old CEO. The advisor, therefore, correspondingly shifts her assessment of the company’s probability of success under the old CEO downwards.

Following our pre-registration, we use this setting to address four sets of questions. First, what are the strategies that advisors use to construct narratives? Here, we examine how advisors resolve the tension they face between the twin objectives of *moving* the investor’s beliefs away from their prior and constructing a narrative that *fits* the data well and is, therefore, convincing. Second, is narrative persuasive effective? To answer this question, we test whether investors who meet an advisor with a conflict of interest end up holding beliefs that are more biased than investors who meet an advisor who has aligned interests. Third, what are the properties of narratives that are more convincing? Since investors can assess the veracity of a narrative by comparing it to the objective data, we evaluate whether better-fitting narratives are more persuasive. To achieve this, we construct an index that orders narratives according to their empirical fit, conditional on the observed historical data. We then ask whether investors’ beliefs are more consistent with the narrative they receive when this narrative has a higher likelihood of having generated the objective data.

Fourth, can policy interventions be used to protect individuals from being harmed by narrative persuasion? Using three treatment conditions, we consider three candidate policies aimed

at protecting investors. In *DISCLOSURE*, investors are fully informed about the incentives of every advisor that they meet. This implies that they know exactly when they face an advisor with a conflict of interest and when they face an advisor whose incentives are perfectly aligned with their own. This may protect investors by allowing them to be more skeptical of advice received from conflicted advisors. In *INVESTORPRIOR*, investors are encouraged to take some time to assess the objective company data themselves prior to meeting with their advisor. The intuition behind this intervention is that investors who have conducted a careful assessment of the objective data themselves may be less willing to believe narratives that do not fit the data well. Finally, in *PRIVATEDATA* advisors are not provided with access to the objective data that the investors see. While advisors are still more informed than investors about the company and the true underlying process, they are now unable to tailor their narrative to the information set of the investors. This may protect investors, since it makes it more difficult for advisors to propose biased narratives that they fit *ex post* to the precise realization of the data in the investor's information set.

We correspondingly document four sets of results. Taken together these results are largely in line with several of the core ideas developed in *S&S*. First, focusing on advisors, we find that those with a conflict of interest do try to take advantage of the opportunity to persuade by transmitting a biased narrative to the investor. Furthermore, they engage in a fairly sophisticated form of self-interested narrative persuasion in which they don't only distort their assessment of the company's probability of success under the new CEO (the target belief that they wish to influence), but also adjust their assessment of the other auxiliary narrative components in order to try to make their narrative more convincing. Second, we show that conflicted advisors are successful in shifting the beliefs of investors—investors that meet a conflicted advisor form beliefs that are further from the truth than those that meet an aligned advisor. Furthermore, investors that meet an up-advisor form more optimistic beliefs about the company than those that meet a down-advisor. Third, we show that when advisors construct a narrative that fits the data better, investors form beliefs that are closer to the advisor's narrative.

Fourth, our evaluation of the three potential policy interventions reveals that none of them is successful in providing protection to the average investor. Their beliefs are equally far from the truth in these three treatments as in *BASELINE*. However, these average results obscure an interesting (non-preregistered) finding. In the *DISCLOSURE* treatment, investors know exactly when they meet a conflicted advisor and do become more skeptical of narratives received from such advisors. This does protect investors when they meet a conflicted advisor who is lying to them. However, sometimes even conflicted advisors choose to tell the truth and not construct a biased narrative (possibly due to truth-telling preferences). Investors cannot easily distinguish truth-telling conflicted advisors from lying conflicted advisors. Therefore, when they become more skeptical in *DISCLOSURE*, they also become more skeptical of the information received from truth-telling conflicted advisors and this harms them. On average, these more skeptical investors, therefore, do no better in the *DISCLOSURE* treatment. In the *PRIVATEDATA* treatment we find similar increases in investors' skepticism and find that investors indeed become more

skeptical of messages which are lies. This is in line with the idea that lying becomes more difficult for advisors when they cannot send tailored narratives. Overall, the findings from the policy interventions highlight that narrative persuasion is difficult to protect against. Being able to compare a message to objective data can add (perceived) credibility to the message—when a narrative fits objective data well, it can seem compelling.

Finally, our experiment can also be viewed through the lens of a more traditional cheap talk model in which advisors and investors are assumed to be strategically sophisticated. We, therefore, also provide a theoretical analysis of our setting from this perspective and derive predictions for behavior that differ from those prescribed by the S&S framework. The empirical data that we collect in our experiment are more in line with the predictions of S&S. This suggests that in our setting investors focus more on the fit of the narrative relative to the data when deciding whether they find it persuasive, rather than engaging in strategic thinking. This evidence that individuals neglect to fully engage in sophisticated strategic thinking, and instead reduce the complexity of the problem by reformulating it as a slightly different problem that they can solve more easily to arrive at their decision-relevant beliefs is consistent with an array of well-documented behavioral biases including cursedness (Eyster and Rabin, 2005), selection neglect (Jehiel, 2018; Barron, Huck, and Jehiel, 2019; Enke, 2020), correlation neglect (Eyster and Weizsacker, 2016; Enke and Zimmermann, 2019; Laudenbach, Ungeheuer, and Weber, 2019). Furthermore, relative to the more complex everyday scenarios where narratives may be used to influence individuals, in our experiment, participants are better informed about the structure of the game and incentives of everyone involved, which should imply a bias towards strategic thinking.

The remainder of the paper proceeds as follows. Section 2 discusses the relationship to the extant literature. Section 3 develops the theoretical framework. Section 4 describes the experimental design. In Section 5, we present the results and Section 6 contains a concluding discussion.

## 2 Relationship to the Literature

Our results contribute to several strands of literature. First, many other academic disciplines have attributed a central role to narratives in understanding human behavior, including the analysis of ideology and belief systems in political science, sociology and psychology (Mannheim, 2015 [1936]; Converse, 2006 [1964]; Bruner, 1991; Haidt, 2007, 2012; Charnysh, 2021), discourse analysis and narrative analysis in sociology (Foucault, 1972; Franzosi, 1998; Polletta, Chen, Gardner, and Motes, 2011), and narrative analysis in literary and cultural studies (Koschorke, 2018; Herman and Vervaeck, 2019). Economics has been substantially slower in adopting this perspective with the orthodox economic model assuming that individuals hold in mind the correct model of the world—i.e., that they interpret the data they observe through the lens of this correct model. However, recent work in economic theory has ar-



gued that narratives play an important role in shaping economic outcomes. This literature has begun to explore the consequences of individuals holding (possibly incorrect) subjective models of the world (Spiegler, 2016; Shiller, 2017; Heidhues, Kőszegi, and Strack, 2018; Bénabou, Falk, and Tirole, 2020; Eliaz and Spiegler, 2020; Spiegler, 2020a,b; Mailath and Samuelson, 2020; Schwartzstein and Sunderam, 2021; Aina, 2021; Olea, Ortoleva, Pai, and Prat, 2021; Schumacher and Thysen, 2022; Ispano, 2022). Despite this activity in economic theory, there remains a scarcity of direct empirical evidence on the role of narratives in economics.<sup>3</sup>

To the best of our knowledge, we are the first to document experimental evidence on the role of narratives as a tool for persuasion. We do this by analyzing the decision problems faced by both the narrative-sender and the narrative-recipient. In doing so, we contribute evidence towards understanding a class of situations where narratives may play a key role—strategic settings in which one individual may transmit a narrative to another in order to influence how they interpret objective data. This constitutes an important class of situations, since many modern policy decisions are influenced by discussions around the interpretation of rapidly expanding accumulation of publicly accessible data. Our paper shows how individuals with a vested interest may leverage such discussions by constructing *ex post* interpretations of public information, thereby persuading others to adopt their preferred interpretation. The possibility to compare the narrative to public information that is known to be objectively true may lend credibility to the narrative, even when it is known that the narrative was constructed *ex post*. This allows individuals with a vested interest to construct a narrative that serves their own purposes, but also appears to be coherent with the objective public information.

Second, our work relates closely to the sender-receiver literature in which a better-informed sender sends a message to a receiver, and the receiver takes an action that influences the welfare of both (Crawford and Sobel, 1982). While this work has given rise to a large body of experimental work on cheap talk models (see, e.g., Blume, DeJong, Kim, and Sprinkle, 1998; Blume, DeJong, Neumann, and Savin, 2002; Wang, Spezio, and Camerer, 2010) and also on communication with evidence in the form of disclosure games (see, e.g., King and Wallin, 1991; Hagenbach and Perez-Richet, 2018; Jin, Luca, and Martin, 2021), our work differs from this previous literature due to the focus on the interpretation of objective public data. Specifically, advisors in our experiment send messages that not only provide information on the payoff-relevant parameter but also on non payoff-relevant parameters, which they might choose in ways that justify their payoff-relevant parameter choice. Since these messages provide a lens for interpreting the data, they can be evaluated against it. Our setting reflects many common scenarios in modern everyday life, where huge quantities of information are increasingly available and there is scope for manipulating the beliefs of others by shaping how they interpret this information. We discuss the relationship between the narrative persuasion theoretical frame-

---

<sup>3</sup>There are some noteworthy recent exceptions to this that have analyzed the role of narratives empirically in contexts very different from the one considered in this paper (Andre, Haaland, Roth, and Wohlfart, 2022; Harrs, Müller, and Rockenbach, 2021; Morag and Loewenstein, 2021; Hagmann, Minson, Tinsley, et al., 2021; Laudenbach, Weber, and Wohlfart, 2021; Andre, Pizzinelli, Roth, and Wohlfart, 2022; Barron, Harmgart, Huck, Schneider, and Sutter, 2022; Hillenbrand and Verrina, 2022). These are discussed in more detail below.

work and the sender-receiver theoretical approach in detail in Section 3.5. When discussing the results of the experiment, we also test the data against predictions of the most persuasive equilibrium of the cheap talk game underlying our setup.

Third, our findings on the difficulty of protecting investors from harmful persuasion relate to a string of papers showing that the disclosure of conflicts of interest may backfire (see, e.g., Cain, Loewenstein, and Moore, 2005; Malmendier and Shanthikumar, 2007; Loewenstein, Sah, and Cain, 2012; Sah, Loewenstein, and Cain, 2013). This literature delineates several mechanisms through which disclosure may be undermined.<sup>4</sup> For example, Cain et al. (2005) show that senders react to the introduction of disclosure rules by shifting the message that they send to receivers even further from the truth and receivers fail to adequately account for this reaction. Our experimental design rules out this mechanism as the advisors see identical instructions in our BASELINE and DISCLOSURE treatments. However, our results reveal a new channel that may undermine the effectiveness of disclosure rules. When there is a sizable fraction of honest advisors amongst the set of advisors who have a conflict of interest, the effectiveness of disclosure rules may be reduced by the fact that they make investors equally skeptical of advice received from both honest and dishonest advisors.

Finally, our work is related to a very recent empirical literature that explores how narratives (broadly construed) shape behavior. For example, Andre et al. (2022) study households' subjective beliefs about the responsiveness of key economic variables to macroeconomic shocks and Andre et al. (2022) provide causal evidence on how individuals construct narratives to explain the evolution of inflation rates and how these narratives in turn influence the interpretation of new information. Laudenbach et al. (2021) show that investor's beliefs about the autocorrelation of aggregate stock returns can be improved by providing them with information about the correct underlying model. Turning to COVID-19, Harris et al. (2021) examine how optimistic versus pessimistic narratives about the pandemic affected economically relevant behavior. In the domain of pro-social behavior, Barron et al. (2022) show that when parents believe certain narratives about refugees, this can affect the pro-social behavior of their children, while Hillenbrand and Verrina (2022) also show that stories can be used to influence prosocial behavior. Finally, Morag and Loewenstein (2021) find evidence that the act of telling a story about an owned object increases one's valuation of the object.

Our study differs from this work in several important ways. First, different to this literature, we focus on the particular phenomenon of the use of narratives in a strategic setting, where one individual wishes to use a narrative to persuade another in their interpretation of objective information. The papers mentioned above are broadly interested in a completely different class of situations where narratives also play an important role. Second, while some contributions in this literature conceptualize narratives in a broad sense, including stories and informal models,

---

<sup>4</sup>One example of a context in which mandatory and voluntary conflict of interest disclosures may be effective is provided by Sah and Loewenstein (2014). The authors show that when advisors have the opportunity to choose to credibly avoid the conflict of interest entirely, they may select into a non-conflicted environment and thereby avoid having to signal that they have a conflict of interest. This benefits advisees.



we focus on a particular conceptualization of a narrative as a subjective model explaining a particular process (Andre et al., 2022, adopt a similar approach, but build on the machinery of directed acyclic graphs (DAGs)). Third, while much of this work does not try to fully account for the full information set of the individuals being studied (due to addressing completely different types of research questions), our experimental design provides us with full control over subjects’ information sets, and allows us to introduce several layers of exogenous variation, which provides the opportunity to analyze the comparative statics we are interested in.

### 3 Theoretical Framework

In this section, we develop a theoretical framework that we use as a lens to zoom in on specific features of the investor-advisor setup that we then study empirically using our experiment. The framework draws heavily on the one proposed by S&S. In contrast to traditional game-theoretic approaches, this framework dispenses with equilibrium reasoning by assuming that the narrative-recipient (in our case, the investor) credulously adopts a narrative if it explains the observed historical data sufficiently well. This captures the idea that individuals may focus on the narrative’s fit when deciding whether to adopt a new narrative rather than engaging in equilibrium reasoning. At the end of the section, we also discuss the predictions of a model in which investors are strategically sophisticated and compare the differences between the two theoretical approaches in more detail.

#### 3.1 Model persuasion setup

We consider a setup with an investor (“he”) and an advisor (“she”). The setup will closely follow our baseline experimental design. In this setting, the investor’s goal is to form an accurate belief about a company’s future success probability. To form that belief, the investor may draw on the advisor’s advice and the historical data.

**Historical data and the data generating process** The investor and advisor both have access to a time series of the historical performance data from a company. For each year  $t$  in the data set, the company can either have a success-year, which we denote by  $s_t = 1$ , or a failure-year, which we denote by  $s_t = 0$ . The history  $h$  is the sequence of successes and failures from years 1 to 10;  $h \equiv (s_t)_{t=1}^{10}$ .

Underlying the historical data is a data generating process consisting of three parameters. First, the data-generating process contains a structural change parameter  $c^T$  which divides the years observed in the data set into a *pre* and a *post* period. (In the experiment, this structural change is framed as a change in the company’s CEO.) This structural change takes place at some point between years 2 and 8. Second, the parameter  $\theta_{pre}^T$  denotes the company’s success probability in each of the years 1 up to  $c^T$ . Finally, the parameter  $\theta_{post}^T$  denotes the company’s success probability in the years after  $c^T$ . The true underlying model is thus given

by  $(c, \theta_{pre}, \theta_{post}) \in \mathcal{M} \equiv \{2, \dots, 8\} \times [0, 1]^2$ . The investor and advisor both know that the true underlying model is part of this set.

**Investor** The investor is uncertain about the true model,  $m^T$ , governing the success and failure of the company during the period observed in the historical data set. He will form a belief about it. Based on that belief, the investor will make an assessment,  $\theta_{post}^I$ , of the company's probability of success in the *post* period.

The investor can draw on several pieces of information to form his assessment. First, he can use the information contained in the historical data set. Based on this information, the investor forms a subjective model or default narrative—his own private initial interpretation of the data—which we denote by  $m^{I,0} \in \mathcal{M}$ . This subjective model is best thought of as being exogenously given as part of the investor's endowment. S&S discuss a number of focal cases of interest; a default which does not allow the investor to draw any information from the data and a default which is equal to the true model. Both of these benchmark cases seem slightly unsatisfactory in our setting. Since the investor knows that a specific statistical process (from a known class of models) has generated the company history, he should be able to infer some information about the company's quality from observing the history. This does not mean, however, that the investor is able to infer the truth. Different true models can generate the same history. Therefore, assuming that the investor can infer the truth from the observed history would be a very strong and rather unrealistic assumption to make. Instead, we will treat the investor's default narrative as a random variable that is distributed according to a density function  $f(m)$  which has full support on  $\mathcal{M}$ .<sup>5</sup>

Second, in addition to the default model, the investor also receives advice. This advice arrives in the form of message  $m^A \in \mathcal{M}$ , sent by the advisor. The investor observes  $m^A$  before making an assessment of  $\theta_{post}$ . The investor may construct his assessment in two ways—either he reports the corresponding parameter value contained in his own default model or he reports the one contained in the advisor's message. We say that the investor adopts the advisor's message whenever  $\theta_{post}^I = \theta_{post}^A$ .

One key ingredient of S&S's setup is the assumption that the investor adopts the advisor's narrative only after it passes a “Bayesian hypothesis test”. This means that the investor adopts the narrative suggested by the advisor if the observed history is at least as likely under the advisor's proposed narrative as under the investor's default narrative;

$$\Pr(h|m^A) \geq \Pr(h|m^{I,0}). \quad (1)$$

In our setting, this is equivalent to saying that the investor picks the narrative with the better fit as measured by the log likelihood function. We will denote the log likelihood function by  $\ell(m)$  and the narrative in  $\mathcal{M}$  that maximizes the likelihood function by  $m^{DO}$ . This narrative

---

<sup>5</sup>A more general definition would condition the prior distribution on  $h$  to account for the fact that the investor's default model can be different for different realizations of  $h$ . Since we will not study comparative statics with respect to changes in  $h$  we refrain from conditioning here and in the following definitions to simplify notation.

is *data-optimal* (DO) in the sense that the investor will always adopt it upon receiving it as a message.

**Advisor** The advisor's objective is to send a message that induces the investor to make an assessment that is as close as possible to the advisor's desired assessment. Depending on the investor's assessment  $\theta_{post}^I$  and the advisor's desired investor assessment  $\varphi$ , the advisor receives utility

$$U(\theta_{post}^I, \varphi) = 1 - (\varphi - \theta_{post}^I)^2.$$

This utility is maximized if  $\theta_{post}^I = \varphi$ . The exact value of  $\varphi$  depends on the advisor's type. The up-advisor wants the investor to make the highest possible assessment, and thus has  $\varphi = 1$ . The down-advisor has  $\varphi = 0$ , i.e. this type wants the investor to make the lowest possible assessment. The aligned advisor wants the investor to make an accurate assessment; for this type  $\varphi = \theta_{post}^T$ . When sending a message, the advisor does not know the investor's default model and therefore cannot be sure whether the investor will adopt or not. Her message thus induces a lottery over investor assessments where the advisor knows that the investor will adopt the  $\theta_{post}^A$  if the fit of  $m^A$  to the objective data is high enough. The probability of the investor adopting the advisor's model is then given by a c.d.f.  $G(\ell(m^A))$ , which increases in the advisor's model fit,  $\ell(m^A)$ , and which has full support on the interval  $(-\infty, \ell(m^{DO})]$ . We can derive this function directly from the investor's default model distribution  $f(m)$ .<sup>6</sup> The advisor chooses  $m^A$  to solve:

$$\max_{m^A \in M} \mathbb{E}[U(\theta_{post}^I, \varphi)|m^A], \text{ where}$$

$$\mathbb{E}[U(\theta_{post}^I, \varphi)|m^A] = G(\ell(m^A))U(\theta_{post}^A, \varphi) + (1 - G(\ell(m^A)))\mathbb{E}[U(\theta_{post}^{I,0}, \varphi)|\ell(m^A) < \ell(m^{I,0})]. \quad (2)$$

The advisor thus chooses a message which maximizes a convex combination of the utility obtained from the investor's assessment when the investor adopts the advisor's message (the first term above) and when he does not adopt (the second term above). When constructing the message, the advisor cannot be sure which assessment the investor will make when he does not adopt. The advisor thus has to form an expectation about the consequences of the investor not adopting, which in the maximization problem is given by the conditional expectation term  $\mathbb{E}[U(\theta_{post}^{I,0}, \varphi)|\ell(m^A) < \ell(m^{I,0})]$ .

### 3.2 Discussion of the model

At the core of the model is the advisor's problem of constructing a message that the investor will adopt and that also induces an assessment that is close to the advisor's objective. The advisor will often face a tension between these two motives (fit and movement) and, as we

---

<sup>6</sup>In particular,  $G(l) \equiv \int_{-\infty}^l g(s) ds$  where  $g(l) \equiv \int_{m \in M} I(\ell(m) = l) f(m) dm$ . To establish that  $G$  has full support on  $(-\infty, \ell(m^{DO})]$ , note that  $\ell((c^{DO}, \theta_{pre}^{DO}, \theta_{post}))$  is continuous for values  $\theta_{post} \in (0, 1)$  and that  $\ell(\cdot)$  will always cover the full range of values between  $-\infty$  and  $\ell(m^{DO})$ . This observation together with the assumption that  $f$  has full support on  $\mathcal{M}$  implies full support of  $G$ .

will discuss in the next part, this tension leads to systematic predictions about the structure of the advisor’s message. The theoretical framework that we use to derive these predictions embeds several assumptions. Here, we discuss the merits of the main assumptions made.

**Common knowledge about the set of possible true underlying models.** The model considers a restricted set of possible data generating processes,  $\mathcal{M}$ , which is common knowledge. This restricts the investor’s and advisor’s attention to narratives characterized by three parameters; the structural break, and the *pre* and *post* success probabilities. Therefore, the investor is not “maximally open to persuasion” (S&S) as other types of narratives that are outside  $\mathcal{M}$  are ruled out. The investor knows that exactly one structural change occurred within the company but is unsure about exactly when it happened. He is also uncertain about the consequences of the change. Constraining the set of true underlying models in this particular way allows us to characterize the kind of messages that the advisor will send to the investor.

**The investor’s decision rule.** The investor adopts a message whenever it provides a higher empirical fit than the default narrative. The investor is thus *credulous* as he does not think through the strategic incentives of the advisor when judging a message. In addition, he is also *skeptical* in that he only adopts a message if it provides a better explanation of the historical data. As we will explain further below, we can generalize this rule to allow the investor to be skeptical to different degrees (by applying a “fit penalty” to messages received from advisors who might not be trustworthy). The credulity assumption plays a key role in generating the prediction that the advisor can, through carefully choosing  $c$  and  $\theta_{pre}$ , construct a narrative that induces the investor to adopt the  $\theta_{post}$  parameter value sent by the advisor.

**The investor’s default model.** In contrast to S&S, we do not assume that the advisor necessarily knows the investor’s default model when constructing the message. This seems more realistic for our context since we assume that the investor’s default model is characterized by a random variable.<sup>7</sup> When the advisor does not know the default, she might fail to always induce the investor to adopt her message. This is in contrast to the full knowledge case studied by S&S where the advisor optimally constructs a message that the investor always adopts. Therefore, the advisor in our setting has to form a belief about what the investor’s assessment will be if he does not adopt the narrative contained in the advisor’s message.

### 3.3 The structure of the advisor’s message

The setup above generates predictions about how the advisor will construct her message. We will focus on the impact of the motive to move the investor’s belief first. To do that, we will focus on the set of messages with the same fit  $\bar{\ell}$ . Denote this set of messages by  $\mathcal{M}(\bar{\ell})$ . Our

---

<sup>7</sup>One can think of an advisor who expects to be matched with an investor drawn from the population of investors. The advisor knows the distribution of default narratives held by investors in the population, but does not know that of the specific investor she will advise.

first result states that, among all messages in  $\mathcal{M}(\bar{\ell})$ , the advisor will choose the message  $m^*(\bar{\ell})$  whose  $\theta_{post}$  value is closest to the advisor's objective.

**Proposition 1.** *Among all possible messages with the same message fit  $\bar{\ell}$ , the advisor chooses the message which minimizes the distance between  $\theta_{post}^A$  and  $\varphi$ ;*

$$m^*(\bar{\ell}) \in \arg \min_{m \in \mathcal{M}(\bar{\ell})} (\varphi - \theta_{post})^2.$$

We can see how this result follows from Equation (2). Fixing the message fit at  $\bar{\ell}$ , the only moving part in the expected utility function is the utility that the advisor will receive if the investor adopts the message. Therefore, the advisor will prefer the message which moves the investor's assessment closest to her target in case of adoption. A second observation is that the advisor will choose the  $c^A$  and  $\theta_{pre}^A$  parameters to maximize the message fit *conditional* on the chosen  $\theta_{post}^A$ . This occurs because the advisor is only directly incentivized to *move* the investor's  $\theta_{post}$ -assessment in a certain direction; she does not have any direct incentive to shift the investor's beliefs about the other two parameters. Therefore, if we hold  $\theta_{post}^A$  fixed, improving the *fit* of the message is the sole criterion driving the advisor's choice of the two remaining parameters. Essentially, the advisor chooses these non payoff-relevant message components to construct a narrative that makes the  $\theta_{post}$  component of the message appear more plausible in view of the historical data.

**Proposition 2.** *In the advisor's optimal message  $m^* = (c^*, \theta_{pre}^*, \theta_{post}^*)$ ,  $c^*$  and  $\theta_{pre}^*$  maximize the log likelihood function conditional on  $\theta_{post}^*$ ;*

$$(c^*, \theta_{pre}^*) \in \arg \max_{(c, \theta_{pre}) \in \{2, \dots, 8\} \times [0, 1]} \ell(c, \theta_{pre}, \theta_{post}^*).$$

Since each  $\theta_{post}^A$  will yield a pair,  $(c^A, \theta_{pre}^A)$ , that maximizes the fit (possibly non-uniquely), the advisor can then compare these messages along the continuum of  $\theta_{post}^A$ s to choose the message that best trades off fit against movement.

Based on this reasoning, the set of models that can be part of the advisor's message are represented by what we call the "likelihood frontier". To provide an illustrative example, Figure 2 plots the up- and down-advisor's likelihood frontier for a specific history,  $h = (0, 1, 1, 0, 1, 1, 1, 1, 0, 1)$ . The graph in the figure plots the highest message fit (as measured by the log likelihood function) that the advisor can obtain for each possible value of  $\theta_{post}$ . The log likelihood function takes on its maximum value when the message equals the data-optimal model,  $(c^{DO} = 4, \theta_{pre}^{DO} = 2/4, \theta_{post}^{DO} = 5/6)$ .<sup>8</sup> At this point, the up- and the down-advisor's likelihood frontiers, as illustrated by the red and blue lines, meet. The up-advisor's likelihood frontier also includes all  $\theta_{post}$  values between the data-optimum and 1, as each of these messages can be rationalized under some intensity of the tradeoff between movement

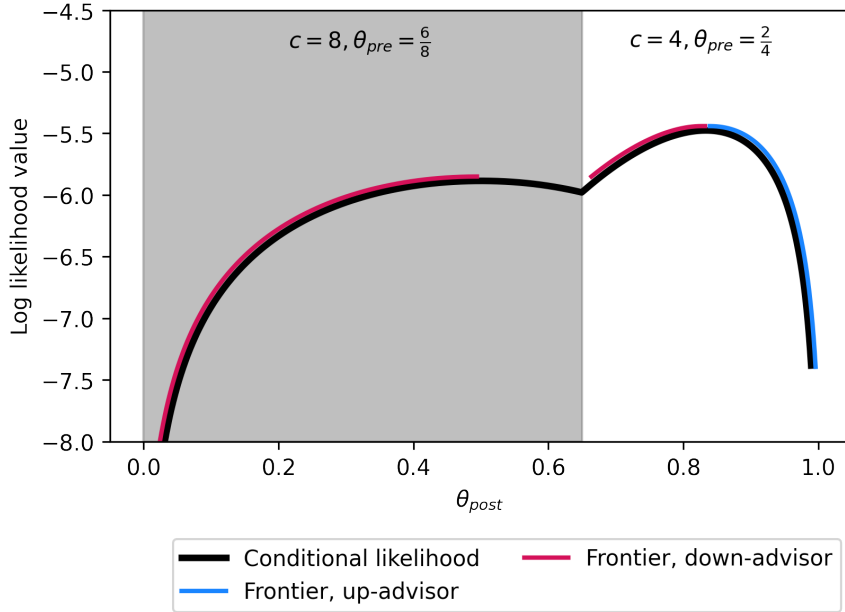
---

<sup>8</sup>Note that, conditional on  $c = 4$ , the data-optimal  $\theta_{pre}$  and  $\theta_{post}$  are simply equal to the proportion of successes in their respective periods.

and model fit. The likelihood frontier of the down-advisor is instead discontinuous because a range of messages with intermediate  $\theta_{post}$  parameter values around 0.6 is dominated by messages with lower  $\theta_{post}$  values which are both closer to the down-advisor's objective of 0 and provide a better fit.

One can use this figure to think about how the advisor resolves the trade-off between movement and fit. If the advisor believes that the investor likely holds a default model which is close to the data-optimal model, then she will send a message where  $\theta_{post}^A$  is close to  $\theta_{post}^{DO}$ . The advisor does this because she believes that the investor will compare the message to a default model that already fits the data well, and this limits the movement that the advisor is able to induce. If, instead, the advisor believes that the investor likely holds a default model that does not fit the data well, the up-advisor will increase  $\theta_{post}^A$  while the down advisor will decrease it.<sup>9</sup>

Figure 2: Likelihood frontiers of up- and down-advisors for an example history



*Notes:* The figure plots the likelihood frontiers and conditional log-likelihood function for all possible values of  $\theta_{post}$  and example history  $h = (0, 1, 1, 0, 1, 1, 1, 0, 1)$ . The  $c$  and  $\theta_{pre}$  values at the top of the figure maximize the conditional maximum likelihood in the respective range of  $\theta_{post}$  values. It is also worth noting that not all histories can be visually represented as simply as this one—for example, this history induces only two possible optimal  $c$  values in the frontiers of the up and down-advisors, which facilitates the relatively simple visual representation.

Figure 2 also illustrates how the advisor's motive to move the investor's assessment,  $\theta_{post}^I$ , towards the advisor's target,  $\varphi$ , induces the advisor to systematically deviate from the data-optimal structural change parameter,  $c^{DO}$ . To see this, consider the likelihood frontier of the down-advisor. As the down-advisor lowers  $\theta_{post}$  from the data-optimal value,  $\theta_{post}^{DO} = 5/6$ , it becomes optimal for her to move the structural change from year 4 towards year 8 (i.e., when

<sup>9</sup>Similarly, the optimal  $\theta_{post}^A$  of the aligned advisor lies between  $\theta_{post}^{DO}$  and the true parameter,  $\theta_{post}^T$ . The  $\theta_{post}$  sent by the aligned advisor will move closer to the data-optimal model as the trade-off between movement and fit becomes sharper.



she wishes to induce a  $\theta_{post}$  that is in the darker shaded region). A later structural change better rationalizes a low  $\theta_{post}$  than an earlier structural change—while the proportion of successes in the post period is 5/6 under  $c = 4$ , it declines to 1/2 under  $c = 8$ . This is because the history of the company under consideration contains 4 successful years between year 4 and year 8.

More generally, we can think about the motives that an advisor has for deviating from the data-optimal structural change parameter  $c^{DO}$  in the following way: Consider an advisor who is forced to send  $c^{DO}$  as part of her message. This advisor will then choose a value for  $\theta_{pre}$  that maximizes the likelihood conditional on  $c^{DO}$ , and which we denote by  $\hat{\theta}_{pre}(c^{DO})$ . Her chosen parameter for the post period,  $\theta_{post}^*$ , will then maximize the expected utility function conditional on  $c^{DO}$  and  $\hat{\theta}_{pre}(c^{DO})$  (i.e.  $\theta_{post}^* = \arg \max_{\theta_{post}} \mathbb{E}[U(\theta_{post}^I, \varphi) | (c^{DO}, \hat{\theta}_{pre}(c^{DO}), \theta_{post})]$ ). Now consider providing the advisor with the flexibility to adjust  $c$ , keeping  $\theta_{post}^*$  fixed. When would the advisor prefer to send a model  $(c', \hat{\theta}_{pre}(c'), \theta_{post}^*)$  instead of  $(c^{DO}, \hat{\theta}_{pre}(c^{DO}), \theta_{post}^*)$ ? We provide two results, one for the case where the advisor considers a lower  $c$  and one where the advisor considers a higher  $c$ .

**Proposition 3.** *Denote the optimal message of an advisor who is forced to send  $c^{DO}$  by  $\tilde{m} = (c^{DO}, \hat{\theta}_{pre}(c^{DO}), \theta_{post}^*)$ . Consider an advisor who chooses between this message and a message  $m' = (c', \hat{\theta}_{pre}(c'), \theta_{post}^*)$  where  $c' < c^{DO}$ .*

1. *The up-advisor prefers to send  $m'$  over  $\tilde{m}$  only if*

$$\frac{\sum_{t=c'+1}^{10} s_t}{10 - c'} > \frac{\sum_{t=c^{DO}+1}^{10} s_t}{10 - c^{DO}}.$$

2. *The down-advisor prefers to send  $m'$  over  $\tilde{m}$  only if*

$$\frac{\sum_{t=c'+1}^{10} s_t}{10 - c'} < \frac{\sum_{t=c^{DO}+1}^{10} s_t}{10 - c^{DO}}.$$

When the advisor lowers  $c$ , she is essentially shifting years from the *pre* period into the *post* period within the narrative under consideration. The statement above says that the advisor will only do this if it shifts the fraction of successful years in the *post* period closer to the advisor's desired assessment (i.e., an increase in the fraction of successes for the up-advisor and a decrease for the down-advisor).

The logic of the advisor's problem is slightly different when she considers deviating to an alternative  $c'$  that is larger than  $c^{DO}$  and, therefore, she is now shifting years from the *post* period into the *pre* period. As the following proposition shows, an up-advisor will now only choose the alternative threshold if it reduces the *number* of failures, and a down-advisor only if it reduces the number of successes, in the *post* period.

**Proposition 4.** *Denote the optimal message of an advisor who is forced to send  $c^{DO}$  by  $\tilde{m} = (c^{DO}, \hat{\theta}_{pre}(c^{DO}), \theta_{post}^*)$ . Consider an advisor who chooses between this message and a message  $m' = (c', \hat{\theta}_{pre}(c'), \theta_{post}^*)$  where  $c' > c^{DO}$ .*

1. The up-advisor prefers to send  $m'$  over  $\tilde{m}$  only if

$$\sum_{t=c'+1}^{10} (1-s_t) < \sum_{t=c^{DO}+1}^{10} (1-s_t).$$

2. The down-advisor prefers to send  $m'$  over  $\tilde{m}$  only if

$$\sum_{t=c'+1}^{10} s_t < \sum_{t=c^{DO}+1}^{10} s_t.$$

Two motives direct the advisor's choice of  $c$ . First, the advisor wants to minimize any discrepancy between the empirical proportion of successes implied by  $c$  and the  $\theta_{post}^*$  she sends. Since  $\theta_{post}^*$  is larger than the data-optimal value for the up-advisor, this motivates the up-advisor to choose a  $c$  which increases the empirical proportion of successes in the *post* period. Conversely, it provides the down-advisor with a motive to choose a  $c$  which decreases the empirical proportion of successes in the *post* period. We can see how this motive guides the advisor's decision when she considers lowering  $c$ , as shown in Proposition 3. Second, when  $\theta_{post}^*$  takes an extreme value—e.g., it is close to 1 for an up-advisor—then the advisor wants to minimize the *number* of failures (as opposed to the fraction) in *post* to justify this extreme value. Similarly, a down-advisor with a  $\theta_{post}^*$  close to 0 wishes to minimize the number of successes. This follows from the non-linearity of the log likelihood function which implies that, e.g., for a large  $\theta_{post}$ , any failure in the post period receives a large penalty. Intuitively, any failure becomes increasingly difficult to explain as the success probability increases. This is the rationale behind Proposition 4: When the up-advisor considers increasing  $c$ , she will always do so if her  $\theta_{post}^*$  is high enough and if increasing  $c$  decreases the numbers of failures in the *post* period. A similar logic applies for the down-advisor.

The two propositions provide necessary conditions for the advisor to deviate from reporting the data-optimal structural change. Whether the advisor actually prefers to deviate will depend on the distance between  $\theta_{post}^*$  and  $\theta_{post}^{DO}$ . If the distance is very small (i.e., if there is a sharp tradeoff between fit and movement) then the advisor will be unlikely to deviate to the alternative  $c$ ; since the utility-optimal message under the data-optimal cutoff is close to the data-optimal message, such a deviation will decrease the empirical fit. As the distance between  $\theta_{post}^*$  and  $\theta_{post}^{DO}$  increases, it becomes more likely that coupling  $\theta_{post}^*$  with a different structural change parameter value will lead to an improved message fit relative to coupling it with  $c^{DO}$ .

One noteworthy implication of the advisor's desire to construct a narrative that supports the veracity of the  $\theta_{post}^A$  she sends is that it will push  $\theta_{pre}^A$  into the opposite direction; for example, as the up-advisor increases the proportion of successes and minimizes the number of failures in the *post* period, this will tend to increase the proportion of failures in the *pre* period. This dynamic will lead to a negative correlation between  $\theta_{pre}$  and  $\theta_{post}$  in the messages of misaligned

advisors. Even if the true underlying parameters are drawn independently under the true data-generating process, the advisor's objective to construct messages which embody a compromise between movement and fit will lead to messages that shift the  $\theta_{pre}$  and  $\theta_{post}$  parameters in opposite directions.

### 3.4 Interventions that aim to protect the investor from persuasion

In the theoretical framework, the investor adopts or rejects the advisor's narrative based on a rule or heuristic: When the advisor's narrative makes more sense to him when held up to the data in comparison to the prior narrative that he previously held, he will adopt it, otherwise he will not. Given this rule-based decision-making, the theory does not endogenously predict changes in the investor's behavior based on changes in the decision environment. Through the lens of the narrative persuasion framework, however, we can ask which factors prevent the investor from adopting the advisor's narrative. To the extent that we think of persuasion being harmful, interventions which address these factors might thus protect the investor from harmful persuasion.

**Restricting access to the history.** According to the theory, the advisor tries to send a narrative which fits the history well. The investor can be persuaded by such a narrative because he disregards the fact that the advisor constructed the narrative ex-post, after observing the data. If access to the data is restricted such that only the investor may access it, the advisor loses the opportunity to tailor the narrative to the data. As a consequence, the average expected fit of the advisor's narrative will decrease. The investor will in turn be less likely to adopt the narrative proposed by the advisor if he has exclusive access to the history.

**Encouraging the investor to make sense of the history.** In the narrative persuasion framework, the advisor can sometimes convince the investor to adopt a different narrative only because the investor's default narrative fits poorly, so that there is an alternative narrative with a better fit that the advisor can potentially send. If the investor were instead to always adopt the data-optimal narrative as a default, then no better-fitting alternative narrative exists and the investor would never move away from his default. To try to improve the fit of the investor's default, one potential intervention is one that encourages the investor to reflect on his own interpretation of the observed history *before* being exposed to the advisor's narrative. Encouraging a more carefully chosen default might improve its fit and bring it closer to the data-optimum, which in turn would make the investor more immune to adopting the advisor's narrative.<sup>10</sup>

---

<sup>10</sup>In the theoretical framework, the default narrative is distributed according to a density  $f(m)$ , which implies some distribution of the default narratives' likelihood values and which we denote by  $G(l)$ . Encouraging a more carefully chosen default changes its prior density to  $\tilde{f}$  and the corresponding distribution of likelihood values to  $\tilde{G}$ . One can think about encouragement of a more carefully chosen default as inducing a density which is more concentrated around narratives close to the data-optimal narrative, resulting in a distribution  $\tilde{G}$  that first-order stochastically dominates  $G$ .

**Disclosing the advisor’s incentives.** The investor in the narrative persuasion framework is non-strategic. He selects among narratives based purely on fit, without taking the advisor’s incentives into account when deciding whether to adopt or not. This non-strategic approach to decision-making does not imply that the investor cannot be *skeptical*. As discussed by S&S, the investor might have a more or less demanding narrative adoption criterion. For example, he might penalize the fit of narratives received from the advisor relative to his default (or he might only penalize narratives received from advisors when he knows that they have a conflict of interest). We can capture this by modifying the adoption criterion provided in Equation (1) to:

$$\Pr(h|m^A) \geq \Pr(h|m^{I,0}) + s,$$

where  $s \geq 0$  is a parameter that quantifies the investor’s degree of skepticism; a strictly positive parameter value implies that the investor only adopts a narrative which explains the data substantially better (and not merely better) than the default narrative. Revealing to the investor that the advisor’s incentives are misaligned with his own might raise the investor’s awareness that the message might potentially be inaccurate. In turn, this could lead to the investor becoming more skeptical of messages received from a misaligned advisor. Therefore, an intervention that discloses the advisor’s incentives to the investor could make it more difficult for a misaligned advisor to persuade the investor.

### 3.5 An alternative approach to persuasion: cheap talk

An appealing feature of our experiment is that it lends itself to different theoretical approaches, which allows us to examine which theory is more consistent with the behavior we observe. In Appendix E, we dispose of the assumption that the investor can always be persuaded by messages which provide a sufficiently good fit. We instead assume that the investor is strategically sophisticated. Therefore, the investor takes into account all the information contained in the observed history and thinks about the incentives of the advisor types they might potentially meet. Formally, this transforms the setup into a cheap talk game between the advisor and the investor, where the investor’s prior over the true  $\theta_{post}$  is common knowledge. While it is relatively well known that there are no informative equilibria if the advisor (the “sender” in a traditional cheap talk setup) has state-independent preferences (e.g. Little, 2022), the experiment introduces uncertainty about whether the advisor’s preferences are state-dependent (the aligned advisor) or not (the up- and down-advisor). We show that this introduces some scope for persuasive communication: An equilibrium exists which is characterized by an interval of admissible values of  $\theta_{post}$  around the investor’s prior belief about  $\theta_{post}$ . In equilibrium, all messages sent by the advisor include a  $\theta_{post}$  parameter inside this interval and the investor will always adopt.

The strategic framework highlights another potential constraint on persuasion: sophisticated investors will not adopt “extreme” messages, i.e., messages that either contain a high or

a low  $\theta_{post}$  value outside the interval. A key difference from the S&S framework is that in the strategic framework the  $\theta_{post}$  parameter is never part of a broader narrative where the advisor uses the  $c$  and  $\theta_{pre}$  components of their message to make their communicated  $\theta_{post}$  seem more compelling. The reason for this is that, if everyone is strategic, talk about  $c$  and  $\theta_{pre}$  is cheap; no type of advisor has an incentive to report information about these parameters truthfully. Therefore, whichever equilibrium strategies advisors follow in choosing the payoff-relevant parameter values, the resulting interval of admissible  $\theta_{post}$  values is uniquely determined for any historical data set and invariant to changes in the non payoff-relevant parameter equilibrium strategies.<sup>11</sup>

## 4 Experimental Design

We conduct an experiment where participants take on the role of either an investor or a financial advisor. The experiment is framed as an investment game, where the investor’s goal is to accurately assess a company’s future success probability. To make an accurate assessment, the investor can draw on two sources of information. First, the investor observes historical performance data for the company, which contains information about whether the company achieved success or failure in each of the past ten years. Second, the investor receives advice from an advisor. The experiment consists of ten rounds with a new company evaluated in every round.

### 4.1 BASELINE treatment

In each round of the experiment, the investor is randomly matched with an advisor. Decision making proceeds in two steps. First, the advisor observes the historical data from a hypothetical company. This data shows whether the company was “successful” or “unsuccessful” in each of the past ten years. In the experiment, the years are labeled from 1 to 10. The advisor then constructs a message to send to the investor, which consists of the three parameters underlying the company’s success and failure during the ten-year period. Second, the investor then receives the historical data and the advisor’s message simultaneously, with both types of information being presented side-by-side on the same screen. Based on this information, the investor assesses the company’s current success probability.

#### 4.1.1 Information environment

**The data-generating process.** Both the investor and the advisor are told that, in each year of the company’s ten-year history, the probability of the company being successful was determined by an underlying fundamental parameter,  $\theta$ . The investor and advisor both know that

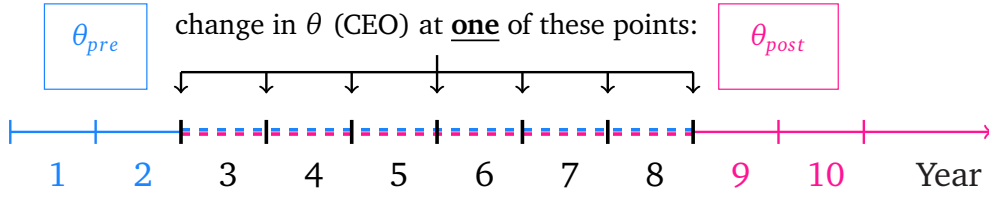
---

<sup>11</sup>Of course, there are multiple equilibria where the investor does not follow the message off the equilibrium path for certain values of  $\theta_{pre}$  and  $c$ . However, any equilibrium induces the same equilibrium allocations.

$\theta$  can take on values between 0 and 1. This fundamental changed exactly once during the ten years—i.e., this parameter was redrawn exactly once at some point between Year 2 and Year 8 (inclusive). The investor and advisor are truthfully told that both the initial probability of success ( $\theta_{pre}$ ) and the current probability of success ( $\theta_{post}$ ) were drawn from a uniform distribution,  $\theta_x \sim U[0, 1]$ . Likewise, they are also truthfully told that the year of the structural change was drawn from a discrete uniform distribution—i.e., with an equal probability of drawing each of the years 2 to 8. All parameter values are independent of one another.

Figure 3 illustrates the structure of the historical data.

Figure 3: Structure of the historical data



As shown in the figure, the last two periods in the historical dataset are commonly known to be: (i) governed by a different probability of success to the first two periods, and (ii) informative about the current and future success probability of the company. This is because participants are certain that the current CEO was in charge of the company for at least the last two years (and at most the last eight years).

**The Advisor’s Additional Information.** The advisor has full information about the underlying data generating model—i.e. the advisor knows the true values of the three fundamental parameters ( $c^T, \theta_{pre}^T, \theta_{post}^T$ ). The investor knows that the advisor has this additional information.

**Feedback.** Investors and advisors go through the ten rounds of the experiment without receiving any feedback during the experiment. For example, advisors do not receive feedback about their matched investors’ assessments and neither advisors nor investors receive intermediate feedback about their payoffs. We do this to minimize learning and the potential interdependence of choices across the ten rounds of the experiment.

#### 4.1.2 Choices and incentives

In each round of the experiment, the advisor sends a message to an investor which consists of the parameters ( $c, \theta_{pre}, \theta_{post}$ ). When composing the message, the advisor can send any parameter values which are plausible—i.e., they cannot propose parameters that are outside feasible set. In particular, this means that the advisor has to choose a value for  $c$  which lies



between 2 and 8.<sup>12</sup> There are no further constraints on the messages that the advisor can send—i.e., the advisor is not required to send a truthful message. Upon receiving the message and inspecting the data, the investor submits his own estimate of  $\theta_{post}$ .

Participants in the role of the investor are incentivized to estimate  $\theta_{post}$  as close as possible to  $\theta_{post}^T$ . We use the binarized scoring rule (BSR; Hossain and Okui, 2013) to ensure that investors will maximize their expected payment when reporting their expectation about  $\theta_{post}$  truthfully.

Participants in the role of the advisor are assigned to one of three incentive conditions. In all three conditions, the advisor’s payment depends exclusively on their matched investor’s  $\theta_{post}$ -assessment. Under the three conditions, the advisor is either: (a) an up-advisor whose payoff increases in the investor’s estimate of  $\theta_{post}$ , (b) a down-advisor whose payoff decreases in the investor’s estimate of  $\theta_{post}$ , or (c) an aligned advisor whose payoff increases in the accuracy of the investor’s estimate of  $\theta_{post}$ . We use a strategic version of the BSR to determine the payoffs of each type of advisor. This strategic version differs in two ways from the standard version of the BSR. First, the belief report that is relevant for determining the probability of receiving the bonus payment is made by the investor, not the advisor (i.e., the beliefs of the investor determine the advisor’s payment). Second the BSR of up- and down-advisors compares the  $\theta_{post}$  reported by the investor to extreme  $\theta_{post}$  values, namely  $\theta_{post} = 1$  or  $\theta_{post} = 0$ , to determine the advisor’s payment. This induces these advisors to want investors to hold high or low beliefs and differs from the standard BSR which typically compares the reported belief to the truth—i.e.,  $\theta_{post}$ . The aligned advisor’s BSR is equivalent to the investor’s BSR (i.e., their incentives are perfectly aligned). This strategic version of the BSR is, therefore, useful for inducing particular preferences in one individual over the beliefs held by another individual.

The assignment of advisors to incentive conditions in our experiment is random, with each incentive condition being equally likely. Importantly, each advisor is assigned to a particular incentive condition at the beginning of the experiment and stays in that incentive condition throughout the ten rounds of the experiment.

**Strategic Information about Incentives:** Investors are fully informed about the different types of advisors that they may face. Specifically, they are told about the three types of advisors and that the chance of being matched with each type is 1/3 in every round of the experiment. However, they are not informed about the specific incentives of the advisor that they are matched with in a particular period.

Advisors know the incentives of investors. In all treatment conditions, advisors are also always told that investors may or may not know their matched advisor’s incentives.<sup>13</sup>

---

<sup>12</sup>The year of the structural change in the experiment was framed as denoting the *first* year under the new CEO. Therefore, advisors could actually choose numbers between 3-9. For expositional clarity and coherence between the discussion of the theory and the experiment, throughout the paper we will continue using the convention that the structural change parameter denotes the *last* year under the old CEO. All variables in the analysis have been re-coded in a way that is consistent with the “last year under the old CEO” convention.

<sup>13</sup>This design feature allows us to keep the advisor’s instructions constant between the BASELINE, SKEPTICISM, and DISCLOSURE treatment. We discuss the additional treatments further below.

### 4.1.3 General Comments about the Design

There are two features of our experimental design that warrant further explanation. First, to introduce an asymmetry in expertise about the companies between the advisor and investor in a controlled way, we opted to inform the advisor about the true underlying DGP ( $c^T, \theta_{pre}^T, \theta_{post}^T$ ) of each company. This serves to provide an opportunity for gains from communication for the investor, since the advisor is more informed. Depending on the advisor's incentives, the advisor might sometimes try to deceive the investor into reporting an overly optimistic or pessimistic belief about  $\theta_{post}$ . Specifically, the advisor can use the other dimensions of the report,  $c$  and  $\theta_{pre}$ , as supporting evidence in trying to shift this belief about  $\theta_{post}$ . Informing the advisor about the true model also fixes normative expectations about what the advisor *should* do. This replicates an essential feature of advisor-investor relations in real life; even though it would be unrealistic to assume that financial advisors know the underlying fundamentals of the firms and markets they analyze, they often know more than the investor—e.g., they might know the industry consensus or have access to additional information or better information-processing tools. Furthermore, advisors are typically expected to provide advice that is accurate to the best of their knowledge. Informing the investor about the true model is a simple intervention that controls for these (first- and higher-order) normative expectations, making it clear that an advisor who deviates from reporting the true DGP is doing so intentionally with the aim of persuading the investor.<sup>14</sup>

Second, in most of our treatment conditions (including our BASELINE treatment), we chose not to elicit investor's prior beliefs about the default model (i.e., the belief based on seeing only the historical data). We have three reasons for this. The first reason is that we wish to study scenarios in which advisors present data to investors at the same time as they communicate their theory explaining the data, as opposed to situations where the receiver first constructs their own personal theory of the data. This conjunction of receiving the data along with a potential sense-making explanation mimics situations in which the data arrives alongside a ready interpretation from an interested party. The second reason is that we wish to explicitly study whether being encouraged to form a personal theory of the data *prior* to receiving a potential explanation from an advisor has a protective function that helps to insulate investors from persuasion. One of our intervention treatments discussed below encourages investors to form their own subjective assessment after seeing the data but before receiving the advisor's message. Our third reason is that omitting this initial elicitation stage from most treatments helps to simplify the experiment and promote participant understanding.

---

<sup>14</sup>Fixing these advisor beliefs about the true underlying DGP also avoids introducing an additional layer of endogeneity that would be present if advisors first formed their own assessment of the data and only then constructed a message to the investor. In addition to the endogeneity introduced, doing this would require additionally eliciting advisors beliefs about the underlying data generating process, which would add complexity to the experiment. Nevertheless, we believe that this is a promising avenue for future research and discuss this further in our conclusion section.

## 4.2 Intervention treatments

To study potential mechanisms that might protect investors from harmful persuasion, we introduce three treatment conditions that each vary a specific feature of the BASELINE setting. The treatments reflect the interventions whose theoretical implications we discussed in section 3.4 above.

**PRIVATE DATA:** To investigate whether having access to private data serves a protective role against persuasion, we vary whether the advisor observes the historical performance dataset. In particular, both the investor and advisor in this treatment know that the advisor does not observe the historical performance dataset when choosing their message. The advisor, therefore, knows the true underlying parameters of the data generating process, and is still able to try to persuade the investor by sending an inaccurate message, but is unable to tailor the message to the data that the investor observes. This may make it more difficult for the advisor to send a message that is both deceptive and persuasive.

**DISCLOSURE:** To investigate whether knowing their specific matched advisor’s incentives makes investors skeptical, we introduce a treatment that discloses advisor incentives to investors. In each round of the experiment, on the decision screen, advisors in DISCLOSURE learn whether they received a message from an up-, down- or aligned advisor.

**INVESTOR PRIOR:** In this treatment, we examine the effect of being encouraged to form a default (or prior) theory about the data generating process *before* entertaining theories received from others. Specifically, instead of receiving the historical data and the advisor’s message simultaneously, and only then forming a belief about the data generating process, in this treatment investors will first receive only the data. We will then ask them to report their prior belief about the data generating process (i.e.,  $c$ ,  $\theta_{pre}$ , and  $\theta_{post}$ ). Thereafter, investors receive the advisor’s message, and we elicit their final assessment of  $\theta_{post}$ .

This treatment will allow us to evaluate whether being encouraged to try to make sense of the data oneself first serves a protective function against persuasion using models.<sup>15</sup>

Since we are mainly interested in how the interventions potentially change the model adoption of investors, we keep the changes to the advisor instructions as minimal as possible. In particular, advisors in BASELINE, DISCLOSURE, and INVESTOR PRIOR see exactly the same instructions. This makes it possible for us to hold advisor behavior constant across these treatments, and therefore being able to attribute potential treatment effects to changes in investor behavior. It also provides us with clean data on how advisors craft messages for many different

---

<sup>15</sup>An additional benefit of this treatment is that the reported prior beliefs will provide us with some descriptive information about the types of subjective models that investors construct in the absence of messages from advisors. It also allows us to examine updating of beliefs.

combinations of the true underlying model and the observed historical data set to study message formation in detail. Since the PRIVATE DATA treatment studies an intervention which constraints advisors in their ability to tailor models to the data, this treatment necessarily changes the advisor’s instructions in addition to introducing changes in the investor’s instructions.

### 4.3 Procedures

The experiment was pre-registered in the AEA registry (AEARCTR-0009103) and conducted via the Prolific platform. We devoted substantial attention in the design of the experiment towards ensuring that we explained this setup to participants as clearly and intuitively as possible to ensure maximum understanding. We also included several understanding questions that participants were required to answer correctly before proceeding. Section F contains screenshots of the instructions that investors received in our BASELINE treatment.

Participants took part in the experiment in groups of 6. Within each group, 3 participants were randomly assigned to the role of the sender (advisor) and 3 are assigned to the role of the receiver (investor). Each advisor was then randomly assigned to one of the three incentive conditions (i.e., there will be one advisor from each of the three incentive conditions within each group of 6). Both advisors and investors kept their role for the duration of the experiment. Upon clicking on the link to participate in the study, participants were randomly allocated to one treatment. Therefore, the randomization to treatments controls for potential weekday and daytime effects.<sup>16</sup>

In each of the ten rounds of the experiment, each investor was randomly matched with an advisor within their group of six (i.e., the three investors were randomly matched with the three advisors). All matched investor-advisor pairs saw data generated by the same true model in each round of the experiment. Specifically, we drew ten triplets of fundamentals,  $(c^T, \theta_{pre}^T, \theta_{post}^T)$ , before the first session. The sequence in which participants were exposed to each underlying true model was constant in all sessions and treatments of the experiments. Conditional on these fundamentals, however, the observed historical data of success and failure of the company was drawn independently for each investor-advisor pair and round.

In addition to a participation fee of £3.50, participants were paid for one randomly chosen round of the experiment. The additional bonus that the investors and advisors could earn was £3.75, with the probability of earning that bonus depending on the various binarized scoring rules evaluated at the investor’s assessment. After finishing all ten rounds of the experiment, participants answered a short demographic questionnaire. Participants took around 20-25 minutes for the experiment.

---

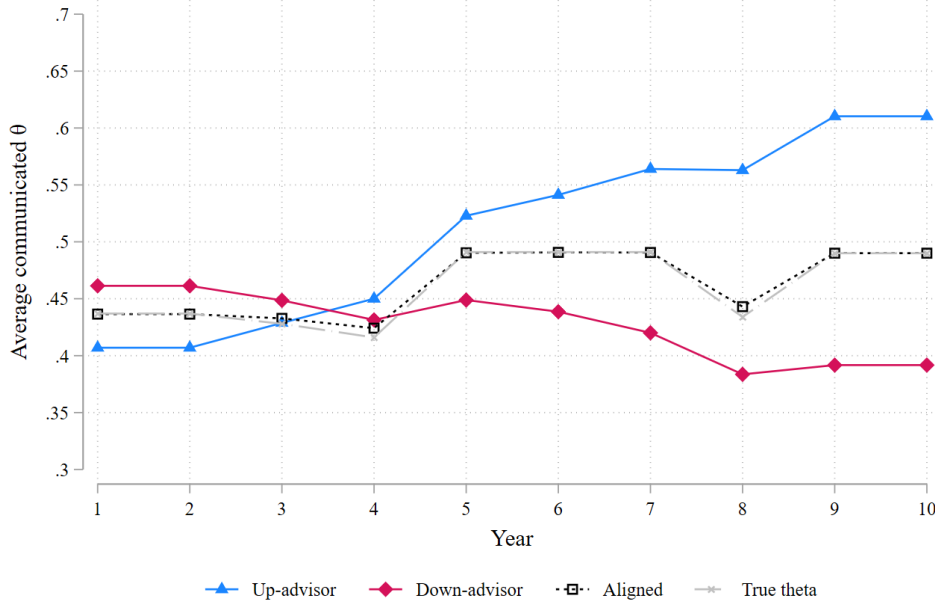
<sup>16</sup>We collected the advisor data for each group one day before we collected the investor data. Therefore, participants were not randomly allocated to a particular role conditional on the session. We did this so that participants did not have to wait for their group members to finish their assessments or messages before moving on to the next round in an effort to minimize attrition.

## 5 Results

### 5.1 Advisor Narrative Construction

Following Schwartzstein and Sunderam (2021), one can think of “model persuaders” facing a tradeoff between *movement* and *fit*. If advisors wish to induce substantial *movement* in investors’ beliefs, this means that they want to propose a narrative (subjective model) that deviates from the truth. However, they also want investors to find the proposed narrative compelling when compared to the data—narratives that *fit* the objective data well are likely to be more compelling. Here, we try to better understand how our advisors construct the narratives that they transmit to investors. To do this, we ask the following questions. First, do advisors with a conflict of interest (i.e., misaligned incentives) construct narratives that are biased towards their own self-interest? Second, what strategies do advisors use in constructing these narratives?

Figure 4: Average narrative communicated by advisors (by incentive type)



#### 5.1.1 Do advisors construct self-interested narratives?

Figure 4 provides an illustration of the raw data. It depicts the average narrative transmitted by the advisors of each incentive type. Specifically, every narrative sent by an advisor implies a probability of success of the company,  $\theta$ , in each of the ten years—this is given by the  $\theta_{pre}$  from the period before the CEO change and the  $\theta_{post}$  from the period after the CEO change. To obtain Figure 4, we take the average  $\theta$  for each year across all message sent by advisors of each type. We can see that up-advisors (denoted by the blue line) construct messages that imply a *higher*  $\theta$  in year 10 than down-advisors (denoted by the red line). Conversely, up-advisors send messages with a *lower*  $\theta$  in year 1 than down-advisors. As one might expect, the messages sent by the aligned advisors imply  $\theta$ ’s between those of the two misaligned advisor types.

Table 1 provides further statistical evidence in support of the patterns shown in the figure. It reports the results from two pre-registered regressions that test whether advisors with misaligned incentives transmit messages that deviate further from the truth than advisors with aligned incentives. In accordance with the advisor’s twin goals of movement and fit, we hypothesize that the  $\theta_{post}$  and  $\theta_{pre}$  parameters sent by misaligned advisors are further from the truth.

**Hypothesis 1a.** [PR.6a] *The distance between the sender’s message and the truth of the post report,  $|\theta_{post}^A - \theta_{post}^T|$ , is larger for misaligned senders than for aligned senders.*

**Hypothesis 1b.** [PR.6b] *The distance between the sender’s message and the truth of the pre report,  $|\theta_{pre}^A - \theta_{pre}^T|$ , is larger for misaligned senders than for aligned senders.*

Column (1) shows that the average misaligned advisor reports a  $\theta_{post}$  that is 13pp further from the truth than the average aligned advisor. Similarly, column (2) shows that advisors also shift the  $\theta_{pre}$  component of the narrative 6pp further from the truth when they hold misaligned incentives. Thus, we do not only find statistical evidence that advisors adjust  $\theta_{post}$  in response to the incentives they face, but also find evidence that is consistent with a more sophisticated narrative construction strategy, where advisors shift their assessment of the company’s historical success probability,  $\theta_{pre}$ , in order to try to improve the fit of their narrative and make it more compelling to the investor.

Table 1: Distance from the truth of narratives proposed by misaligned vs aligned advisors

	(1) $ \theta_{post}^A - \theta_{post}^T $	(2) $ \theta_{pre}^A - \theta_{pre}^T $
Misaligned advisor = 1	12.72*** (0.702)	6.492*** (0.660)
Observations	3600	3600
Round FE	Yes	Yes

(i) The dependent variable is the distance between the true  $\theta$  parameter and the corresponding  $\theta$  parameter of the advisor’s message, (ii) Standard errors are clustered at the advisor level, reported in parentheses \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ , (iii) There are 360 clusters, (iv) The sample contains all advisors from the three treatment conditions which share the exact same set of instructions, (v) For each advisor we have 10 observations—one for each round.

### 5.1.2 What strategies do advisors use to construct their narratives?

The results reported so far outline some general patterns in the message construction of advisors. We now examine the strategies followed by advisors in closer detail by using the exogenous variation provided by the different randomly generated company histories. This allows us to provide evidence on the underlying mechanisms generating the broader patterns that we observe in advisor behavior.



**Using  $\theta_{pre}$  and  $c$  to construct a convincing narrative:** We use the repeated nature of the experiment to classify advisors by the different strategies they might take when constructing their narratives. One measure of narrative production skill is whether and how frequently advisors adjust the non payoff-relevant parameters in ways that support their persuasion target. A proxy for this is the advisors' choice of  $c$ : an advisor who chooses to deviate from the true parameter  $c^T$  to an alternative value that better justifies a high (up-advisor) or low (down-advisor) parameter value of  $\theta_{post}$  uses the structural break parameter to their advantage.

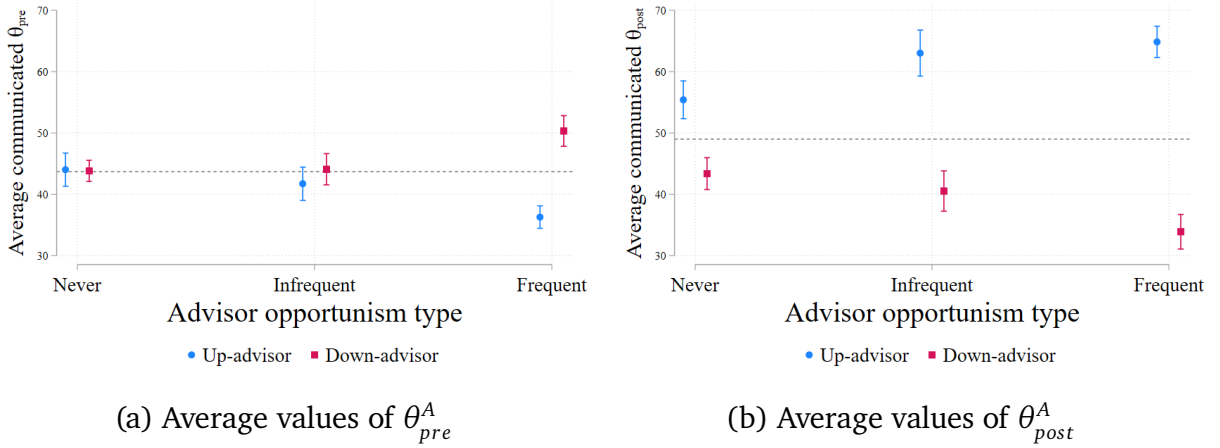
We generate a measure of "opportunism" for each misaligned advisor in the experiment by calculating how frequently they chose such an advantageous  $c$ -value over the course of the ten rounds. We identify three roughly equally sized groups of misaligned advisors based on this measure: misaligned advisors who always transmit the true  $c^T$ , misaligned advisors who choose an advantageous structural break at least once but fewer than 50% of the time, and misaligned advisors who choose an advantageous structural break at least 50% of the time. We call these groups the never-, infrequent-, and frequent-opportunists, respectively.<sup>17</sup>

Based on this classification, we can ask what kind of messages advisors of each opportunism type send. Figure 5 shows the average  $\theta_{pre}$  and  $\theta_{post}$  parameter values sent by each advisor type. The left panel shows the average parameter values in the *pre* period and the right panel shows the average parameter values in the *post* period. A number of insights emerge: First, both up- and down-advisors who are never-opportunists still moderately bias their  $\theta_{post}$ -reports towards their persuasion goal. However, more opportunistic types bias their  $\theta_{post}$ -reports by more. Second, never-opportunists do not on average bias their  $\theta_{pre}$ -report away from the true parameter value, in line with the idea that they are not constructing narratives in a sophisticated way. We can see that, on the opposite, the difference in average  $\theta_{pre}$ -reports is driven by the frequent opportunists, who are the only types which systematically bias  $\theta_{pre}$  in an opposite direction to  $\theta_{post}$ .

---

<sup>17</sup>Among the 240 misaligned advisors who received the BASELINE instructions, 79 are never-, 80 are infrequent-, and 81 are frequent-opportunists.

Figure 5: Average  $\theta^A$ s, by opportunism type



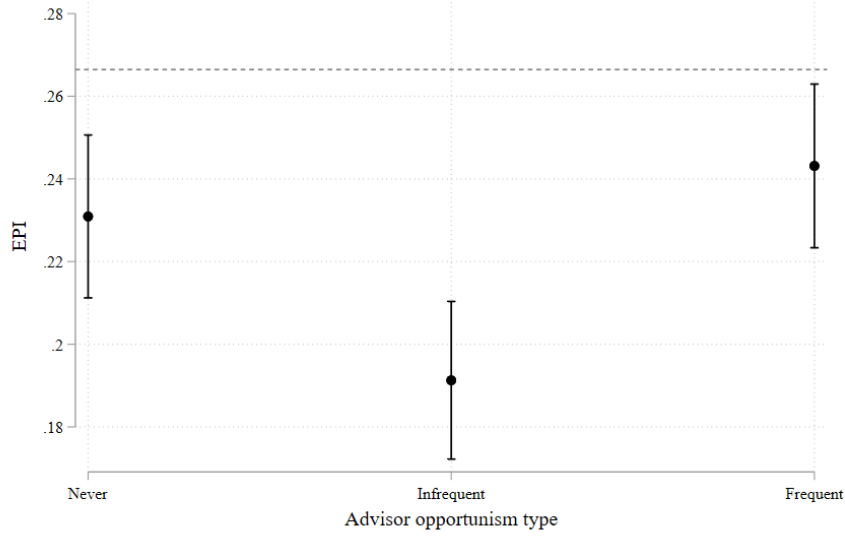
Note: The dashed line denotes the average true  $\theta_{pre}$  encountered by advisors in BASELINE. Error bars are 95% confidence intervals derived from regressions which cluster standard errors at the advisor level.

These patterns also affect narrative quality or message fit, which we measure with what we call the Empirical Plausibility Index (EPI). To derive this value, we calculate the likelihood value of the narrative sent by the advisor in relation to the relevant realization of the historical data set. The EPI is then equal to this likelihood value divided by the likelihood value obtained by the data-optimal narrative for the relevant history.<sup>18</sup> Therefore, the EPI takes on values between 0 and 1. A value of 1 can be obtained by a narrative that is equal to the data-optimal (best-fitting) narrative and a minimum value of 0 is obtained if the advisor sends the worst-fitting narrative.<sup>19</sup> Figure 6 shows that both never and frequent opportunists achieve similar levels of message fit, while infrequent opportunists construct messages of lower fit. Never and frequent opportunists achieve model fits close to the true model for different reasons: Messages of never opportunists are often close to the true model, which makes their fits similar. Frequent opportunists on the other hand introduce a high bias in  $\theta_{post}$ , but, by adjusting the non payoff-relevant parameters, they achieve comparable model fits. The infrequent opportunists are the ones who bias  $\theta_{post}$  by relatively high amounts without adjusting the non payoff-relevant parameter values, which leads to worse message fits.

<sup>18</sup>For a more detailed discussion of the construction of the EPI, please refer to our pre-registration document, where the EPI is discussed on pages 8-9 in Section 3 and also on pages 19-20 in Appendix Section A.

<sup>19</sup>For each history, the lowest possible value is always equal to zero. This is because there exists a narrative with a likelihood value of zero for any history.

Figure 6: Average EPI of advisor messages, by opportunism type



Note: The dashed line denotes the average EPI of the true data generating model encountered by advisors in BASELINE. Error bars are 95% confidence intervals derived from regressions which cluster standard errors at the advisor level.

**Balancing persuasiveness against the truth (aligned advisors):** We also examine the role played by the belief movement and fit motives in the message construction of the aligned advisor. The aligned advisor knows that the investor will compare the narrative she sends to the objective data to assess how convincing it is. In the absence of truth-telling preferences, the aligned advisor has no interest in reporting the true model, but rather wants to send a message that: (i) fits the data well, and (ii) induces a belief that is close to the truth. If the message that fits the data best induces a belief in the investor that is “close” to the truth,  $\theta_{post}^T$ , the advisor may wish to send this data-optimal narrative to the investor. This logic suggests that when the exogenous variation in the historical data is such that that true model does not actually fit the data well—i.e., the data-optimal value  $\theta_{post}^{DO}$  is far from the true value  $\theta_{post}^T$ —aligned advisors will send a message that contains a  $\theta_{post}^A$  value that is further from  $\theta_{post}^{DO}$ . In other words, we hypothesize that the average aligned advisor will follow a strategy that involves sending a  $\theta_{post}^A$  that is a weighted average of the truth,  $\theta_{post}^T$ , and the data-optimal narrative,  $\theta_{post}^{DO}$ .

**Hypothesis 2a.** [PR.7a] *The distance between the data-optimal model and the aligned advisor’s message,  $|\theta_{post}^A - \theta_{post}^{DO}|$ , increases in the distance between the truth and the data optimal report  $|\theta_{post}^T - \theta_{post}^{DO}|$ .*

**Gravitational pull of the truth is weaker for misaligned advisors:** For the misaligned advisors, the true model should not play a role unless truth-telling preferences influence the narratives they construct. Misaligned advisors face monetary incentives to draw the investor’s belief away from the truth. They are constrained only by the investor’s information set (i.e., the historical data) and their own truth-telling preferences. If they hold no truth-telling preferences,

they will completely disregard the truth and it will play no role in influencing the narrative they construct. In the following hypothesis we check (a) whether truth-telling preferences influence misaligned advisors and (b) whether the size of this influence (pull towards the truth) is smaller than it is for aligned advisors.

**Hypothesis 2b.** [PR.7b] *The distance between the data-optimal model and the misaligned advisor's report is governed to a lesser extent by the size of  $|\theta_{post}^T - \theta_{post}^{DO}|$  than in the aligned advisor's report.*

We test both hypotheses by estimating the following model for advisors from the pooled BASELINE, SKEPTICISM, and INVESTORPRIOR treatments (the three treatments where advisors receive identical instructions):

$$|\theta_{post}^A - \theta_{post}^{DO}| = \beta_0 + \beta_1 \mathbb{I}(\text{Misaligned}) + (\beta_2 + \beta_3 \mathbb{I}(\text{Misaligned})) \cdot |\theta_{post}^T - \theta_{post}^{DO}| + \rho_r + \varepsilon$$

In the equation above,  $\mathbb{I}(\text{Misaligned})$  is an indicator function which takes a value of 1 if the advisor's incentives are misaligned,  $\rho_r$  are round fixed effects and  $\varepsilon$  is an error term.<sup>20</sup> Table 2 reports the results. We test Hypothesis 2a by examining  $\beta_2$ . Since  $\beta_2$  is statistically greater than 0, we find evidence in support of Hypothesis 2a. Specifically, the aligned advisors' is biased away from the data-optimal model towards the truth. This indicates that aligned advisors are motivated both by their monetary incentives and also by truth-telling preferences. The magnitude of this coefficient suggests that truth-telling is the dominant approach adopted by aligned advisors.

---

<sup>20</sup>Since the true model is held constant within each round of the experiment, the  $\rho_r$  parameters absorb both round and true model fixed effects. We account for repeated observations by clustering errors at the advisor level when studying advisor outcomes. When studying investor outcomes, we instead cluster at the Interaction Group level to account for potential additional Interaction Group spillovers. It is worth noting that since advisors receive no feedback at all during the experiment, the within Interaction Group spillovers are more limited in scope than usual in experiments where subjects interact in groups. In our experiment, interaction between players only operates in one direction: from advisors to investors via the messages. Investors also do not receive any feedback on the outcomes of their decisions prior to the end of the experiment.

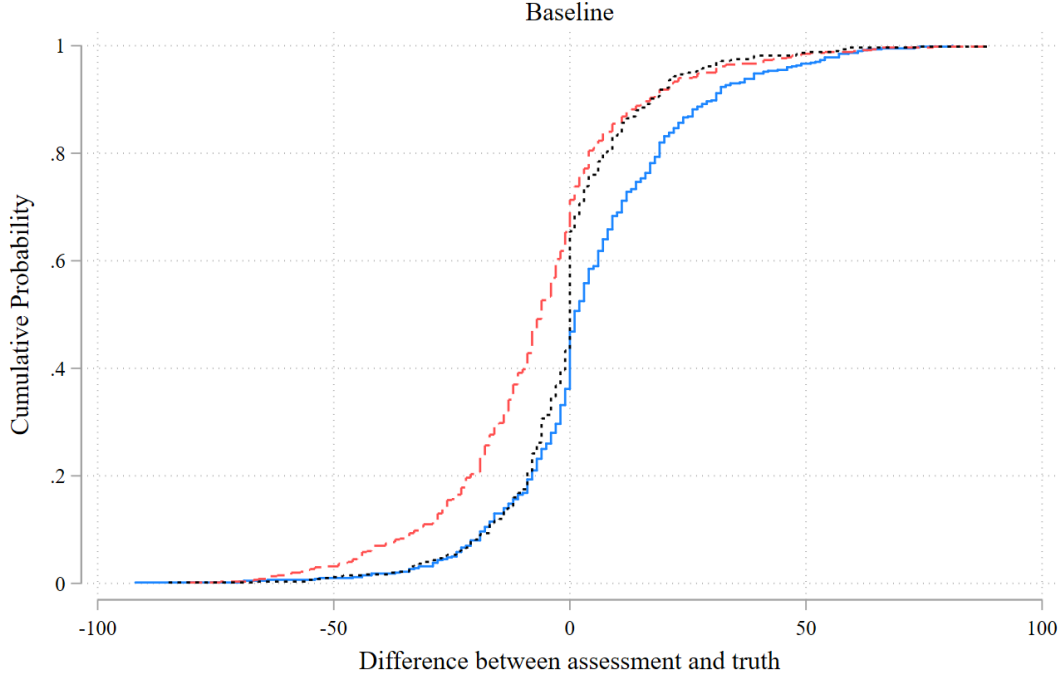
Table 2: The influence of the truth on advisor narratives

	(1) $ \theta_{post}^A - \theta_{post}^{DO} $
$\beta_1$ : Misaligned advisor = 1	13.33*** (0.864)
$\beta_2$ : $ \theta_{post}^T - \theta_{post}^{DO} $	0.974*** (0.0149)
$\beta_3$ : Misaligned advisor $\times  \theta_{post}^T - \theta_{post}^{DO} $	-0.411*** (0.0322)
Round FE	Yes
$H_0 : \beta_2 + \beta_3 > 0$ , p-value	0.001
Observations	3600

(i) The dependent variable is the distance between the advisor's report,  $\theta_{post}^A$ , and the true value  $\theta_{post}^T$ , (ii) Standard errors are clustered at the advisor level, reported in parentheses \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ , (iii) There are 360 clusters, (iv) The sample contains all advisors from the three treatment conditions which share the exact same set of instructions, (v) For each advisor, we have 10 observations—one for each round.

In support of Hypothesis 2b, we find that misaligned advisors respond less strongly to the truth than aligned advisors ( $\beta_3 < 0$ ). However, misaligned advisors do not ignore the truth completely—on average, they do still adjust their narratives towards the truth, even though they are not incentivized to do so ( $\beta_2 + \beta_3 > 0$ ). One potential explanation for this is that (some) advisors hold truth-telling preferences that are sufficiently strong to induce them to tell the truth in some rounds. We find support for this when we calculate the number of rounds in which each advisor lied (see Figure 12 in the appendices for the distribution of this measure). We see that while the vast majority of misaligned advisors lied in more than five rounds, fewer than 40% lied in all ten rounds. This suggests that a majority of advisors hold some truth-telling preferences.

Figure 7: Difference between investor belief,  $\theta_{post}^I$ , and the truth,  $\theta_{post}^T$  (by advisor type)



Notes: (i) The figure plots the cdf of the measure  $\theta_{post}^I - \theta_{post}^T$  for all investor-rounds where the investor is matched with a particular advisor type, (ii) The red dashed line shows the cdf for investor-rounds where the investor is matched with down-advisor, the black dotted lines shows the cdf for investor-rounds where the investor is matched with aligned advisor, and the blue solid line shows the cdf for investor-rounds where the investor is matched with up-advisor, (iii) The figure uses data from investors in the BASELINE.

## 5.2 Persuasion of Investors

A key question that this paper aims to address is whether persuasion using narratives is effective—are advisors successful in distorting investors’ beliefs by proposing biased interpretations of the available objective data? Figure 7 provides an initial visual answer to this question by plotting the cdf of the distance between investors’ beliefs,  $\theta_{post}^I$ , and the truth,  $\theta_{post}^T$ . This is done separately for each advisor type. Specifically, we plot three cdfs—one for all advisor-investor interactions in which an investor is matched with a down-advisor (red, dashed line), one for interactions with a aligned advisor (black, dotted line), and one for interactions with an up-advisor (blue, solid line). The figure shows that the reported beliefs of investors who are matched with an up-advisor stochastically dominate those of investors matched with a down-advisor. This indicates that being matched with an advisor with a conflict of interest does result in a shift in investors’ beliefs towards the self-interest of the advisor.

**Are advisors successful in using narratives to persuade investors?** To explicitly test whether advisors are able to shift the beliefs of the average investor through the narratives they send, and to obtain a measure of the size of this effect, we evaluate the following hypothesis.



**Hypothesis 3.** [PR.1] In *BASELINE*, the distance between the investor’s assessment and the truth is larger when the advisor’s incentives are misaligned than when the advisor’s incentives are aligned.

Table 3 reports the results from the regression testing this hypothesis. We observe that when an investor is matched with an advisor who has a conflict of interest, the investor ends up holding beliefs that are 5pp further from the truth,  $\theta_{post}^T$ . This implies that advisors are successful in distorting the way that investors interpret the objective data through the narratives that they send. This is harmful for investors.

Table 3: Movement of investor beliefs when matched with a misaligned advisor

	(1) $ \theta_{post}^I - \theta_{post}^T $
Misaligned advisor = 1	5.111*** (0.679)
Observations	1800
Round FE	Yes

(i) The dependent variable is the absolute distance between the investor’s belief,  $\theta_{post}^I$ , and the true value  $\theta_{post}^T$ , (ii) Standard errors are clustered at the Interaction Group level, reported in parentheses \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ , (iii) For each of the investors, we have 10 observations—one for each round, (iv) The regression is estimated using data from investors in the *BASELINE* treatment.

**What types of narratives do investors follow?** To examine this question, we relate our measure of narrative quality—the EPI—to investor assessments. We ask whether investors tend to report beliefs,  $\theta_{post}^I$ , that are closer to those transmitted by advisors,  $\theta_{post}^A$ , when the advisor sends a model with a high EPI.

**Hypothesis 4.** [PR.5a] The distance between the advisor’s message and the investor’s assessment decreases in the EPI.

To test this hypothesis, we regress the distance between the advisor’s message,  $\theta_{post}^A$ , and the investor’s report on the EPI of the advisor’s narrative, controlling for round fixed effects and clustering at the Interaction Group level. The results are reported in Table 4. We see that an improvement in the fit of the advisor’s narrative to the objective data from the worst-fitting narrative to the best-fitting narrative is associated with the investor’s belief moving 15pp closer to the advisor’s message,  $\theta_{post}^A$ . This suggests that investors find narratives that fit the data well to be more compelling.

Table 4: Investor conformity and the fit of the advisor’s narrative

	(1) $ \theta_{post}^I - \theta_{post}^A $
Advisor Model Fit (EPI)	-14.59*** (1.892)
Misaligned advisor	0.691 (0.668)
Observations	1800
Round FE	Yes

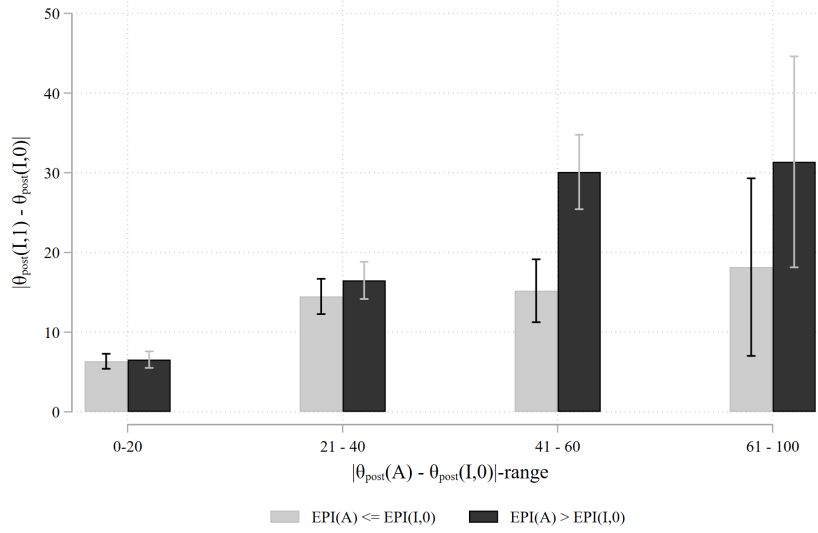
(i) The dependent variables is the absolute distance between the investor’s belief,  $\theta_{post}^I$ , and the advisor narrative,  $\theta_{post}^A$ , (ii) Standard errors are clustered at the Interaction Group level, reported in parentheses \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ , (iii) For each of the investors, we have 10 observations—one for each round, (iv) The regression is estimated using data from investors in the BASELINE treatment.

**Is this “investor conformity-narrative fit” relationship due to belief updating?** One concern that may be raised regarding the relationship documented in Table 4 is that it may not be causal. Taken alone, this result does not imply that the better fit of the advisor’s narrative *causes* the investor to shift their belief towards the advisor’s message. To see this, consider an investor who holds an ex-ante belief about the correct model after seeing the historical data but before receiving the advisor’s message. Now, there are two possible reasons for the negative correlation between  $|\theta_{post}^{I,1} - \theta_{post}^A|$  and  $EPI(m^A)$  that we find in Table 4. The first potential explanation is about updating—when an advisor proposes a better-fitting narrative, the investor shifts their beliefs more, resulting in a smaller gap between the investor’s posterior belief,  $\theta_{post}^{I,1}$ , and advisor’s message,  $\theta_{post}^A$ . The second potential explanation is that this is a spurious relationship that is generated by investors preferring better-fitting prior beliefs. Consider an investor who initially holds a default model that is close to the data-optimal model and who never updates upon receiving the advisor’s model. It would still be possible to observe a negative correlation between the advisor model fit and distance between the advisor’s message and the investor’s assessment: If the advisor sends a model with a high fit, the  $\theta_{post}^A$  will likely be closer to the investor’s default  $\theta_{post}^{I,0}$  than if the advisor sends a model with a low fit, simply because both parameter values will be correlated due to the similarity of their narrative’s EPI. This would lead to the observed negative correlation even though the investor does not update. This example points to a potential endogeneity, which, if the investor’s prior model is likely to fit the data well, can lead to a spurious correlation between  $|\theta_{post}^{I,1} - \theta_{post}^A|$  and  $EPI(m^A)$ .

Our interest lies predominantly in the first channel—detecting a causal effect of the advisor’s narrative fit on the investor’s beliefs. Therefore, we conduct additional analyses to test directly whether the advisor’s EPI affects belief updating. By looking at belief updating, we are controlling for the potential influence of the investor’s prior belief and thereby removing the influence of the second channel. To do this, we use the data collected in our INVESTORPRIOR

treatment where we have information on the investors' prior beliefs.

Figure 8: Belief updating of investors



Notes: (i) The figure uses data from the INVESTORPRIOR treatment, (ii) The y-axis shows the average absolute distance that investors update, (iii) The x-axis disaggregates the data into categories according to the difference between the fit of the advisor's message and the investor's default model.

Figure 8 provides a visual illustration of investor belief updating, depending on whether the empirical fit of the advisor's proposed narrative,  $EPI(m^A)$ , is better or worse than the fit of the investor's default model,  $EPI(m^{I,0})$ . The y-axis shows the average absolute distance that investors update,  $|\theta_{post}^{I,1} - \theta_{post}^{I,0}|$ , and the x-axis disaggregates the data into categories according to the distance between the advisors message and the investor's prior belief,  $|\theta_{post}^A - \theta_{post}^{I,1}|$ . The black bars show updating when the advisor's narrative fits the data better than the investor's prior, while the grey bars show updating when the investor holds a prior that fit the data better than the advisor's proposed narrative. The figure shows that investors update their beliefs more when the advisor proposes a model that fits the data better than their prior. This is particularly the case when the distance between the advisor's proposed  $\theta_{post}^A$  and the investor's prior  $\theta_{post}^{I,0}$  is large. One potential explanation for why investors are less skeptical when updating towards a message where the difference between the message and the assessment is small might be that investors perceive adopting the advisor's model in this case as less risky than in the case where the difference is large.

Table 5: Belief updating and narrative fit

	(1) $ \theta_{post}^{I,1} - \theta_{post}^{I,0} $	(2) $ \theta_{post}^{I,1} - \theta_{post}^{I,0} $	(3) $ \theta_{post}^{I,1} - \theta_{post}^{I,0} $	(4) $ \theta_{post}^{I,1} - \theta_{post}^{I,0} $
$I(EPI^A > EPI^{I,0})$	3.465*** (0.835)	3.350*** (0.852)	-2.203* (1.172)	-1.393 (1.190)
Misaligned sender	0.0117 (1.090)	-0.165 (1.204)	-0.733 (0.747)	-0.681 (0.810)
$ \theta_{post}^{I,0} - \theta_{post}^A $			0.266*** (0.0530)	0.363*** (0.0547)
$I(EPI^A > EPI^{I,0}) \times  \theta_{post}^{I,0} - \theta_{post}^A $			0.238*** (0.0729)	0.173** (0.0717)
Dependent variable mean	11.102	12.35	11.102	12.35
Incl. opposite updaters	Yes	No	Yes	No
Round FE	Yes	Yes	Yes	Yes
Incl. aligned advisors	Yes	Yes	Yes	Yes
Observations	900	779	900	779

(i) The outcome variable in the regressions in this table is the absolute distance that investors update,  $|\theta_{post}^{I,1} - \theta_{post}^{I,0}|$  (ii) The variable  $I(EPI^A > EPI^{I,0})$  is an indicator variable that takes a value of one when the advisor's narrative fits the data better than the investor's prior, (iii) In columns (2) and (4), we remove observations in which the investor updates their belief in the opposite direction to the message sent by the advisor, (vi) Standard errors are clustered at the Interaction Group level, reported in parentheses \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ , (v) For each of the investors, we have 10 observations—one for each round.

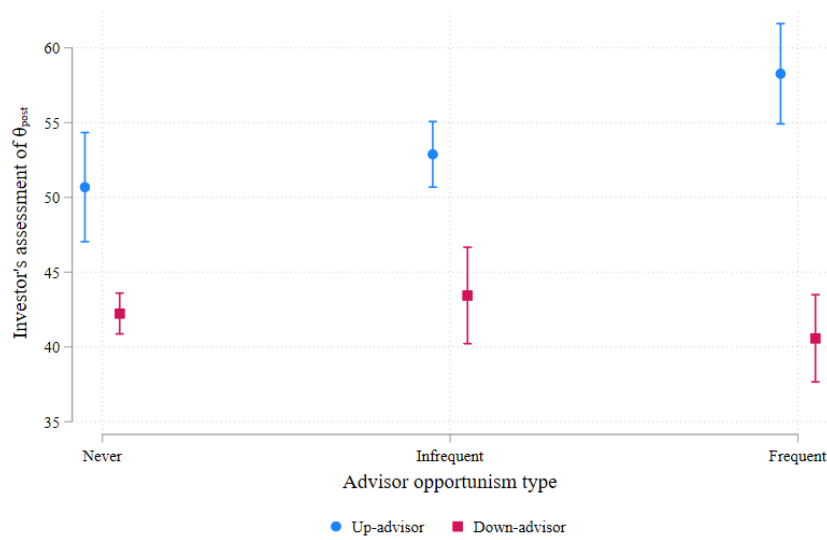
Table 5 presents regression results showing the impact of the fit of the advisor's proposed narrative relative to that of the investor's prior on belief updating. In all four columns, the outcome variable is the absolute amount by which the investor updated their  $\theta_{post}$ -belief. Column (1) shows that investors update their beliefs by approximately 3pp more when the advisor proposes a narrative that fits the data better than their prior belief about the underlying model. In column (3), the coefficient on the interaction term shows that when the advisor's narrative fits better, the investor updates their beliefs by more. Specifically, it shows that as the gap between the advisor's proposed  $\theta_{post}^A$  and the investor's prior,  $\theta_{post}^{I,0}$ , gets larger, an investor who meets an advisor that proposes a better-fitting narrative updates more than an investor who meets an advisor who proposes a worse-fitting narrative. As a robustness exercise, columns (2) and (4) estimate the same specifications as columns (1) and (3) respectively, with the exception that we remove investors who update in the opposite direction to the message received from their advisor.<sup>21</sup> Taken together, these results show that the fit of the advisor's narrative plays an important role in influencing investor belief updating. This provides support for the influence of the first channel discussed above.

**More opportunistic advisors move investor beliefs closer to their persuasion target.** Above we ask which features make individual messages successful and show that the message fit matters. In this section, we ask what types of investors have more persuasive success.

<sup>21</sup>Table 7 in the Appendices report the results from an additional set of robustness exercises. The table examines the influence of the fit of the advisor's narrative when using a more continuous measure of narrative fit. Columns (1) and (2) includes the level of the EPI of the advisor's narrative, while columns (3) and (4) look at the distance between EPI of the advisor's narrative and the investor's prior. Looking at the interaction terms in all four columns shows that the results are consistent with those in Table 5, essentially showing that as the gap between the advisor's message and the investor's prior gets larger, the higher the EPI of the advisor's narrative, the more does the investor update their beliefs. This is consistent with the investor moving towards the advisor's message by more substantially when the advisor's narrative has a high EPI.

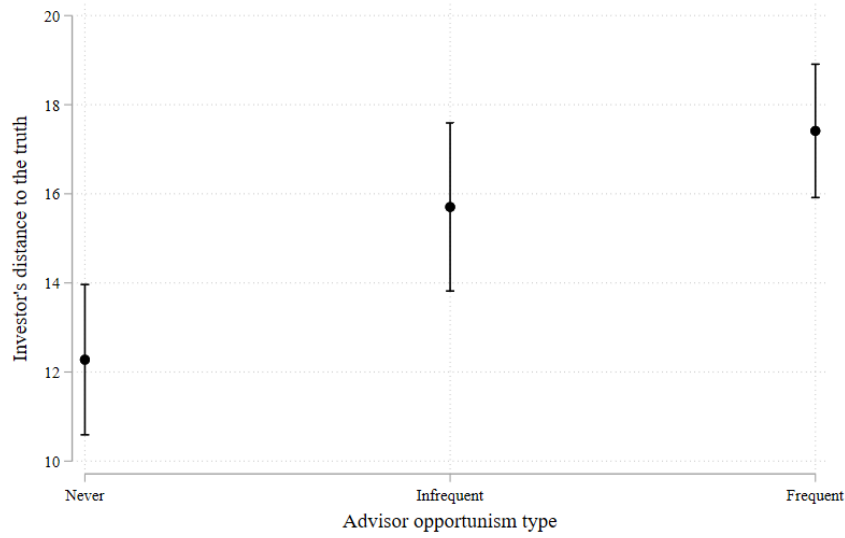
We examine if the opportunism type of the advisor that an investor is matched to—whether they never, infrequently, or frequently choose an advantageous  $c$ —influences the investor’s assessment. We present two results. Figure 9 shows that investors’ assessments are biased more towards extreme values as the advisor they are matched to becomes more opportunistic. This is in line with the “investor conformity-narrative fit” channel where investors update more towards messages that have a higher empirical fit. Since frequent opportunists send messages with comparatively high fits (see Figure 6), their persuasion attempts should successfully shift investor beliefs. Figure 10 presents complementary evidence suggesting that investor assessments become worse, and investor therefore worse off, as their matched advisor becomes more opportunistic.

Figure 9: Investor assessment, by advisor type



Notes: The figure includes observations from investors who are matched to a misaligned advisor in BASELINE. Error bars are 95% confidence intervals derived from regressions which cluster standard errors at the matching group level.

Figure 10: Investor distance from the truth, by advisor type



*Note:* The figure includes observations from investors who are matched to a misaligned advisor in BASELINE. Error bars are 95% confidence intervals derived from regressions which cluster standard errors at the matching group level.

### 5.3 Evaluating Potential Protective Interventions

The discussion above has shown that narrative persuasion can harm investors when they meet an advisor who has a conflict of interest. In this section, we ask whether investors can be protected from this type of persuasion. To do this, we evaluate three potential interventions by comparing investor behavior in each of our three treatment conditions to the behavior observed in BASELINE. Each of these three treatments was designed to capture the core features of a natural option for an intervention. The first asks what happens when advisors' incentives are fully disclosed, the second essentially nudges investors to carefully evaluate the data themselves prior to meeting their advisor, and the third considers a scenario where advisors do not have full knowledge of the data that investor's see.<sup>22</sup>

Table 6 examines whether the interventions help investors to form beliefs that are closer to the truth. The (\*a) columns of the table report the results from regressing the absolute distance between investors' beliefs and the truth on an indicator variable for the particular intervention being considered. The regressions only consider rounds in which investors are matched with advisors with misaligned advisors, since investors do not require protection when they are

<sup>22</sup>There are several ways to think about the PRIVATEDATA treatment. In the context of financial advice, one can think of the investor having access to a subset of the information that the advisor has, but that the advisor does not know which subset this is and, therefore, cannot tailor their message to the investor's information set. However, in other narrative persuasion contexts where the data in question is personal data, the persuader may not have access to the information that the receiver has at all. For example, for medical advice, tailored marketing, or political persuasion, the persuader may wish to tailor their narrative to the individual. This can be done if the persuader has access to a wealth of personal information about their target (e.g., data collected from an individual's browsing history). For such scenarios, the PRIVATEDATA treatment has a different interpretation. It considers the effectiveness of policy interventions that assign ownership of personal information to the individual.

matched with an advisor with perfectly aligned incentives. The coefficient associated with “Intervention=1” in each of the (\*a) columns shows the average effect of the intervention denoted in the column header.<sup>23</sup>

Surprisingly, we see that none of the three interventions had a statistically significant protective effect for the average investor. For the INVESTORPRIOR and PRIVATE DATA treatments, this can also be seen visually in Figure 15, which shows that the distribution of the distance between investors’ beliefs and the truth in these treatments is very similar to that in BASELINE. However, the figure also shows that investor behavior changes substantially in DISCLOSURE, where there is far less of a gap between the beliefs of investors who are matched with up- and down-advisors. This difference in behavior is not surprising because one would expect that investors who have their advisor’s conflict of interest disclosed to them will become more skeptical and be less influenced by the narrative received from these conflicted advisors.

Table 6: Evaluating the impact of interventions aimed at protecting investors

	DISCLOSURE $ \theta_{post}^I - \theta_{post}^T $		INVESTORPRIOR $ \theta_{post}^I - \theta_{post}^T $		PRIVATE DATA $ \theta_{post}^I - \theta_{post}^T $	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Intervention = 1	-0.713 (1.001)	2.403 (1.549)	0.454 (0.924)	1.241 (1.117)	-0.124 (0.750)	-0.0775 (1.192)
Advisor lied=1		9.340*** (1.012)		9.200*** (1.024)		9.419*** (1.018)
Intervention × Advisor lied		-3.974** (1.633)		-0.764 (1.425)		0.116 (1.558)
Control BASELINE mean	15.274	15.274	15.274	15.274	15.274	15.274
Round FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1800	1800	1800	1800	1800	1800

(i) The dependent variable is the distance between the true  $\theta_{post}^T$  parameter and the corresponding belief held by the investor  $\theta_{post}^I$ , (ii) Standard errors are clustered at the Interaction Group level, reported in parentheses \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ , (iii) The regressions are estimated using data from investors who are matched with misaligned advisors (i.e., rounds in which investors are matched with aligned advisors are excluded), (iv) Each column uses data from the BASELINE treatment as well as the relevant treatment mentioned in the column header, (v) The results in columns (\*a) relate to Hypotheses 2, 3 and 4 from the pre-registration.

**Why does increased skepticism not protect investors in DISCLOSURE?** The discussion above leads to the question of why this increased skepticism in DISCLOSURE does not protect the average investor. One potential explanation is the following. We know from Figure 12 that

<sup>23</sup>These regressions correspond to Hypotheses 2, 3, and 4 of the pre-registration:

**Hypothesis 5.** [PR.2] When matched with an advisor with misaligned incentives, the distance between the investor’s assessment and the truth is smaller in DISCLOSURE than in BASELINE.

**Hypothesis 6.** [PR.3] When matched with a advisor with misaligned incentives, the distance between the investor’s assessment and the truth is smaller in INVESTORPRIOR than in BASELINE.

**Hypothesis 7.** [PR.4] When matched with a advisor with misaligned incentives, the distance between the investor’s assessment and the truth is smaller in PRIVATE DATA than in BASELINE.



even advisors with a conflict of interest tell the truth a substantial fraction of the time (some are honest most of the time). Investors who become more skeptical about the narrative that they receive from an advisor with a conflict of interest cannot easily distinguish advisors who are being honest from those that are being dishonest. This implies that they also become skeptical of narratives received from advisors who are being honest. This would lead skeptical investors to do worse than less-skeptical investors when matched with honest advisors, but better than less-skeptical investors when matched with dishonest advisors. Column (1b) provides some support for this explanation by showing that investors are indeed protected when they are matched with an advisor who is lying to them (negative coefficient on the interaction term). In contrast, the coefficient on the “Intervention=1” variable is positive, suggesting that they are harmed when matched with an honest advisor (although, this variable is not statistically significant). Therefore, we find that it is difficult to protect investors from this type of persuasion using narratives. Even when investors are explicitly told that they face an advisor with a conflict of interest, they are not better off than when they were uncertain about the advisor’s interests.

**When do investors follow advisors’ messages?** To better understand exactly how the treatments are influencing investor behavior, we replicate Table 6 but now consider as an outcome variable the distance between the investors’ beliefs and the advisors’ message. The results are reported in Table 8 in the Appendices. This table provides insight into what determines whether investors follow the message of their advisor. The table reveals several interesting insights. First, the coefficient point estimates in the (\*a) columns show that in all three treatments, investors beliefs are 2-3pp further from the message that they receive from their advisor relative to in the BASELINE treatment (although, this coefficient not statistically significant for INVESTORPRIOR). This indicates that the intervention treatments are leading investors to rely less on their advisors’ messages. However, in combination with the results discussed above, it seems that investors are not able to make their beliefs more accurate. Second, the coefficient on the “Advisor lied” variable is highly significant and shows that investors in BASELINE report beliefs that are 4pp further from their advisors message when the advisor lies. This provides strong evidence that investors are able to detect advisor lying to some degree. Third, the interaction term in column (6) shows that investors are even less likely to follow the messages of advisors who lie in the PRIVATEDATA treatment. Here, investors’ beliefs are over 7pp further away from advisors’ messages when the advisor lies. This makes sense, since advisors in PRIVATEDATA cannot tailor their lies to the data that the investor observes.

Turning to the aligned advisors, Table 9 in the Appendices reports the effect of the treatment interventions for investors matched with aligned advisors. The (\*a) columns report the results for the distance between the investor’s belief and the truth, while the (\*b) columns consider the distance between the investor’s belief and the advisor’s message. The results are largely as one would expect. We observe a large influence of the DISCLOSURE treatment. Specifically, when investors learn that they are matched with an aligned advisor, this shifts their beliefs 5pp closer

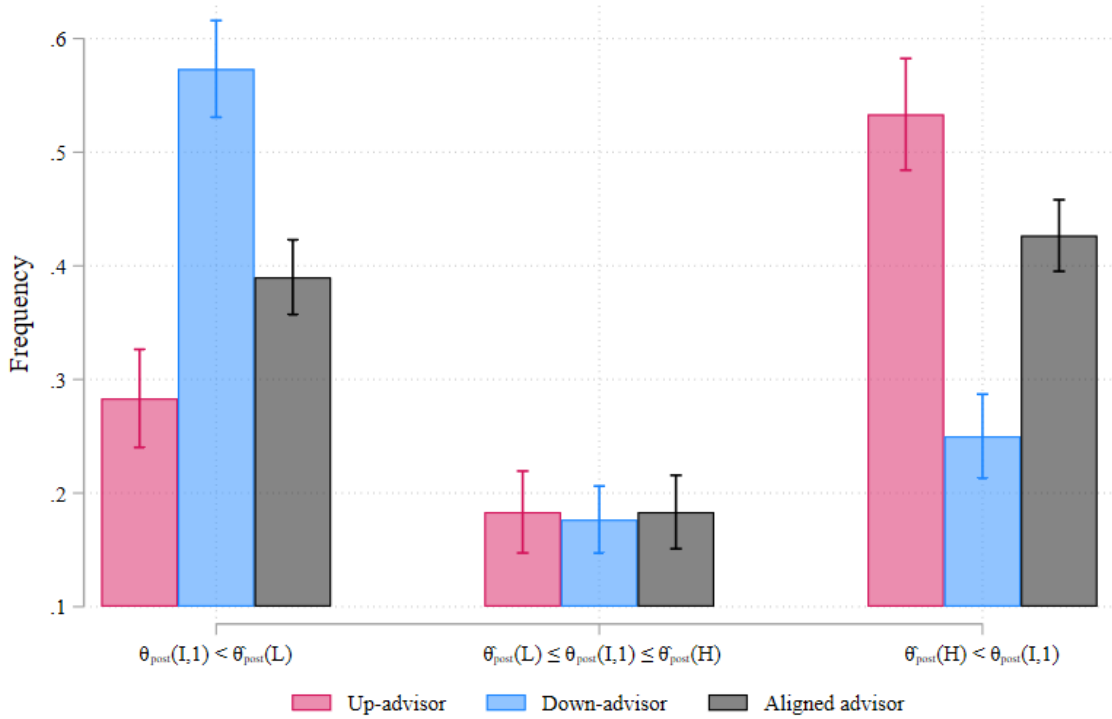
to their advisor’s message and also 5pp closer to the truth. This halves the average distance from the truth for investors matched with aligned advisors in BASELINE. The other treatment interventions have no impact on the beliefs of investors matched with aligned advisors.

## 5.4 Consistency of the experimental data with Nash equilibrium

We can also test for the existence of patterns in the data that are predicted by the persuasive Nash equilibrium of the underlying cheap talk game. As we discuss in Section 3.5 and Appendix E, in the most informative equilibrium of the game underlying the BASELINE treatment, some persuasive communication around the investor’s prior expectation of  $\theta_{post}$  is possible; the Nash equilibrium predicts the existence of a lower threshold,  $\theta_{post}^L$ , and an upper threshold,  $\theta_{post}^H$ , with  $\theta_{post}^L < \mathbb{E}[\theta_{post}] < \theta_{post}^H$ . The investor adopts a message only if it includes a  $\theta_{post}$  parameter on the interval between both thresholds.

We can numerically solve for these thresholds, which are uniquely determined for every historical data set observed by participants in the experiment. With these thresholds, we can ask how often investors adopt messages that are outside the threshold interval—this should never happen in equilibrium. Figure 11 shows how often investors in BASELINE report an assessment that is either (i) lower than  $\theta_{post}^L$ , (ii) between  $\theta_{post}^L$  and  $\theta_{post}^H$ , or (iii) larger than  $\theta_{post}^H$ , conditional on advisor type. We observe that the majority of investors make assessments which are outside of the range predicted by Nash equilibrium. Moreover, advisor messages are more persuasive than predicted: Being matched to a down-advisor significantly increases the proportion of assessments lower than  $\theta_{post}^L$  while being matched to an up-advisor significantly increases the proportion of assessments above  $\theta_{post}^H$ . The advisor effects are large: under both up- and down-advisors, the majority of investors makes an assessment that is below (down-advisor) or above (up-advisor) the interval range. Taken together, these results indicate that behavior is not very consistent with sophisticated strategic thinking.

Figure 11: Frequency of investor assessments that fall inside a certain interval, by advisor type



Notes: Error bars are 95% confidence intervals that were derived from standard errors which are clustered at the matching group level.

## 6 Concluding Discussion

We provide empirical evidence showing how narratives can be used as a tool for persuasion, with one individual shaping how another interprets objective data. Since the advisor can construct the narrative *ex post*, she is able to tailor it to the public data. This *ex post* tailoring means that the advisor is able to construct a narrative that fits the data well and in turn can present this coherence with the objective information as supporting evidence for the veracity of the narrative. In line with this idea, we document systematic patterns in the strategies used by advisors to construct the narratives they send. Advisors in the experiment manage to construct messages which influence the assessments of investors, and this is especially the case for messages that fit the historical data well. As a result, misaligned advisors manage to successfully bias investors' interpretation of the data in ways that benefit the advisor.

The results from the interventions show that narrative persuasion is difficult to protect against, with none of the interventions we consider bringing investors' assessments closer to the truth on average. This finding for the DISCLOSURE-BASELINE comparison is reminiscent of the results discussed by Cain et al. (2005) and Sah et al. (2013), who examine disclosure of conflicts of interest in settings closer to a standard sender-receiver game (i.e., without narratives). Cain et al. (2005) find that disclosing incentives can backfire because it changes the behavior of both senders and receivers in a particular way. In their experiment, senders dis-

tort the messages they send even further from the truth when their incentives are disclosed. Receivers do not sufficiently discount this increased bias in the messages, implying that the net effect of disclosure harms investors. The ineffectiveness of disclosure that we find in our experiment is a consequence of a different mechanism. First, in our setting we focus only on the investor side, since advisors in our BASELINE and DISCLOSURE treatments receive identical instructions. Therefore, we abstract away from any backfiring mechanism that operates via the advisor. We also do not allow advisors to choose whether to disclose their own incentives. Such voluntary disclosure could trigger a (perceived) credibility boost and is explored in Sah et al. (2013). Therefore, we rule out several mechanisms considered in previous work and document a new channel through which disclosure may backfire—namely, that investors who become more skeptical of misaligned advisors’ messages are insufficiently able to distinguish misaligned advisors who still offer honest advice from misaligned advisors who offer dishonest advice. This means that investors may benefit from the introduction of disclosure when they meet a dishonest advisor, but can be harmed by disclosure when in fact the advisor they meet is honest despite having misaligned incentives. In this, our results highlight a new pathway through which disclosure may backfire.

Our PRIVATE DATA intervention ensures that the advisor cannot tailor the message to the realized data by restricting access to the data to only the investor. While this intervention does not draw the investor’s beliefs closer to the truth, it does change investor behavior. In particular, investors in PRIVATE DATA form beliefs that are significantly further away from the advisor’s message than in BASELINE. This increase in the distance from the advisors message is concentrated among messages where the advisor is lying. Taken together, these results suggest that it does in fact become more difficult for advisors to convince investors of their lies once they are not able to tailor them to the data. The results are in line with the ideas discussed in the theoretical framework of S&S. One potential explanation for why this intervention does not help investors move their beliefs closer to the truth despite reducing the persuasiveness of advisor lies is that investors seem to not be sufficiently good at inferring the true underlying model on their own.

While our experiment is designed to study persuasion through the lens of S&S, an advantage of our design is that we can test whether our results are consistent with the Nash equilibrium predictions of the underlying cheap talk game. We find that investors make assessments which are outside the interval of values predicted by the most persuasive Nash equilibrium. One of the reasons for this could be that investors draw heterogeneous inferences from the evidence provided in the form of the historical data and this can result in them sometimes holding prior expectations which are outside the interval. However, our evidence also suggests that, through their messages, advisors succeed in installing beliefs in investors which are outside the interval. This suggests that communication in our experiment is more persuasive than predicted by Nash, a result which has also been documented in different experimental settings the literature (Cai and Wang, 2006). One interesting question is whether the relative complexity of our decision environment encourages investors adopt the credulous-but-skeptic decision

rule described by S&S.<sup>24</sup> If this is the case, then we would expect S&S's narrative persuasion theoretical framework to provide the more appropriate tool for analyzing situations with complex data sources. This is the case in many important life decisions, such as when buying a house or investing during the trough of a major recession. In such scenarios, being exposed to a proposed narrative may shape how the individual processes the complex and potentially overwhelming wealth of information they have access to.

Our analysis provides an early empirical contribution to understanding some of the mechanisms that govern narrative persuasion. The data we collect is very rich and contains several layers of exogenous variation. This allows us to document interesting systematic regularities in the way that advisors construct their narratives and also to learn about when and how investors' beliefs are influenced by receiving such narratives. However, given the breadth and importance of the topic, there is a need for further research to paint a more complete picture of how narrative persuasion is influenced by other contextual factors. As is often the case when exploring new research areas, our analysis has raised many new questions. These questions could provide promising avenues for further research. The following provides an outline of some of these avenues.

**Narrative construction as a personal skill.** Our results indicate that narrative persuasion can be highly effective. Even though participants in our experiment are likely to be relatively inexperienced in constructing convincing narratives, they are able to employ fairly sophisticated strategies to manipulate others' beliefs. In everyday applications, expert persuaders might not only be successful in their role as advisor because of the expert knowledge they possess but also because they are able to skillfully relate the narratives they construct to a selected subset of the huge quantity of available objective data. Some individuals might also be particularly creative in constructing new types of narratives. Therefore, selection pressures on the narrative construction skill-domain might make the effects of narrative persuasion even more pernicious in real-world contexts. Furthermore, the relevance of narrative persuasion extends far beyond the domain of financial advice. Examples where persuasion using narratives may play an important role in everyday life abound, from political persuasion by politicians and lobbyists on virtually every policy issue to lawyers who weave a story through the evidence to persuade a jury of their case to businesses who carefully sculpt a marketing story to persuade consumers to buy their product. It would be interesting to investigate whether narrative-construction skills are particularly developed amongst individuals in these professions (either due to selection of individuals with that trait into the profession or due to learning the skill within the profession).

**Narrative persuasion in less constrained real world scenarios.** In many real-world settings, the available data that individuals might draw upon to learn about a particular company or fund is normally larger and more complex than in our experiment. Furthermore, the set of

---

<sup>24</sup>In our setting forming a Bayesian prior based on the data is non-trivial, which implies that assuming common priors is a fairly strong assumption to make.

possible narratives available to advisors is also typically unconstrained. Advisors may select which variables to include in the narrative they propose, as well as the relationship between these variables, with more flexibility.<sup>25</sup> In addition, advisors often have a degree of flexibility to choose what data they want to reveal (or highlight) to their advisees, thereby hiding or obscuring information that does not support their favored narrative. These relaxations of the environment might increase the effectiveness of persuasion using narratives. Our experiment is not designed to answer this question and research in this direction would be valuable.

**How would behavior change if advisors did not know the true underlying model?** Advisors in our experiment were provided with the true underlying model. This design choice establishes the advisor as an expert with superior knowledge and provides us with the control that we use to measure discrepancies between the advisor's own beliefs about the true underlying model and the messages she sends. It also ensures that advisor's beliefs are exogenously assigned, removing one layer of potential endogeneity from the analysis. However, it is informative to think about how our setting relates to different real world contexts in which advisors may or may not be aware of the true underlying model. In some contexts, an advisor may privately be aware of information about the underlying process, but also be aware that the advisee is not. In such cases, the advisor faces a choice that is very similar to that in our experiment. She has hard information about the truth, but knows that the advisee does not. Here, the normative prescription is clear—the morally desirable choice would be to reveal what she knows. In other real-world contexts, the advisor may not have concrete information about the process underlying the observable data. The advisor must then draw inference from the data based on her expertise. Now, there are two channels through which such an advisor might construct a biased narrative based on having incentives that are not aligned with the investor. Either the advisor deceives herself and then transmits her truly held belief to the investor. Such an advisor draws biased inference from the data by actually believing in a narrative that is distorted due to her private incentives. Or, alternatively, the advisor forms an unbiased assessment of the data and believes in one narrative but chooses to transmit a different narrative to the investor. Our experiment rules out the first channel (self-deception) and focuses on the second.

Taken as a whole, our results indicate that narratives can provide an effective tool for persuasion. Bad actors may exploit this opportunity in a wide range of economically important settings, with the proliferation of social media and rapid expansion of access to data potentially exacerbating the problem. Given these concerns, and the fact that it is non-trivial to protect individuals from this form of persuasion, further work that helps to develop a deeper understanding of the psychological mechanisms involved and that identifies the most effective protection strategies would be valuable.

---

<sup>25</sup>See Andre et al. (2022) for a neat example of how individuals might construct narratives in more complex settings by selecting a subset of the available variables and constructing causal links between them.

## References

- Aina, C. (2021). Tailored Stories. *Mimeo*.
- Akerlof, G. A. and D. J. Snower (2016). Bread and Bullets. *Journal of Economic Behavior & Organization* 126, 58–71.
- Allcott, H., L. Boxell, J. Conway, M. Gentzkow, M. Thaler, and D. Yang (2020). Polarization and Public Health: Partisan Differences in Social Distancing During the Coronavirus Pandemic. *Journal of Public Economics* 191, 104254.
- Andre, P., I. Haaland, C. Roth, and J. Wohlfart (2022). Narratives about the macroeconomy. *Working Paper*.
- Andre, P., C. Pizzinelli, C. Roth, and J. Wohlfart (2022). Subjective Models of the Macroeconomy: Evidence from Experts and a Representative Sample. *Review of Economic Studies*.
- Barron, K., H. Harmgart, S. Huck, S. Schneider, and M. Sutter (2022). Discrimination, Narratives and Family History: An Experiment with Jordanian Host and Syrian Refugee Children. *Review of Economics and Statistics*.
- Barron, K., S. Huck, and P. Jehiel (2019). Everyday econometricians: Selection neglect and overoptimism when learning from others. *WZB Discussion Paper*.
- Bénabou, R., A. Falk, and J. Tirole (2020). Narratives, Imperatives, and Moral Persuasion. *Working Paper*.
- Blume, A., D. V. DeJong, Y.-G. Kim, and G. B. Sprinkle (1998). Experimental evidence on the evolution of meaning of messages in sender-receiver games. *The American Economic Review* 88(5), 1323–1340.
- Blume, A., D. V. DeJong, G. R. Neumann, and N. Savin (2002). Learning and communication in sender-receiver games: an econometric investigation. *Journal of Applied Econometrics* 17(3), 225–247.
- Braghieri, L., R. Levy, and A. Makarin (2022). Social media and mental health.
- Bruner, J. (1991). The Narrative Construction of Reality. *Critical Inquiry* 18(1), 1–21.
- Cai, H. and J. T.-Y. Wang (2006, July). Overcommunication in strategic information transmission games. *Games and Economic Behavior* 56(1), 7–36.
- Cain, D. M., G. Loewenstein, and D. A. Moore (2005). The dirt on coming clean: Perverse effects of disclosing conflicts of interest. *The Journal of Legal Studies* 34(1), 1–25.
- Charles, C. and C. Kendall (2022). Causal narratives. *Mimeo*.



- Charnysh, V. (2021). Remembering past atrocities—good or bad for attitudes toward minorities? *Mimeo*.
- Chater, N. and G. Loewenstein (2016). The under-appreciated drive for sense-making. *Journal of Economic Behavior & Organization* 126, 137–154.
- Converse, P. E. (2006). The Nature of Belief Systems in Mass Publics. *Critical review* 18(1-3), 1–74.
- Crawford, V. P. and J. Sobel (1982). Strategic Information Transmission. *Econometrica* 50(6), 1431–1451.
- Eliaz, K. and R. Spiegler (2020). A Model of Competing Narratives. *American Economic Review* 110(12), 3786–3816.
- Eliaz, K., R. Spiegler, and Y. Weiss (2021). Cheating with Models. *Working Paper*.
- Enke, B. (2020). What You See is All There Is. *Quarterly Journal of Economics* 135(3), 1363–1398.
- Enke, B. and F. Zimmermann (2019). Correlation neglect in belief formation. *The Review of Economic Studies* 86(1), 313–332.
- Eyster, E. and M. Rabin (2005). Cursed Equilibrium. *Econometrica* 73(5), 1623–1672.
- Eyster, E. and G. Weizsacker (2016). Correlation neglect in portfolio choice: Lab evidence. *Available at SSRN* 2914526.
- Foucault, M. (1972). *The Archaeology of Knowledge*. New York: Pantheon Books.
- Franzosi, R. (1998). Narrative Analysis-or Why (and How) Sociologists Should be Interested in Narrative. *Annual Review of Sociology* 24(1), 517–554.
- Gennaioli, N. and A. Shleifer (2018). *A Crisis of Beliefs*. Princeton University Press.
- Hagenbach, J. and E. Perez-Richet (2018). Communication with evidence in the lab. *Games and Economic Behavior* 112, 139–165.
- Hagmann, D., C. Minson, C. Tinsley, et al. (2021). Personal narratives build trust across ideological divides. *Working Paper*.
- Haidt, J. (2007). The New Synthesis in Moral Psychology. *Science* 316(5827), 998–1002.
- Haidt, J. (2012). *The Righteous Mind: Why Good People are Divided by Politics and Religion*. Vintage.
- Harbaugh, R. and E. Rasmusen (2018). Coarse Grades: Informing the Public by Withholding Information. *American Economic Journal: Microeconomics* 10(1), 210–235.

- Harris, S., L. M. Müller, and B. Rockenbach (2021). How optimistic and pessimistic narratives about covid-19 impact economic behavior.
- Heidhues, P., B. Köszegi, and P. Strack (2018). Unrealistic Expectations and Misguided Learning. *Econometrica* 86(4), 1159–1214.
- Herman, L. and B. Vervaeck (2019). *Handbook of Narrative Analysis*. University of Nebraska Press.
- Hillenbrand, A. and E. Verrina (2022). The Differential Effect of Narratives on Prosocial Behavior. *Games and Economic Behavior* 135, 241–270.
- Hossain, T. and R. Okui (2013). The Binarized Scoring Rule. *Review of Economic Studies* 80(3), 984–1001.
- Ispano, A. (2022). The perils of a coherent narrative. *THEMA Working Paper N. 2022-13*.
- Jehiel, P. (2018). Investment strategy and selection bias: An equilibrium perspective on overoptimism. *American Economic Review* 108(6), 1582–97.
- Jin, G. Z., M. Luca, and D. Martin (2021). Is no news (perceived as) bad news? an experimental investigation of information disclosure. *American Economic Journal: Microeconomics* 13(2), 141–73.
- Karlsson, N., G. Loewenstein, J. McCafferty, et al. (2004). The economics of meaning. *Nordic Journal of Political Economy* 30(1), 61–75.
- King, R. R. and D. E. Wallin (1991). Market-induced information disclosures: An experimental markets investigation. *Contemporary Accounting Research* 8(1), 170–197.
- Koschorke, A. (2018). *Fact and Fiction: Elements of a General Theory of Narrative*. Walter de Gruyter.
- Laudenbach, C., M. Ungeheuer, and M. Weber (2019). How to alleviate correlation neglect. *CEPR Discussion Paper No. DP13737*.
- Laudenbach, C., A. Weber, and J. Wohlfart (2021). Beliefs about the stock market and investment choices: Evidence from a field experiment. *CEBI Working Paper 17/21*.
- Little, A. T. (2022). Bayesian explanations for persuasion. *OSF Preprints*.
- Loewenstein, G., S. Sah, and D. M. Cain (2012). The unintended consequences of conflict of interest disclosure. *Jama* 307(7), 669–670.
- Mailath, G. J. and L. Samuelson (2020). Learning under Diverse World Views: Model-Based Inference. *American Economic Review* 110(5), 1464–1501.

- Malmendier, U. and D. Shanthikumar (2007). Are small investors naive about incentives? *Journal of Financial Economics* 85(2), 457–489.
- Mannheim, K. (2015). *Ideology and Utopia*. USA: Martino Publishing.
- Moore, D. A. and P. J. Healy (2008). The trouble with overconfidence. *Psychological review* 115(2), 502.
- Morag, D. and G. Loewenstein (2021). Narratives and valuations. Available at SSRN 3919471.
- Olea, J. L. M., P. Ortoleva, M. M. Pai, and A. Prat (2021). Competing Models. *Working Paper*.
- Pennington, N. and R. Hastie (1986). Evidence Evaluation in Complex Decision Making. *Journal of Personality and Social Psychology* 51(2), 242.
- Pennington, N. and R. Hastie (1988). Explanation-Based Decision Making: Effects of Memory Structure on Judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14(3), 521.
- Pennington, N. and R. Hastie (1992). Explaining the Evidence: Tests of the Story Model for Juror Decision Making. *Journal of Personality and Social Psychology* 62(2), 189.
- Polletta, F., P. C. B. Chen, B. G. Gardner, and A. Motes (2011). The Sociology of Storytelling. *Annual Review of Sociology* 37, 109–130.
- Roos, M. and M. Reccius (2021). Narratives in economics. *arXiv preprint arXiv:2109.02331*.
- Sah, S. and G. Loewenstein (2014). Nothing to declare: Mandatory and voluntary disclosure leads advisors to avoid conflicts of interest. *Psychological science* 25(2), 575–584.
- Sah, S., G. Loewenstein, and D. M. Cain (2013). The burden of disclosure: increased compliance with distrusted advice. *Journal of personality and social psychology* 104(2), 289.
- Schumacher, H. and H. C. Thysen (2022). Equilibrium contracts and boundedly rational expectations. *Theoretical Economics* 17(1), 371–414.
- Schwartzstein, J. and A. Sunderam (2021). Using Models to Persuade. *American Economic Review* 111(1), 276–323.
- Shiller, R. J. (2017). Narrative Economics. *American Economic Review* 107(4), 967–1004.
- Shiller, R. J. (2019). *Narrative economics*. Princeton University Press Princeton.
- Shiller, R. J. (2020). Popular economic narratives advancing the longest us expansion 2009–2019. *Journal of policy modeling* 42(4), 791–798.
- Sloman, S. A. and D. Lagnado (2015). Causality in Thought. *Annual Review of Psychology* 66(1), 223–247.

- Spiegler, R. (2016). Bayesian Networks and Boundedly Rational Expectations. *Quarterly Journal of Economics* 131(3), 1243–1290.
- Spiegler, R. (2020a). Behavioral Implications of Causal Misperceptions. *Annual Review of Economics* 12(1), 81–106.
- Spiegler, R. (2020b). Can agents with causal misperceptions be systematically fooled? *Journal of the European Economic Association* 18(2), 583–617.
- Wang, J. T.-y., M. Spezio, and C. F. Camerer (2010). Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *American economic review* 100(3), 984–1007.

# APPENDICES

## A Conceptualization of Narratives in Economics

Recently, the economics profession has seen a rapid rise in the interest in incorporating the concept of a “narrative” into economic models. This was highlighted by Robert Shiller’s 2017 Presidential Address at the American Economic Association and the associated publication of his book “Narrative Economics” (Shiller, 2019). Further, in their paper reviewing the existing literature on narratives in economics, Roos and Reccius (2021) show that the number of economics publications containing the term “narrative” in the title [abstract] has been growing sharply for at least the past ten years. Nevertheless, as noted by Roos and Reccius (2021), amongst others, there does not yet exist a commonly accepted definition of what the term “narrative” means in economics. The following examples selected from important early contributions to this literature serve to illustrate this point.

Morag and Loewenstein (2021) view a narrative as “a story [that] places selected events on a timeline and establishes causal links between them” (p. 2).<sup>26</sup> Shiller (2020) argues that economic narratives are “stories that offer interpretations of economic events, or morals, [or] hints of theories about the economy [that] go viral just as diseases do” (p. 792). Similar to Shiller (2020), Bénabou et al. (2020) offer a fairly broad conceptualization of a moral narrative as “... any signal, story, or heuristic that can potentially alter an agent’s beliefs about the tradeoff between private benefits and social costs” (p. 1). Spiegler (2020a) proposes a different approach, placing substantially more structure on what constitutes a narrative. In a series of contributions to this literature, Spiegler draws on Bayesian Network theory to analyze the implications of representing narratives as Directed Acyclic Graphs (DAGs), demonstrating the value and potential of this approach (see, e.g., Spiegler, 2016, 2020a,b; Eliaz and Spiegler, 2020; Eliaz, Spiegler, and Weiss, 2021). In this framework, DAGs indicate (subjective) causal relationships between variables that are relevant in determining a particular outcome of interest. A particular (subjective) DAG can then be thought of as a lens through which an individual interprets the data they observe. Together, the DAG and the data determine the beliefs that the individual holds, and consequently their actions.<sup>27</sup> Similar to Bayesian Networks approach, Schwartzstein and Sunderam (2021) consider a setting in which individuals may interpret an

---

<sup>26</sup>This view is reminiscent of that proposed by Akerlof and Snower (2016), who characterizes narratives as “a sequence of causally linked events and their underlying sources” (p. 58). This conceptualization adopted by Morag and Loewenstein (2021) and Akerlof and Snower (2016) draws on the perspective commonly adopted in psychology (see, e.g., Bruner, 1991; Pennington and Hastie, 1992; Sloman and Lagnado, 2015). Roos and Reccius (2021) offer a refined version of this definition of the concept by placing some additional restrictions on what constitutes a narrative in their view. They propose defining a “collective economic narrative” as “a sense-making story about some economically relevant topic that is shared by members of a group, emerges and proliferates in social interaction, and suggests actions” (p. 13).

<sup>27</sup>In two early empirical contributions to this literature, Andre et al. (2022) and Charles and Kendall (2022) build on this idea by studying how individuals form subjective causal models (represented as DAGs) that individuals hold—Andre et al. (2022) examine subjective causal models of the causes on inflation, while Charles and Kendall (2022) provide a test of the theory in a more controlled environment.

existing data set in different ways. However, while the line of research by Spiegler and coauthors focuses predominantly on subjectivity in the construction of the causal structures (DAGs) that are used to interpret the data, Schwartzstein and Sunderam (2021) instead formalize subjective models as likelihood functions and focus on how individuals select amongst models given the data. One key difference is that the Bayesian Networks approach of, e.g. Eliaz and Spiegler (2020), considers inference given infinitely large data where the true model will maximize the likelihood. In contrast, Schwartzstein and Sunderam (2021) study inference with finite data where the true model may sometimes appear less compelling in the data than an incorrect model. A second key difference is that instead of focusing on mistakes in subjective beliefs about causal relations, the Schwartzstein and Sunderam (2021) framework allows for scenarios where there is general agreement about the causal relations between variables, but mistakes may arise in beliefs about other features of the underlying model.<sup>28</sup> Therefore, while both approaches study mistakes that may arise in the interpretation of data, they shine a spotlight on different types of mistakes. In this, they provide complementary perspectives that illustrate different aspects of how biased narrative construction may occur and how they can influence behavior.

Given the fluidity of the concept of narratives in everyday usage and the relative infancy of its use in economics, it is unsurprising that it has been formalized in various different ways. As the discipline collectively explores the usefulness of the analytical concept and tests different approaches to incorporating it into the existing theoretical framework, it seems like a natural and healthy process to experiment with different formalizations. Furthermore, given the vast array of scenarios to which the concept is commonly applied, it seems likely that even when the concept distillation process has reached maturity, several formalizations may survive and prove useful in parallel.<sup>29</sup> The aim of the discussion below is therefore not to propose one specific definition of a narrative that we view as preferable; rather, our aim is threefold: (i) to discuss the features that are common to the different conceptualizations of the term “narrative” in economics, (ii) to briefly discuss some of the constituent components of narratives that are important to think clearly about when comparing different conceptualizations, and (iii) to be clear and precise about what we mean when we refer to a narrative in this paper.

## A.1 What different conceptualizations of the term “narrative” share

There seems to be one core shared feature that is present across most of the working definitions of the term narrative used in the examples discussed above, namely that a narrative involves

---

<sup>28</sup>For example, in the current paper, there is common knowledge that the CEO determines the success of the company—i.e., one could think of a very simple DAG that states:  $\text{CEO} \rightarrow \text{company success}$ . However, mistakes may arise in the strength of this causal relationship and also with regards to when the strength of this causal relationship changed (i.e., when the CEO changed).

<sup>29</sup>An analogous example of this is provided by the literature on *overconfidence*, where it has proved useful to develop a distinct terminology for three different forms of overconfidence, namely overoptimism (or overestimation), overplacement, and overprecision (Moore and Healy, 2008). Each refers to a distinct and precise version of the concept of overconfidence, which previously were often used loosely and often conflated.

“sense-making”.<sup>30</sup> Specifically, in the examples, the concept is used to refer to providing an explanation for a collection of events.<sup>31</sup> This collection of events can take many forms. Consider the following illustrative examples. One can think of a sequence of historical events, such as those leading up to World War II or those leading up to the 2007 Financial Crisis (see, e.g., Gennaioli and Shleifer, 2018). Here, a narrative explaining the causes of World War II or the Financial Crisis would weave a causal path through the preceding events. One can think of the rise of depression amongst teenagers along with the other contemporaneous changes in society in the last twenty years, such as the rise of social media usage (see, e.g., Braghieri, Levy, and Makarin, 2022). Here, a narrative might posit that the widespread diffusion of social media is causally responsible for the rise in depression amongst teenagers. Finally, one can think of differences in culture across the world. Here, a narrative might propose that differences in weather patterns provide an explanation for some of the differences between Southern and Northern Europeans.

Each of these collections of events can equivalently be thought of as a data base, where the narrative provides an explanation for the data.<sup>32</sup> Each of the examples is analogous to a particular data structure—the first corresponds to a *time series*, the second to a *panel* and the third (arguably) to a *cross-section*. Broadly construed, a narrative is a subjective interpretation of the data—an explanation that makes sense of the data. This analogy between the narrative-creator who tries to make sense of a collection of events and the econometrician or the statistician also highlights some specific features of narratives. First, like the econometrician who must select the relevant variables for her empirical specification, this sense-making effort often involves selecting a subset of variables from a large (possibly infinite) set of possible variables that one views as important to focus on. Second, like there may be unobserved variables in a dataset, an individual constructing a narrative is often missing information and can only work with the events they know about. Third, like econometric models can be used for forecasting the impact of a policy, an individual who constructs a narrative to make sense

---

<sup>30</sup>One conspicuous exception to this is the idea of a moral narrative discussed by Bénabou et al. (2020). This moral narrative any signal or message that shifts an agent’s belief about the externality of a moral action they are considering. While this Bénabou et al. (2020) definition can certainly incorporate narratives that involve making sense of existing data (when this shifts beliefs about a moral trade-off), the definition is far broader. It also considers simple hard evidence, which requires no interpretation, as well as fake news that contains no information as falling under the working definition of a narrative. Therefore, the Bénabou et al. (2020) definition focuses more on the implications of the narrative, while remaining very agnostic on its form. This is in contrast to the other definitions which take a substantially stronger stance on what constitutes a narrative.

<sup>31</sup>For some early discussions of why it would be beneficial to incorporate the notion of a drive for “sense-making” into economic analysis, see, e.g., Karlsson, Loewenstein, McCafferty, et al. (2004) and Chater and Loewenstein (2016).

<sup>32</sup>Note, we are thinking of a *data base* here in a broad sense. Therefore, for example, our working definition includes memory data bases, which contain a set of events stored in an individual’s (or group of individuals’) memory. It also includes the storage of a collection of events via any other format, including in a set of history books or in a spreadsheet.



of existing data may then also be used to forecast future events.<sup>33</sup>

Therefore, the common thread present in the extant literature in economics is that we can consider the following as a broad definition of a narrative:

*“A causal explanation that makes sense of a collection of events.”*

Under this broad definition, a narrative is very similar to a *subjective model*. We do not draw a bright line between the two concepts. When people talk about a narrative, they are very often referring to a particular type of model. However, one characteristic of a narrative under this broad definition that we would like to highlight is that this definition ties the narrative to a particular collection of events (i.e., to a particular data set). In this sense, the narrative does not live on its own independent of any data. This distinguishes even this broad class of narratives from theories or models which can be postulated in the absence of any existing data that they are aiming to explain. Therefore, a narrative can be thought of as a type of model that explains a particular set of events (or data set). A narrative is attached to a fixed data set, while a model may stand alone.

## A.2 The main constituent components of a narrative

To highlight where the different working definitions of a narrative differ, it is informative to break the concept down and consider the main components that one might think of as comprising a narrative. It is important to note that the early contributions to this narratives literature in economics (discussed above) are generally not considering mutually inconsistent working definitions of the concept, but are rather choosing to focus on different elements of what might be considered a narrative. We hope that the following discussion helps to clarify this and illustrate how these various projects provide complementary contributions to the collective scientific endeavour of better understanding the role of narratives in economics.

As noted above, we view a narrative as providing an explanation for a collection of events. The construction of such a sense-making explanation (or narrative) can be thought of as comprising the following parts.

**Selection of the events to be explained:** A key step in constructing a narrative is selecting the collection of events that require an explanation. Here, we consider two dimensions on which narratives may differ that can influence this event selection.

First, it is important to consider whether the narrative aims to explain the causes of one particular outcome of interest (e.g., the causes of World War II or the 2007 Financial Crisis).

---

<sup>33</sup>One key difference between our everyday narrative builder and the econometrician is that the narrative builder is not necessarily constrained by good statistical practices. He may construct a narrative that is as simple or as complex as he wishes. If he constructs the narrative for himself, he is constrained by his own view of what constitutes a plausible narrative, given the data. If he constructs it for others, he is constrained by his perception of what they might find credible. The narrative need not be identified in the data.

For such narratives in this *single-outcome-of-interest* class, the selection criterion for an event to be included in the collection of events to be explained, is that it must be viewed as an important causal antecedent to the event of interest.<sup>34</sup> Many of the types of narratives that we are interested in in economics will fall within this class. This class stands in contrast to the class of narratives where there is no primacy of a single event from the collection of events and the narrative simply aims to provide an explanation of the entire set. For example, a narrative of the cultural evolution in Western society over the 20th century does not necessarily assign primacy to the terminal point, but should provide an explanation that places all major cultural events on an equal footing. We refer to such narratives as *multiple-outcomes-of-interest* narratives.

Second, the narrative may either consider a collection of events as being unique or as being part of a repeating pattern. For example, one narrative may focus on explaining the causes of a particular recession, while another may propose an explanation for common causes of recessions in general. Similarly, one narrative might propose an explanation for high inflation observed in the US in late 2021 and early 2022, while another might propose an explanation for common causes of inflation more generally (see Andre et al., 2022, for further discussion of this example). Therefore, a second important dimension for determining which events are included in the collection to be explained is whether the narrative is a *singular narrative* (single-series) or a *generic narrative* (repeated-series).

**Imposing a (subjective) causal structure that connects events:** Given a selected collection of events, a narrative normally involves imposing a causal structure that connects the events. It is this causal structure selection and formation that is the focus of one central thread of the early theoretical work on narratives, which examine how individuals might select which directed acyclic graph (DAG) to adopt to explain a particular data set (e.g. Spiegler, 2016; Eliaz and Spiegler, 2020). Following on from this, some of the early empirical contributions have also focused on studying how individuals end up holding beliefs that can be represented by a particular DAG (Andre et al., 2022; Charles and Kendall, 2022).

As noted above, one can think of the set of events being explained as variables captured in a data set. When thinking through this lens, it will also often be convenient to use the language of variables instead of events. This literature on causal narratives predominantly focuses on how individuals end up believing in a certain causal structure connecting a set of variables, and what the implications of adopting an incorrect DAG may be. Therefore, the focus is on mistakes that arise in forming a subjective causal structure that connects the variables of interest. This captures an important class of mistakes that we are interested in when thinking about narratives. However, conditional on holding a particular (possibly correct) DAG in mind,

---

<sup>34</sup>It is worth also noting that the outcome of interest that a narrative focuses on may sometimes be explicit (as in the case of causal explanations for World War II or the 2007 Financial Crisis), but it may also sometimes be more subtle or implicit. For example, the biography of a noteworthy individual may involve plotting a causal path through selected events from their life that culminate in (and thereby explain) the noteworthy achievement or event in their life. Here, there is still a single focus that the narrative aims to explain (and which guides event selection).

there is still an additional cognitive step required to operationalize the causal model in forming beliefs that may then guide action choices. The DAG itself does not pin down the functional form nor the parameterization of the relationship between variables. Therefore, even with the correct DAG in mind, mistakes in narrative formation may arise. This is discussed further in the next section.

**Beliefs about the precise parameterization of the relationship between events:** A directed acyclic graph provides a lattice that describes causal links between variables. Variables are linked by edges, which are acyclic, implying there can be no causal path that is circular. However, the DAG does not identify the precise nature of the relationship between two variables (e.g., the sign, functional form and parameterization of the causal relationship). This means that even when there is agreement about the relevant set of variables and about the shape and direction of the causal structure, there is still substantial scope for disagreement about the precise relationship between the variables. In other words, there may still be disagreement about the best explanation for a given collection of events (where the events correspond to variables in the DAG). This implies that there is scope for differences in narratives to arise on different levels—either at the level of constructing the subjective causal structure (DAG) or at the level of forming beliefs about the functional form and parameterization of the relationship between variables. While the work by Spiegler (2016), Eliaz and Spiegler (2020), Andre et al. (2022), and Charles and Kendall (2022) focuses predominantly on the first level of narrative formation (i.e., *narrative construction*), Schwartzstein and Sunderam (2021) and the current paper focus on better understanding the second level of narrative formation (i.e., *narrative calibration*).

### A.3 Narratives in this paper

The discussion above serves to highlight some of the characteristics that distinguish different types of narratives from one another (i.e., single- vs multiple-outcomes of interest and singular vs generic narratives). It also dichotomized the process of narrative formation into two broad stages—narrative construction and narrative calibration. Given the wide array of scenarios that the term narrative is applied to, this is necessarily a partial taxonomy, however, in our view it provides a useful starting point for thinking about how different contributions to this nascent literature relate to one another.

In this paper, we are focusing on singular, single-outcome-of-interest narratives. In this context, we study narrative calibration. This focus on narrative calibration is one feature of our study that differentiates it from the contemporaneous empirical work (see, e.g., Andre et al., 2022; Charles and Kendall, 2022). However, there are also several other features in which we differ. First, we follow Schwartzstein and Sunderam (2021) in focusing on a particular feature of narrative selection, namely *narrative fit*. In this, we study the role played by a narrative being convincing when compared to the data. We, therefore, differ from analyses of other types of narrative selection rules, such as the adoption of “hopeful narratives” studied

theoretically by Eliaz and Spiegler (2020) and empirically by Charles and Kendall (2022). Second, the application studied in our paper is that of an expert advisor who wishes to persuade an investor. While this expert-advisee application is relevant for a broad class of scenarios that we are interested in, it has some unique features. Therefore, it raises a specific set of research questions that are not examined in work studying narratives in other contexts. For example, we study the behavior of both types of individuals in the advice scenario. We examine the narrative construction strategies expert advisors may use to form compelling narratives with the intent to pull investors' beliefs in a particular direction. We also analyze the effectiveness of these narratives in persuading investors. The contemporaneous empirical work focuses on other questions, such as the formation and implications of holding different narratives about inflation in 2021/2 in the US (Andre et al., 2022), and the adoption of hopeful narratives and narrative transmission (Charles and Kendall, 2022). Neither study examines persuasion by a conflicted expert advisor and the set of research questions raised by this context.

Finally, to be clear about what we mean in this paper when we refer to a narrative, we are essentially using the broad definition described above—namely *a causal explanation that makes sense of a collection of events*. However, in our experiment, we will restrict participants to think about a particular set of possible narratives. In doing this, we impose a (true) causal structure that is common knowledge to all participants. By fixing the causal structure of the narrative, we only leave the narrative calibration channel open for advisors to use to persuade investors. This provides us with the experimental control required to construct behavioral benchmarks and address our research questions of interest. Even the remaining flexibility provided when the “narrative construction” channel is closed still provides an extremely rich setting for studying narrative persuasion.

## B Additional Tables

Table 7: Belief updating and narrative fit (with continuous EPI variables)

	(1) $ \theta_{post}^{I,1} - \theta_{post}^{I,0} $	(2) $ \theta_{post}^{I,1} - \theta_{post}^{I,0} $	(3) $ \theta_{post}^{I,1} - \theta_{post}^{I,0} $	(4) $ \theta_{post}^{I,1} - \theta_{post}^{I,0} $
$ \theta_{post}^{I,0} - \theta_{post}^A $	0.398*** (0.0412)	0.464*** (0.0443)	0.382*** (0.0408)	0.451*** (0.0433)
$EPI^A$	-1.953*** (0.507)	-1.556*** (0.554)		
$ \theta_{post}^{I,0} - \theta_{post}^A  \times EPI^A$	0.121*** (0.0374)	0.0907** (0.0359)		
Misaligned sender	-0.875 (0.782)	-0.824 (0.864)	-0.745 (0.765)	-0.607 (0.837)
$(EPI^A - EPI^{I,0})$			-1.987*** (0.674)	-1.894** (0.730)
$ \theta_{post}^{I,0} - \theta_{post}^A  \times (EPI^A - EPI^{I,0})$			0.129*** (0.0281)	0.118*** (0.0329)
Dependent variable mean	11.102	12.35	11.102	12.35
Incl. opposite updaters	Yes	No	Yes	No
Round FE	Yes	Yes	Yes	Yes
Incl. aligned advisors	Yes	Yes	Yes	Yes
Observations	900	779	900	779

(i) The regressions use data from the INVESTORPRIOR treatment, (ii) The outcome variable in the regressions in this table is the absolute distance that investors update,  $|\theta_{post}^{I,1} - \theta_{post}^{I,0}|$  (iii) In columns (2) and (4), we remove observations in which the investor updates their belief in the opposite direction to the message sent by the advisor (e.g. updating upwards after receiving a message where  $\theta_{post}^{I,0} > \theta_{post}^A$ ), (vi) The  $EPI^A$  and  $EPI(m^A) - EPI(m^I)$  variables have been standardized to have mean zero and std. deviation one in order to make the coefficient magnitudes comparable, (v) Standard errors are clustered at the Interaction Group level, reported in parentheses \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ , (vi) For each of the investors, we have 10 observations—one for each round.

Table 8: Evaluating the impact of interventions (distance to advisor message)

	DISCLOSURE $ \theta_{post}^I - \theta_{post}^A $		INVESTORPRIOR $ \theta_{post}^I - \theta_{post}^A $		PRIVATEDATA $ \theta_{post}^I - \theta_{post}^A $	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Treatment	2.038* (1.088)	2.488 (1.619)	1.730 (1.044)	0.914 (1.158)	3.020*** (1.090)	-0.442 (1.179)
Advisor lied		3.855*** (0.911)		3.710*** (0.921)		3.685*** (0.916)
Treatment $\times$ Advisor lied		-0.521 (1.825)		1.227 (1.681)		4.696*** (1.624)
BASELINE mean	11.587	11.587	11.587	11.587	11.587	11.587
Round FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1800	1800	1800	1800	1800	1800

(i) The dependent variable is the distance between the advisor's message  $\theta_{post}^A$  and the corresponding belief held by the investor  $\theta_{post}^I$ , (ii) Standard errors are clustered at the Interaction Group level, reported in parentheses \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ , (iii) The regressions are estimated using data from investors who are matched with misaligned advisors (i.e., rounds in which investors are matched with aligned advisors are excluded), (iv) Each column uses data from the BASELINE treatment as well as the relevant treatment mentioned in the column header.

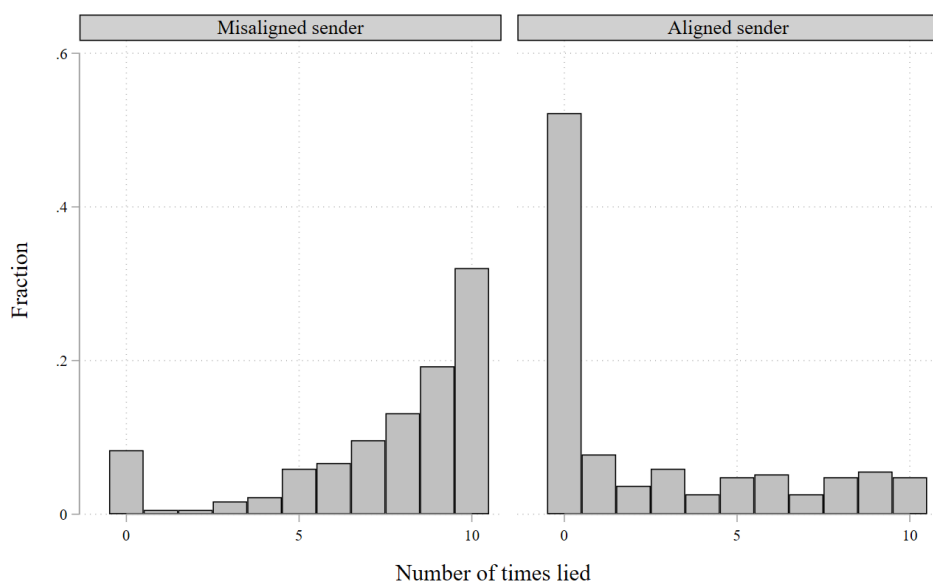
Table 9: Evaluating the impact of interventions (aligned advisor)

	DISCLOSURE $ \theta_{post}^I - \theta_{post}^T $ (1a)	DISCLOSURE $ \theta_{post}^I - \theta_{post}^A $ (1b)	INVESTORPRIOR $ \theta_{post}^I - \theta_{post}^T $ (2a)	INVESTORPRIOR $ \theta_{post}^I - \theta_{post}^A $ (2b)	PRIVATEDATA $ \theta_{post}^I - \theta_{post}^T $ (3a)	PRIVATEDATA $ \theta_{post}^I - \theta_{post}^A $ (3b)
Treatment	-4.597*** (0.994)	-5.075*** (0.934)	-0.0800 (0.972)	-0.278 (1.029)	0.530 (1.110)	0.632 (1.154)
BASELINE mean	10.163	10.082	10.163	10.082	10.163	10.082
Round FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	900	900	900	900	900	900

(i) The dependent variable in the (\*a) columns is the distance between the true  $\theta_{post}^T$  parameter and the corresponding belief held by the investor  $\theta_{post}^I$ , (ii) The dependent variable in the (\*b) columns is the distance between the advisor's message  $\theta_{post}^A$  and the corresponding belief held by the investor  $\theta_{post}^I$ , (iii) Standard errors are clustered at the Interaction Group level, reported in parentheses \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ , (iv) The regressions are estimated using data from investors who are matched with aligned advisors (i.e., rounds in which investors are matched with misaligned advisors are excluded), (iv) Each column uses data from the BASELINE treatment as well as the relevant treatment mentioned in the column header.

## C Additional Figures

Figure 12: Distribution of lying across ten rounds (by advisor type)



Graphs by Aligned sender



Figure 13: Advisor  $\theta_{post}^A$  vs True  $\theta_{post}^T$  (by advisor type)

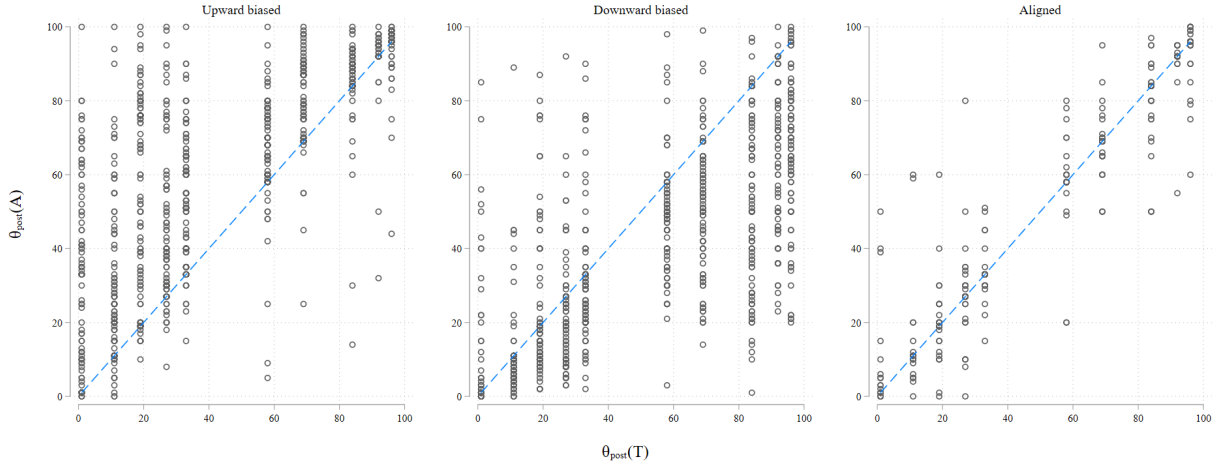


Figure 14: Advisor  $\theta_{pre}^A$  vs True  $\theta_{pre}^T$  (by advisor type)

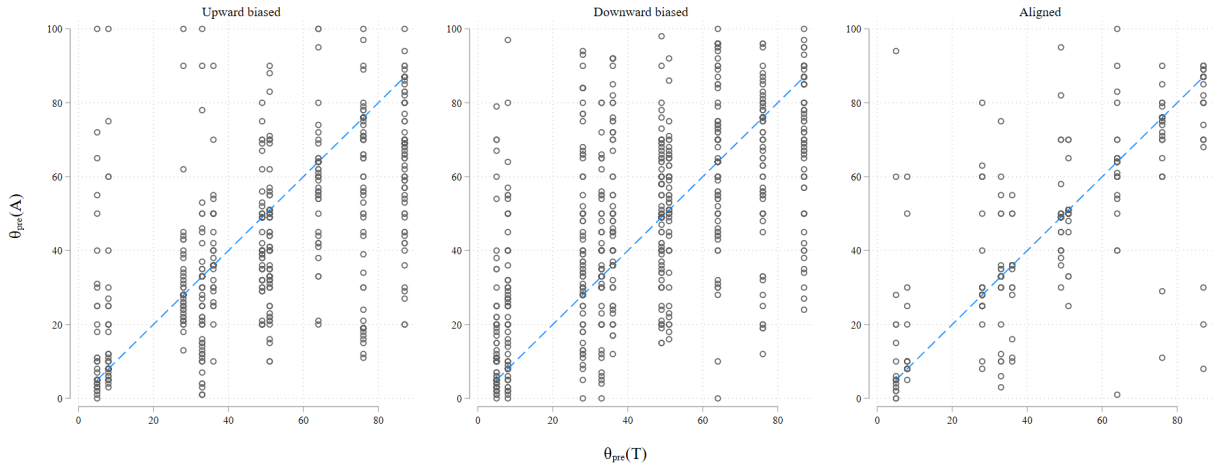
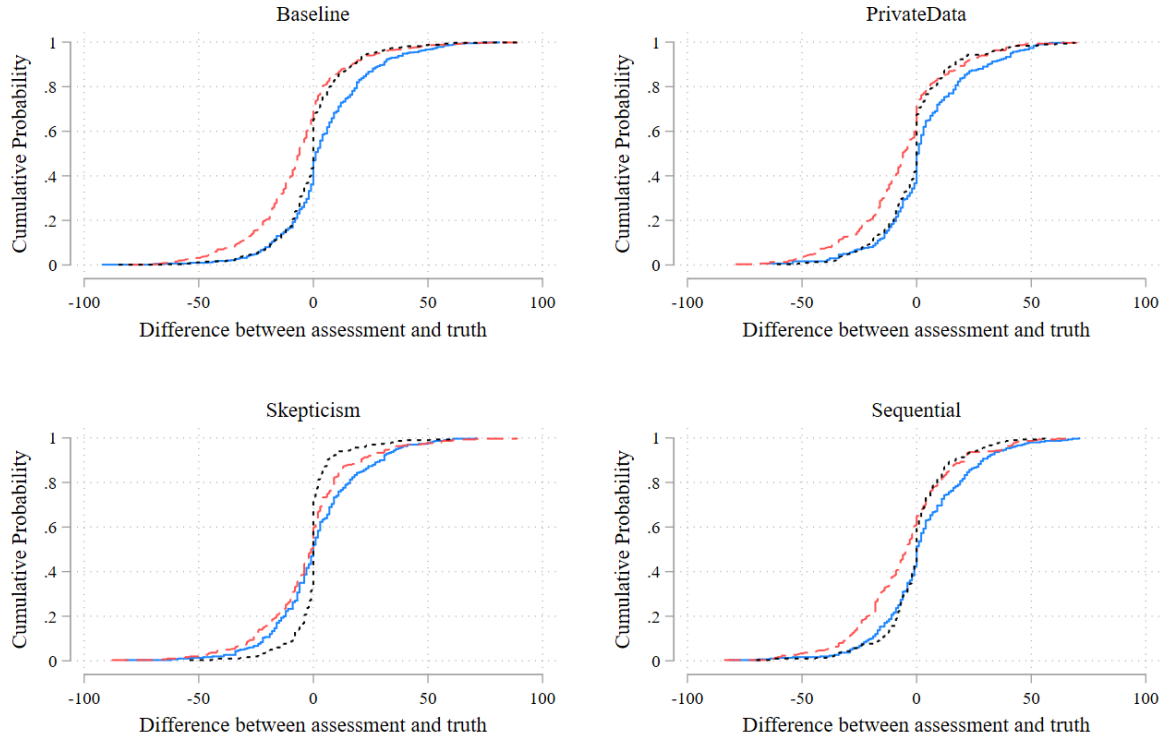
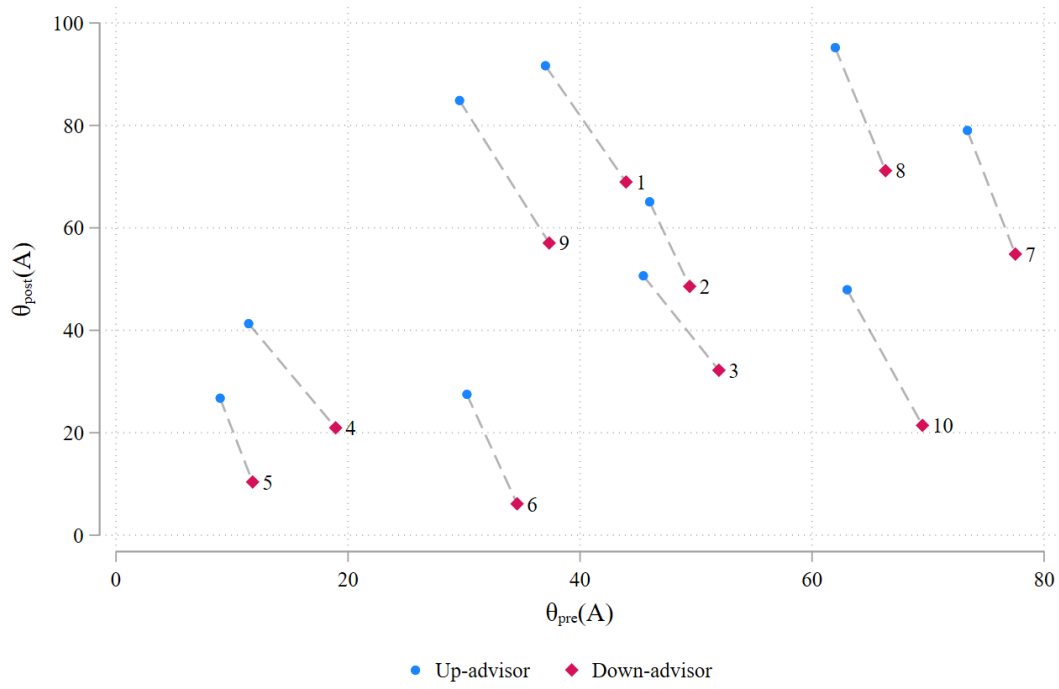


Figure 15: Difference between  $\theta_{post}^I$  and  $\theta_{post}^T$  (by treatment and advisor type).



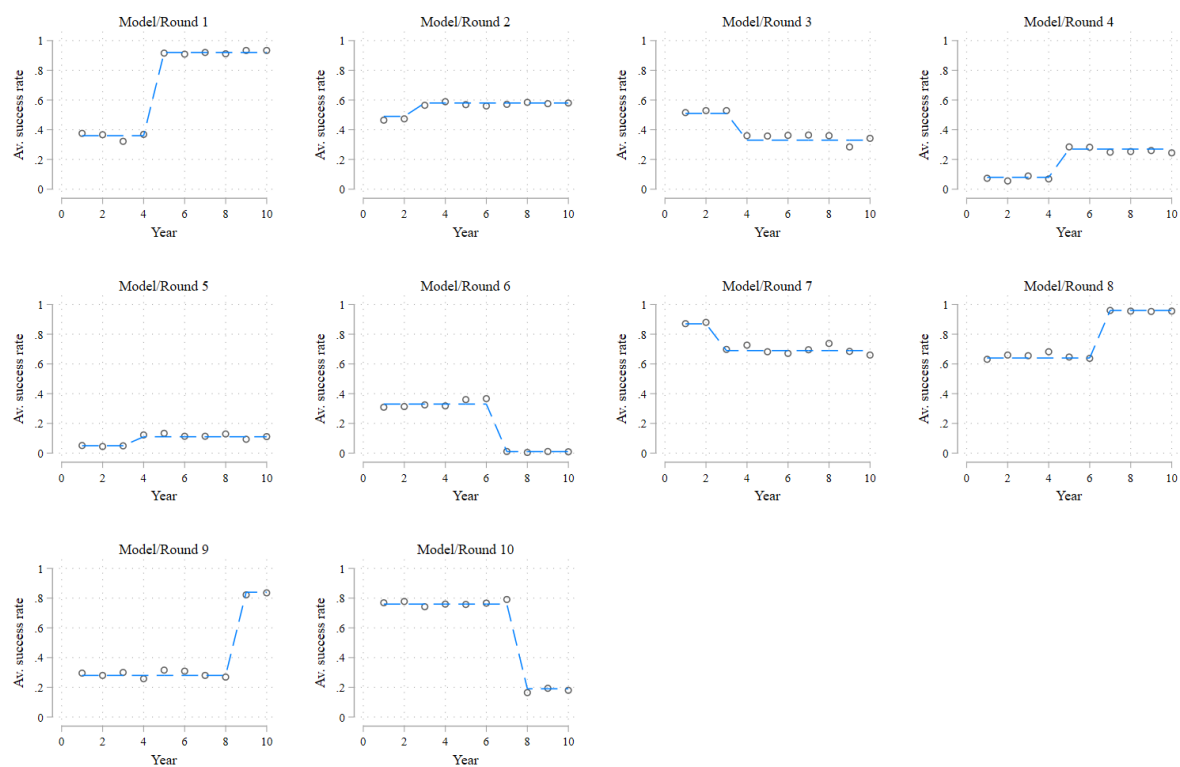
Notes: (i) The figure plots the cdf of the difference between the investor's belief and the truth,  $\theta_{post}^I - \theta_{post}^T$ , for all investor-rounds where the investor is matched with a particular advisor type, (ii) Each of the panels show this for a particular treatment condition, (iii) The red dashed line shows the cdf for investor-rounds where the investor is matched with down-advisor, the black dotted lines shows the cdf for investor-rounds where the investor is matched with aligned advisor, and the blue solid line shows the cdf for investor-rounds where the investor is matched with up-advisor.

Figure 16: Advisor  $\theta_{post}$  and  $\theta_{pre}$  reports in each round (by advisor type)



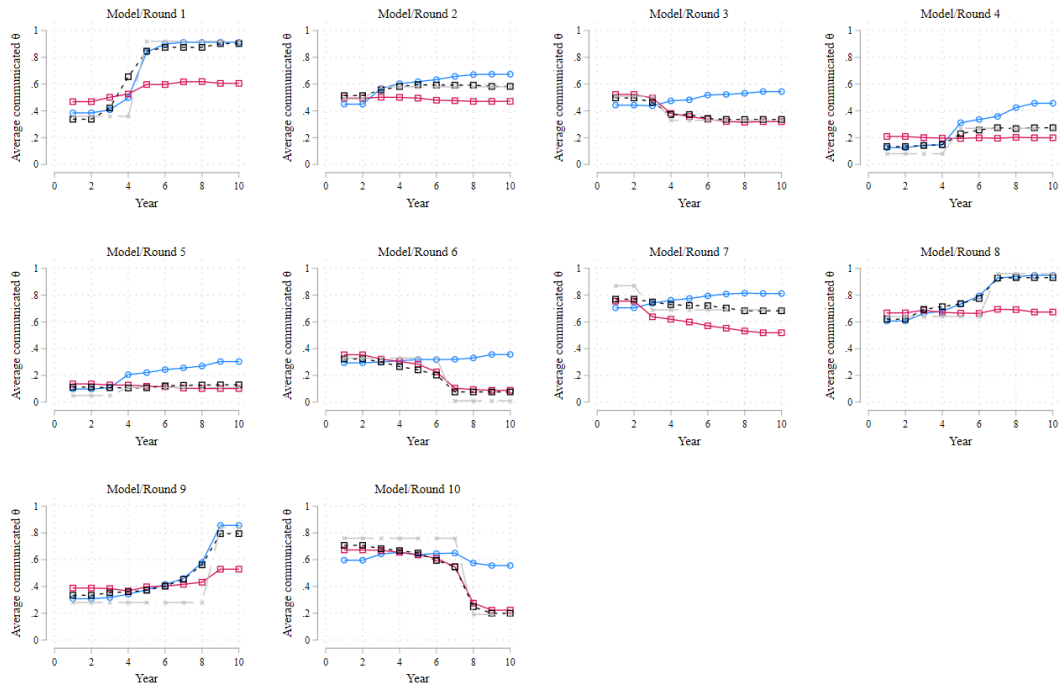
Notes: (i) The numbered labels in the figure denote the 10 rounds of the experiment, (ii) The blue markers show the average  $\theta_{post}$  and  $\theta_{pre}$  report by up-advisors in each round, while the red markers report the same for down-advisor. (iii) The figure shows that down-advisor reports are below and to the right of up-advisor reports, indicating that the advisors move their  $\theta_{post}$  and  $\theta_{pre}$  in opposing directions to construct convincing narratives.

Figure 17: Average observed history vs true model (by round)



Notes: (i) The figure shows the average history observed by investors in each of the rounds in comparison to the true underlying model generating the data.

Figure 18: Average narrative sent by each advisor type (by round)



Notes: (i) The figure shows the average history observed by investors in each of the rounds in comparison to the true underlying model generating the data.

## D Proofs

### D.1 Notation

Throughout the proofs, we will use a number of notational shortcuts. Define by

$$k_{pre}(c) \equiv \sum_{t=0}^c s_t, \quad f_{pre}(c) \equiv c - k_{pre}(c), \quad k_{post}(c) \equiv \sum_{t=c}^{10} s_t, \text{ and } f_{post}(c) \equiv 10 - c - k_{post}(c)$$

the numbers of successes and failures in the *pre* and *post* period for a given  $c$ . We will also use the convention that

$$\tilde{k}_p \equiv k_p(c^{DO}), \quad \tilde{f}_p \equiv f_p(c^{DO}), \quad k'_p \equiv k_p(c'), \text{ and } f'_p \equiv f_p(c').$$

We will sometimes denote differences in the number of successes in *post* under the structural change parameters  $c'$  and  $c^{DO}$  by  $\Delta k = k'_{post} - \tilde{k}_{post}$  and  $\Delta f = f'_{post} - \tilde{f}_{post}$ .

The log likelihood function is equal to

$$\ell(m) = k_{pre}(c) \ln(\theta_{pre}) + f_{pre}(c) \ln(1 - \theta_{pre}) + k_{post}(c) \ln(\theta_{post}) + f_{post}(c) \ln(1 - \theta_{post}).$$

### D.2 Proof of Proposition 1

Sending a model  $m' = (c', \theta'_{pre}, \theta'_{post}) \in \mathcal{M}(\bar{\ell})$  yields utility

$$\mathbb{E}[U(\theta_{post}^I, \varphi) | m'] = G(\bar{\ell}) U(\theta'_{post}, \varphi) + (1 - G(\bar{\ell})) \mathbb{E}[U(\theta_{post}^{I,0}, \varphi) | \bar{\ell} < \ell(m^{I,0})].$$

Note that any alternative model in  $\mathcal{M}(\bar{\ell})$  only changes the value of  $U(\cdot)$  in the first term of the utility function, while the values of all other functions remain fixed, as they only depend on  $\bar{\ell}$ . Therefore, choosing the model that maximizes utility for a given level of the model fit,  $\bar{\ell}$ , is equal to maximizing the utility the advisor receives if the investor adopts the model,  $U(\theta_{post}^A, \varphi)$ , with respect to  $\theta_{post}^A$ . This in turn is equal to minimizing  $(\varphi - \theta_{post}^A)^2$ .

### D.3 Proof of Proposition 2

Denote by  $(\hat{c}(\theta_{post}), \hat{\theta}_{pre}(\theta_{post}))$  the parameter values that maximize the log likelihood function conditional on  $\theta_{post}$ . We can then define the conditional log likelihood function as

$$\ell^C(\theta_{post}) \equiv \ell((\hat{c}(\theta_{post}), \hat{\theta}_{pre}(\theta_{post}), \theta_{post})).$$

The proof will proceed by showing and combining a number of claims. The first claim states that we can always find values of  $c$  and  $\theta_{pre}$  that, if combined with any  $\theta_{post}$ , lead to a message

fit between minus infinity and the value of the conditional log likelihood function evaluated at  $\theta_{post}$ . This claim follows from the continuity of the log likelihood function.

**Claim 1:** For every  $\theta_{post} \in [0, 1]$ , there are always parameter values  $c \in \{2, \dots, 8\}$  and  $\theta_{pre} \in [0, 1]$  so that  $\ell((c, \theta_{pre}, \theta_{post})) = \bar{\ell}$ , where  $\bar{\ell} \in (-\infty, \ell^C(\theta_{post})]$ . If  $\bar{\ell} = \ell^C(\theta_{post})$ , the claim directly follows as the model  $(\hat{c}(\theta_{post}), \hat{\theta}_{pre}(\theta_{post}), \theta_{post})$  induces likelihood value  $\bar{\ell}$ . Now consider  $\bar{\ell}$  taking on a value on the interior of the interval. We know that

$$\bar{\ell} < \ell(\hat{c}(\theta_{post}), \hat{\theta}_{pre}(\theta_{post}), \theta_{post}).$$

Now consider changing  $\hat{\theta}_{pre}$  to a level  $t$ . This will result in the log likelihood taking on value

$$\begin{aligned} \ell((\hat{c}(\theta_{post}), t, \theta_{post})) &= k_{pre}(\hat{c}(\theta_{post}))\ln(t) + f_{pre}(\hat{c}(\theta_{post}))\ln(1-t) \\ &\quad + k_{post}(\hat{c}(\theta_{post}))\ln(\theta_{post}) + f_{post}(\hat{c}(\theta_{post}))\ln(1-\theta_{post}). \end{aligned}$$

Observe that if  $k_{pre} > 0$ , the limit  $\lim_{t \rightarrow 0} \ell((\hat{c}(\theta_{post}), t, \theta_{post})) \rightarrow -\infty$  and that if  $f_{pre} > 0$ , the limit  $\lim_{t \rightarrow 1} \ell((\hat{c}(\theta_{post}), t, \theta_{post})) \rightarrow -\infty$ . As at least one of  $k_{pre}$  or  $f_{pre}$  is strictly positive, at least one limit will always diverge. Since  $\ell(\cdot)$  is continuous in  $t$ , the intermediate value theorem then guarantees the existence of at least one value of  $t$  so that  $\ell((\hat{c}(\theta_{post}), t, \theta_{post})) = \bar{\ell}$ .

The second claim builds on Claim 1, showing that, if  $m^*$  is not on the conditional log likelihood, its  $\theta_{post}^*$  has to be equal to  $\varphi$ .

**Claim 2:** Suppose that  $m^* \notin \mathcal{C}$ . Then,  $\theta_{post}^* = \varphi$ . Suppose by contradiction that  $m^*$  is not in  $\mathcal{C}$  and that  $\theta_{post}^* \neq \varphi$ . Consider permuting  $\theta_{post}^*$  by a small value  $\eta \in \{-\varepsilon, +\varepsilon\}$  to move it closer to the advisor's objective, where  $\varepsilon > 0$  is a small number. That is,  $\theta_{post}' = \theta_{post}^* + \eta$  and  $(\varphi - \theta_{post}')^2 < (\varphi - \theta_{post}^*)^2$ . By Claim 1, we know that a model  $m' = (c', \theta_{pre}', \theta_{post}')$  exists such that  $\ell(m') = \bar{\ell}$  as long as  $\theta_{post}^* \notin \mathcal{C}$ . By Proposition 1, the advisor prefers message  $m'$  to message  $m^*$ , which is contradicts the initial statement.

We proceed with Claim 3 which shows that, if  $\theta_{post}$  is fixed at  $\varphi$ , the advisor will prefer the message with the higher message fit.

**Claim 3:** Consider two messages  $m' = (c', \theta_{pre}', \varphi)$  and  $m'' = (c'', \theta_{pre}'', \varphi)$  and suppose that  $\ell(m') > \ell(m'')$ . The advisor prefers sending  $m'$  over sending  $m''$ . Denote by  $\Delta G$  the difference  $G(\ell(m')) - G(\ell(m''))$ . For notational brevity we will also use  $G'' \equiv G(\ell(m''))$ ,  $\ell' \equiv \ell(m')$ ,  $\ell'' \equiv \ell(m'')$ , and  $\ell^{I,0} \equiv \ell(m^{I,0})$ . We can then denote the expected utility of the

sender from sending  $m'$  as

$$\begin{aligned}
\mathbb{E}[U(\theta_{post}^I, \varphi)|m'] &= (G'' + \Delta G)U(\varphi, \varphi) + (1 - G'' - \Delta G)(\mathbb{E}[U(\theta_{post}^{I,0}, \varphi)|\ell'' < \ell^{I,0}]) \\
&= G''U(\varphi, \varphi) + (1 - G'' - \Delta G)(\mathbb{E}[U(\theta_{post}^{I,0}, \varphi)|\ell'' < \ell^{I,0}] + \Delta GU(\varphi, \varphi)) \\
&> G''U(\varphi, \varphi) + (1 - G'' - \Delta G)(\mathbb{E}[U(\theta_{post}^{I,0}, \varphi)|\ell'' < \ell^{I,0}]) \\
&\quad + \Delta G\mathbb{E}[U(\theta_{post}^{I,0}, \varphi)|\ell^{I,0} \in (\ell'', \ell')] \\
&= G''U(\varphi, \varphi) \\
&\quad + (1 - G'') \times \frac{(1 - G'' - \Delta G)(\mathbb{E}[U(\theta_{post}^{I,0}, \varphi)|\ell'' < \ell^{I,0}] + \Delta G\mathbb{E}[U(\theta_{post}^{I,0}, \varphi)|\ell^{I,0} \in (\ell'', \ell')])}{1 - G''} \\
&= G''U(\varphi, \varphi) + (1 - G'')\mathbb{E}[U(\theta_{post}^{I,0}, \varphi)|\ell(m') < \ell^{I,0}] = \mathbb{E}[U(\theta_{post}^I, \varphi)|m''].
\end{aligned}$$

The inequality above follows from the fact that the investor's prior has full support on  $M$ , so that the set of models among which the investor's default model is if the investor follows message  $m'$  but not message  $m''$  will always include some model with a value  $\theta_{post} < \varphi$  with positive probability, which implies that  $U(\varphi, \varphi) > \mathbb{E}[U(\theta_{post}^{I,0}, \varphi)|\ell^{I,0} \in (\ell'', \ell')]$ . Therefore,  $\mathbb{E}[U(\theta_{post}^I, \varphi)|m'] > \mathbb{E}[U(\theta_{post}^I, \varphi)|m'']$ , which proves the claim.

Combining claims 2 and 3, the statement of the proposition directly follows.

**Claim 4:**  $m^* \in \mathcal{C}$ . By Claim 2, we know that, if the optimal model is not in  $\mathcal{C}$ , then its  $\theta_{post}$ -parameter value is equal to  $\varphi$ . However Claim 3 implies that, among all models in  $M$  with  $\theta_{post} = \varphi$ , the advisor most prefers the model that is in  $\mathcal{C}$ , which implies Claim 4.

## D.4 Proofs of propositions 3 and 4

We will show both statements only for the up-advisor; symmetrical arguments can be made to also show them for the down-advisor.

### D.4.1 Preliminaries

Define a function that returns the log likelihood difference between  $m'$  and  $\tilde{m}$  for a given  $\theta_{post}$  by

$$\begin{aligned}
\Delta\ell(\theta_{post}) &\equiv \Delta k(\ln(\theta_{post}) - \ln(\theta'_{pre})) + \Delta f(\ln(1 - \theta_{post}) - \ln(1 - \theta'_{pre})) \\
&\quad + \underbrace{\tilde{k}_{pre}\ln(\theta'_{pre}) + \tilde{f}_1\ln(1 - \theta'_{pre}) - [\tilde{k}_{pre}\ln(\tilde{\theta}_{pre}) + \tilde{f}_1\ln(1 - \tilde{\theta}_{pre})]}_{=\kappa < 0}.
\end{aligned}$$

Since  $\ell$  is maximal at  $\ell(m^{DO})$ ,  $\Delta\ell(\theta_{post}^{DO}) < 0$ . The derivative is equal to

$$\Delta\ell'(\theta_{post}) = \frac{\Delta k}{\theta_{post}} - \frac{\Delta f}{1 - \theta_{post}}. \tag{3}$$



Furthermore, as  $\theta_{post}$  becomes large,

$$\begin{aligned} \lim_{\theta_{post} \rightarrow 1} \Delta \ell(\theta_{post}) &= \Delta k \left( \lim_{\theta_{post} \rightarrow 1} \ln(\theta_{post}) - \ln(\theta'_{pre}) \right) + \Delta f \left( \lim_{\theta_{post} \rightarrow 1} \ln(1 - \theta_{post}) - \ln(1 - \theta'_{pre}) \right) + \kappa \\ &= -\Delta k \ln(\theta'_{pre}) + \Delta f \left( \lim_{\theta_{post} \rightarrow 1} \ln(1 - \theta_{post}) - \ln(1 - \theta'_{pre}) \right) + \kappa \end{aligned} \quad (4)$$

and therefore  $\lim_{\theta_{post} \rightarrow 1} \Delta \ell(\theta_{post}) \rightarrow -\infty$  if  $\Delta f > 0$  and  $\lim_{\theta_{post} \rightarrow 1} \Delta \ell(\theta_{post}) \rightarrow \infty$  if  $\Delta f < 0$ . If  $\Delta f = 0$ , the limit is positive whenever

$$\begin{aligned} & -\Delta k \ln(\theta'_{pre}) + \tilde{k}_{pre} \ln(\theta'_{pre}) + \tilde{f}_1 \ln(1 - \theta'_{pre}) - [\tilde{k}_{pre} \ln(\tilde{\theta}_{pre}) + \tilde{f}_1 \ln(1 - \tilde{\theta}_{pre})] > 0 \\ \Rightarrow & k'_{pre} \ln(\theta'_{pre}) + f'_1 \ln(1 - \theta'_{pre}) - [\tilde{k}_{pre} \ln(\tilde{\theta}_{pre}) + \tilde{f}_1 \ln(1 - \tilde{\theta}_{pre})] > 0. \end{aligned}$$

When does this condition hold? Define a function

$$g(x) \equiv (\tilde{k}_{pre} + x) \ln \left( \frac{\tilde{k}_{pre} + x}{\tilde{k}_{pre} + \tilde{f}_{pre} + x} \right) + \tilde{f}_{pre} \ln \left( \frac{\tilde{f}_{pre}}{\tilde{k}_{pre} + \tilde{f}_{pre} + x} \right),$$

which has a derivative  $g'(x) = \ln((\tilde{k}_{pre} + x)/(\tilde{k}_{pre} + \tilde{f}_{pre} + x)) < 0$ . For  $\Delta f = 0$ , the limit becomes

$$\lim_{\theta_{post} \rightarrow 1} \Delta \ell(\theta_{post}) = g(-\Delta k) - g(0).$$

Therefore, if  $\Delta f = 0$  the limit as  $\theta_{post} \rightarrow 1$  is positive if  $\Delta k > 0$  and negative if  $\Delta k < 0$ .

#### D.4.2 Proof of Proposition 3

By Assumption ??, the up-advisor prefers  $m'$  over  $\tilde{m}$  if and only if  $\Delta \ell(\theta_{post}^*) \geq 0$ .

If  $c' < c^{DO}$ ,  $\Delta k, \Delta f \geq 0$ , with at least one inequality strict. We consider whether  $\Delta \ell(\theta_{post}^*) \geq 0$  is possible in a number of cases:

**Case 1:**  $\Delta k > 0, \Delta f = 0$ . As  $\theta_{post} \rightarrow 1$ ,  $\Delta \ell(\theta_{post}) > 0$  (see Equation (4) and the discussion afterwards). Furthermore, the derivative in Equation (3) shows that  $\Delta \ell$  is strictly increasing in  $\theta_{post}$ . There is thus one critical value  $\theta_{post}^C > \theta_{post}^{DO}$  so that  $\Delta \ell(\theta_{post}) \geq 0$  whenever  $\theta_{post} \geq \theta_{post}^C$ . The advisor prefers message  $m'$  over  $\tilde{m}$  if and only if  $\theta_{post}^* \geq \theta_{post}^C$ .

**Case 2:**  $\Delta k = 0, \Delta f > 0$ . As  $\theta_{post} \rightarrow 1$ ,  $\Delta \ell(\theta_{post}) < 0$  (see Equation (4) and the discussion afterwards). Furthermore, the derivative in Equation (3) shows that  $\Delta \ell$  is strictly decreasing in  $\theta_{post}$ . As  $\theta_{post}^* \geq \theta_{post}^{DO}$  and  $\Delta \ell(\theta_{post}^{DO}) < 0$ , the advisor never prefers message  $m'$  over  $\tilde{m}$ .

**Case 3:**  $\Delta k > 0, \Delta f > 0$ . As  $\theta_{post} \rightarrow 1$ ,  $\Delta \ell(\theta_{post}) < 0$  (see Equation (4) and the discussion afterwards). Furthermore, the derivative in Equation (3) shows that  $\Delta \ell$  is first increasing and

then decreasing in  $\theta_{post}$ . The derivative changes its sign exactly once at the point

$$\theta_{post}^0 \equiv \frac{\Delta k}{\Delta k + \Delta f}.$$

Rearranging, we find that

$$\theta_{post}^0 > \theta_{post}^{DO} \iff \frac{k'_{post}}{1 - c'} > \frac{\tilde{k}_{post}}{1 - c^{DO}}.$$

As  $\theta_{post}^* \geq \theta_{post}^{DO}$  and  $\Delta\ell(\theta_{post}^{DO}) < 0$ , a necessary condition for the advisor to prefer  $m'$  over  $m^{DO}$  is that  $\Delta\ell(\theta_{post}^{DO})' > 0$ , which is only the case if  $k'_{post}/(1 - c') > \theta_{post}^{DO}$ .

In summary, we find that the advisor prefers  $m'$  over  $m^{DO}$  only in Cases 1 or 3 and only if  $k'_{post}/(1 - c') > \theta_{post}^{DO}$ .

#### D.4.3 Proof of Proposition 4

By Assumption ??, the up-advisor prefers  $m'$  over  $\tilde{m}$  if and only if  $\Delta\ell(\theta_{post}^*) \geq 0$ .

If  $c' > c^{DO}$ , then  $\Delta k, \Delta f \leq 0$  with at least one inequality strict. We consider whether  $\Delta(\theta_{post}^*) \geq 0$  is possible in three cases.

**Case 1:**  $\Delta k < 0, \Delta f = 0$ . As  $\theta_{post} \rightarrow 1$ ,  $\Delta\ell(\theta_{post}) < 0$  (see Equation (4) and the discussion afterwards). Furthermore, the derivative in Equation (3) shows that  $\Delta\ell$  is strictly decreasing in  $\theta_{post}$ . As  $\theta_{post}^* \geq \theta_{post}^{DO}$  and  $\Delta\ell(\theta_{post}^{DO}) < 0$ , the advisor never prefers message  $m'$  over  $\tilde{m}$ .

**Case 2:**  $\Delta k = 0, \Delta f < 0$ . As  $\theta_{post} \rightarrow 1$ ,  $\Delta\ell(\theta_{post}) > 0$  (see Equation (4) and the discussion afterwards). Furthermore, the derivative in Equation (3) shows that  $\Delta\ell$  is strictly increasing in  $\theta_{post}$ . There is thus one critical value  $\theta_{post}^C > \theta_{post}^{DO}$  so that  $\Delta\ell(\theta_{post}) \geq 0$  whenever  $\theta_{post} \geq \theta_{post}^C$ . The advisor prefers message  $m'$  over  $\tilde{m}$  if and only if  $\theta_{post}^* \geq \theta_{post}^C$ .

**Case 3:**  $\Delta k < 0, \Delta f < 0$ . As  $\theta_{post} \rightarrow 1$ ,  $\Delta\ell(\theta_{post}) > 0$  (see Equation (4) and the discussion afterwards). Furthermore, the derivative in Equation (3) shows that  $\Delta\ell$  is first decreasing and then increasing in  $\theta_{post}$ . There is thus one critical value  $\theta_{post}^C > \theta_{post}^{DO}$  so that  $\Delta\ell(\theta_{post}) \geq 0$  whenever  $\theta_{post} \geq \theta_{post}^C$ . The advisor prefers message  $m'$  over  $\tilde{m}$  if and only if  $\theta_{post}^* \geq \theta_{post}^C$ .

In summary, we find that the advisor prefers  $m'$  over  $m^{DO}$  only if  $\Delta f < 0$ .

## E Discussion of the Nash equilibrium

In the following we formally derive equilibria of the cheap talk game that is underlying the investor-advisor setup.

## E.1 Setup

Consider a game between an advisor and an investor. There is an unknown state of the world  $\theta_{post} \in \Theta = [0, 1]$ . This state is distributed with full support on  $\Theta$  according to a commonly known prior distribution  $f(\theta_{post})$ . After learning the true state of the world, the advisor sends a message  $m \in M = [0, 1]$  to the investor. After hearing the advisor's message, the investor makes an assessment  $\theta_{post}^I \in A = [0, 1]$  of the state of the world. Payoffs of both advisor and investor depend on a scoring rule. The advisor's utility function is

$$U^A(\theta_{post}^I, \varphi) = 1 - (\varphi - \theta_{post}^I)^2,$$

where  $\varphi$  varies with the advisor's type. Its value is equal to  $\theta_{post}$  if the advisor is aligned. Such an advisor always gets the maximal payoff if the investor assesses the state of the world accurately. A misaligned advisor is either upward or downward biased. I.e., the up-advisor has a  $\varphi = 1$  and the down-advisor has a  $\varphi = 0$ . The misaligned advisor thus maximizes her payoff if the investor makes the maximum or minimum assessment. To summarize, there are three advisor types  $\varphi \in \{0, \theta_{post}, 1\}$ . The advisor can be of either type with equal probability.

There is only one type of investor who has utility function

$$U^I(\theta_{post}^I, \theta) = 1 - (\theta_{post} - \theta_{post}^I)^2.$$

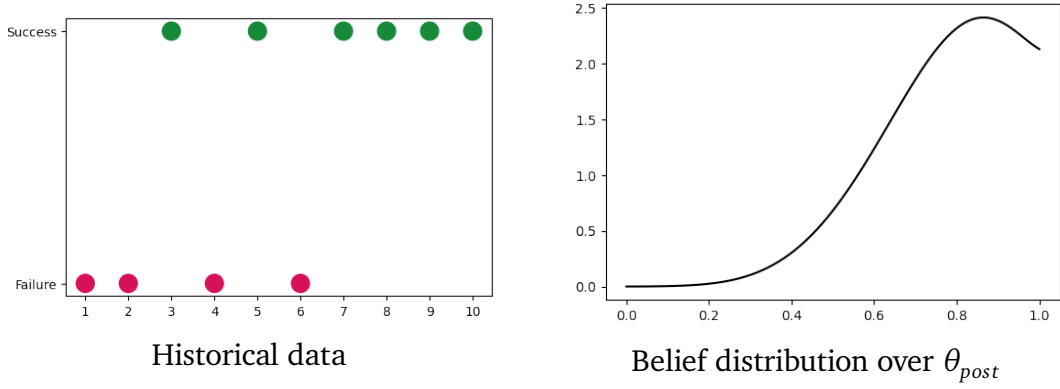
The investor maximizes utility if he makes an accurate assessment.

**Remark: relating theory to design; the case of the historical data** We can think of the historical data, jointly with the information that the three parameter values  $c, \theta_{pre}, \theta_{post}$  are uniformly distributed on  $\{2, 8\} \times [0, 1]^2$  ex-ante, determining the prior belief  $f(\theta_{post})$ . Formally, upon seeing the data, the investor can form a Bayesian posterior belief which is equal to

$$f(\theta_{post}) = \sum_{c=2}^8 \frac{\int_0^1 \mathcal{L}(c, \theta_{pre}, \theta_{post}) d\theta_{pre}}{\sum_{c=2}^8 \int_0^1 \int_0^1 \mathcal{L}(c, \theta_{pre}, \theta_{post}) d\theta_{pre} d\theta}. \quad (5)$$

In the equation above,  $\mathcal{L}(c, \theta_{pre}, \theta_{post}) = \theta_{pre}^{k_{pre}(c)} (1 - \theta_{pre})^{f_{pre}(c)} \theta_{post}^{k_{post}(c)} (1 - \theta_{post})^{f_{post}(c)}$  is the likelihood function, and  $k_p(c)$ ,  $f_p(c)$  denote the number of successes and failures in the *pre* and *post* period for a given structural change parameter value  $c$ . We can explicitly solve for the posterior distribution  $f$  by noting that  $B(k+1, f+1) \equiv \int_0^1 \theta^k (1-\theta)^f d\theta$  is the beta function and  $g(\theta|k+1, f+1) \equiv \theta^k (1-\theta)^f / B(k+1, f+1)$  is the density function of the beta distribution with shape parameters  $k+1$  and  $f+1$ . Substituting the likelihood terms out of Equation (5),

Figure 19: Example of a history and corresponding prior belief over  $\theta_{post}$



we find that

$$f(\theta_{post}) = \sum_{c=2}^8 w_c g(\theta_{post} | k_{post}(c) + 1, f_{post}(c) + 1),$$

$$\text{where } w_c \equiv \frac{B(k_{pre}(c) + 1, f_{pre}(c) + 1) B(k_{post}(c) + 1, f_{post}(c) + 1)}{\sum_{c'=2}^8 B(k_{pre}(c') + 1, f_{pre}(c') + 1) B(k_{post}(c') + 1, f_{post}(c') + 1)}.$$

Therefore, the investor's belief distribution over  $\theta_{post}$  is a mixture of beta distributions with expectation  $\mathbb{E}(\theta_{post}) \in (0, 1)$ . Figure 19 shows the investor's belief for an example historical data set.

## E.2 Equilibrium

In the described game, the advisor's strategy  $m^* : \{\theta, 1\} \times \Theta \rightarrow \Delta M$  maps from the advisor's type and the state of the world into a probability distribution over messages. The investor's strategy  $\theta_{post}^{I*} : M \rightarrow \Delta A$  maps the received message into a distribution over assessments.

We are interested in persuasive equilibria of this game. Following Little (2022), a persuasive equilibrium is an equilibrium in which the investor is sometimes responsive to the advisor's message.

**Definition 1.** A message is persuasive if and only if  $\theta_{post}^{I*}(m) \neq \mathbb{E}(\theta_{post})$ . A persuasive equilibrium is an equilibrium where a persuasive message is sent with strictly positive probability.

The current game has two types of persuasive equilibria.

**Proposition 5.** There exists a persuasive equilibrium in which the advisor sends one of two messages  $m'$  and  $m''$ . The equilibrium is characterized by a threshold  $\hat{\theta} \in \Theta$ . The aligned advisor sends  $m'$  if  $\theta_{post} \leq \hat{\theta}$  and  $m''$  otherwise. The up-advisor always sends  $m''$  and the down advisor always sends  $m'$ . Upon receiving message  $m$ , the investor makes assessment  $\mathbb{E}[\theta | m]$ .

*Proof.* Suppose the described equilibrium exists. The aligned advisor will prefer sending  $m'$  over  $m''$  if

$$(\theta_{post} - \theta_{post}^I(m'))^2 \leq (\theta_{post} - \theta_{post}^I(m''))^2.$$

This implies the existence of a unique threshold  $\hat{\theta} = (\theta_{post}^I(m') + \theta_{post}^I(m''))/2$  so that the aligned advisor sends  $m'$  if and only if  $\theta_{post} \leq \hat{\theta}$ . The investor's best response to message  $m'$  then is to play

$$\theta_{post}^{I*}(m'; \hat{\theta}) = \mathbb{E}(\theta|m') = p(\hat{\theta})\mathbb{E}(\theta|\theta \leq \hat{\theta}) + (1 - p(\hat{\theta}))\mathbb{E}(\theta),$$

where  $p(\hat{\theta}) \equiv F(\hat{\theta})/(1 + F(\hat{\theta}))$  is the probability that the message was sent by the aligned advisor. Upon receiving  $m''$ , the investor's best response is

$$\theta_{post}^I(m''; \hat{\theta}) = \mathbb{E}(\theta|m'') = q(\hat{\theta})\mathbb{E}(\theta|\theta > \hat{\theta}) + (1 - q(\hat{\theta}))\mathbb{E}(\theta),$$

where  $q(\hat{\theta}) \equiv (1 - F(\hat{\theta}))/2$  is the probability that the message  $m''$  was sent by the aligned advisor. Therefore, in equilibrium,

$$\hat{\theta} = \frac{1}{2} [\theta_{post}^{I*}(m'; \hat{\theta}) + \theta_{post}^I(m''; \hat{\theta})].$$

Define a function  $\Phi(\theta) \equiv \theta - 1/2 [\theta_{post}^{I*}(m'; \theta) + \theta_{post}^I(m''; \theta)]$ . An equilibrium obtains where  $\Phi(\hat{\theta}) = 0$ . Since

$$\Phi(0) = 0 - \frac{1}{2} [\mathbb{E}(\theta) + \mathbb{E}(\theta)] = -\mathbb{E}(\theta) \text{ and } \Phi(1) = 1 - \frac{1}{2} [\mathbb{E}(\theta) + \mathbb{E}(\theta)] = 1 - \mathbb{E}(\theta)$$

and as  $\mathbb{E}(\theta) \in (0, 1)$ , the intermediate value theorem tells us that at least one equilibrium exists.  $\square$

**Proposition 6.** *There exists a persuasive equilibrium which is characterized by two unique thresholds  $\hat{\theta}^L$  and  $\hat{\theta}^H$ , with  $\hat{\theta}^L < \mathbb{E}(\theta_{post}) < \hat{\theta}^H$ . The up-advisor always sends  $\hat{\theta}_{post}^H$  and the down-advisor always sends  $\hat{\theta}_{post}^L$ . The aligned advisor sends  $\hat{\theta}^H$  if  $\theta_{post} \geq \hat{\theta}^H$  and sends  $\hat{\theta}^L$  if  $\theta_{post} \leq \hat{\theta}^L$ . If  $\theta_{post} \in (\hat{\theta}^L, \hat{\theta}^H)$ , the aligned advisor sends  $\theta_{post}$ . Upon receiving any message  $m \in [\hat{\theta}^L, \hat{\theta}^H]$  the investor's assessment is  $m$ . Upon receiving a message  $m \notin [\hat{\theta}^L, \hat{\theta}^H]$ , the investor's assessment is  $\mathbb{E}(\theta_{post})$ .*

*Proof.* Given the investor's strategy, it is optimal for both types of misaligned advisors to send the message which induces the lowest or highest possible assessment. The aligned advisor's strategy is also optimal: it induces the true assessment whenever the true state is between both thresholds and conditional on the true state not being between both thresholds, it is optimal for the aligned advisor to send the message which induces the lowest or highest assessment. As messages only fall within both thresholds if they are equal to the true state, it is optimal for the investor to adopt them. Upon receiving message  $\hat{\theta}^L$ , the investor's optimal response is

$$\theta_{post}^{I*L}(\hat{\theta}^L) = \mathbb{E}(\theta_{post}|\hat{\theta}^L) = p(\hat{\theta}^L)\mathbb{E}(\theta|\theta \leq \hat{\theta}^L) + (1 - p(\hat{\theta}^L))\mathbb{E}(\theta),$$

where  $p(\hat{\theta}^H) \equiv F(\hat{\theta}^H)/(1 + F(\hat{\theta}^H))$  is the probability that the message was sent by the aligned

advisor. This function maps from  $\Theta$  into  $\Theta$  and therefore must have at least one fixed point  $\theta_{post}^{I*L}(\hat{\theta}^L) = \hat{\theta}^L$ . Since  $0 < \theta_{post}^{I*L}(0) = \mathbb{E}(\theta_{post}) = \theta_{post}^{I*L}(1) < 1$ , the fixed point must be interior. To see that the fixed point is unique, take the derivative

$$\theta_{post}^{I*L'}(\hat{\theta}) = p'(\hat{\theta})[\mathbb{E}(\theta|\theta \leq \hat{\theta}) - \mathbb{E}(\theta)] + p(\hat{\theta}) \frac{\partial \mathbb{E}(\theta|\theta \leq \hat{\theta})}{\partial \hat{\theta}}.$$

Noting that

$$\frac{\partial \mathbb{E}(\theta|\theta \leq \hat{\theta})}{\partial \hat{\theta}} = \frac{f(\hat{\theta})}{F(\hat{\theta})} [\hat{\theta} - \mathbb{E}(\theta|\theta \leq \hat{\theta})] \text{ and } p'(\hat{\theta}) = -(1 - p(\hat{\theta}))^2 f(\hat{\theta})$$

we can plug in and rearrange to arrive at

$$\theta_{post}^{I*L'}(\hat{\theta}) = f(\hat{\theta})(1 - p(\hat{\theta}))[\hat{\theta} - \mathbb{E}(\theta|\theta \leq \hat{\theta}) - (1 - p(\hat{\theta}))(\mathbb{E}(\theta) - \mathbb{E}(\theta|\theta \leq \hat{\theta}))].$$

It is straightforward to verify that  $\theta_{post}^{I*L'}(0) < 0$  and  $\theta_{post}^{I*L'}(1) > 0$ . The second bracket term,  $(1 - p(\hat{\theta}))(\mathbb{E}(\theta|\theta \leq \hat{\theta}) - \mathbb{E}(\theta))$ , is decreasing in  $\hat{\theta}$  while the first bracket term,  $\hat{\theta} - \mathbb{E}(\theta|\theta \leq \hat{\theta})$ , increases if  $\frac{\partial \mathbb{E}(\theta|\theta \leq \hat{\theta})}{\partial \hat{\theta}} < 1$ . This is the case, as  $f(\theta)$  is a mixture distribution of different beta distributions: For a mixture distribution where the conditional expectations of the individual components are  $\mathbb{E}_1(\theta|\theta \leq \hat{\theta})$ ,  $\mathbb{E}_2(\theta|\theta \leq \hat{\theta})$ , ..., and where the density functions are weighted by  $w_1, w_2, \dots$  we have

$$\mathbb{E}(\theta|\theta \leq \hat{\theta}) = \sum_i w_i \mathbb{E}_i(\theta|\theta \leq \hat{\theta}) \Rightarrow \frac{\partial \mathbb{E}(\theta|\theta \leq \hat{\theta})}{\partial \hat{\theta}} = \sum_i w_i \frac{\partial \mathbb{E}_i(\theta|\theta \leq \hat{\theta})}{\partial \hat{\theta}}.$$

As the beta distribution belongs to the family of log-concave distributions,  $\frac{\partial \mathbb{E}_i(\theta|\theta \leq \hat{\theta})}{\partial \hat{\theta}} < 1$ ,<sup>35</sup> which implies  $\frac{\partial \mathbb{E}(\theta|\theta \leq \hat{\theta})}{\partial \hat{\theta}} < 1$ . Therefore,  $\theta_{post}^{I*L'}(\hat{\theta})$  switches its sign only once from negative to positive, i.e., it is quasiconvex. This implies the existence of a unique fixed point  $\hat{\theta}^L < \mathbb{E}(\theta_{post})$ .

Similarly, when receiving message  $\hat{\theta}^L$ , the investor's optimal response is

$$\theta_{post}^{I*H}(\hat{\theta}^H) = \mathbb{E}(\theta_{post}|\hat{\theta}^H) = q(\hat{\theta}^H)\mathbb{E}(\theta|\theta \geq \hat{\theta}^H) + (1 - q(\hat{\theta}^H))\mathbb{E}(\theta),$$

$q(\hat{\theta}) \equiv (1 - F(\hat{\theta})) / (2 - F(\hat{\theta}))$  is the probability that the message  $m''$  was sent by the aligned advisor. For this function,  $0 < \theta_{post}^{I*H}(0) = \mathbb{E}(\theta_{post}) = \theta_{post}^{I*H}(1) < 1$ , suggesting that at least one critical threshold exists. Analogous steps as we have taken before show that  $\theta_{post}^{I*H'}(0) > 0$  and  $\theta_{post}^{I*H'}(1) < 0$  and that  $\theta_{post}^{I*H}(\hat{\theta})$  is quasiconcave. These properties ensure a unique fixed point  $\hat{\theta}^H > \mathbb{E}(\theta_{post})$ . □

Among the two persuasive equilibria the second is more informative. It restricts influential communication to an interval around the investor's prior expectation. As the utilities of the mis-

<sup>35</sup>See, e.g., Lemma 1 in Harbaugh and Rasmusen (2018).

aligned advisors are state-independent, they will always send the messages at the boundaries of the interval.

**Remark: uniqueness** The second equilibrium is essentially unique: For any historical data set the thresholds which determine the bounds of the interval are unique. Therefore, every equilibrium of the second type leads to the same economic allocations. However, different strategies can typically support the same equilibrium allocations. First, there is the off-equilibrium threat about the action that the investor will take when hearing a message outside the interval. Any off-equilibrium action that is between both thresholds supports the described equilibrium. Second, in the experiment the advisor's message space is actually larger as it also comprises of  $\theta_{pre}$  and  $c$ . These parameters are payoff-irrelevant for all agents and therefore no persuasive communication about them will be possible. However, they might matter as off-equilibrium threats. For example, the investor might have the strategy to only adopt a message if  $\theta_{post}^A$  is on the inside of the interval and  $c^A$  and  $\theta_{pre}^A$  maximize the likelihood function conditional on  $\theta_{post}^A$ . Then there is an equilibrium where advisors always send these conditional likelihood maximizers in their message. However, in the cheap talk game this is a matter of equilibrium selection as off path threats can rationalize any strategy advisors might have over the additional message components. No matter how agents select among the different equilibria, the resulting allocations remain identical.

**Remark: second-order uncertainty about advisor type** In the experiment, the advisor is told that the investor “*may or may not*” know the advisor's type. Therefore, the advisor knows that there are two possible worlds. In world one, the investor knows the advisor's type. A misaligned advisor knows that the investor will not follow the message in this world because, as the misaligned advisor's preferences are state-independent, there is no persuasive equilibrium. We described persuasive equilibria in the world two where the investor does not know the advisor's incentives. Suppose that agents here play the second equilibrium. If the misaligned advisor now has to settle for a message strategy that can potentially be useful in both possible worlds, it is optimal for to follow the strategy described before. In the worst case the investor will never follow the advisor's message (world one). In the best case (world two) the investor will follow it and given the investor's strategy in the described equilibrium there is no better message that the misaligned advisor can send. Therefore, adding the uncertainty about the investor's beliefs about the advisor does not change the misaligned advisor's behavior in equilibrium. Similarly, the investor's strategy when he knows that he is matched to an aligned advisor remains a best reply to the aligned advisor's strategy in the equilibrium described before. Conversely, the aligned advisor has no incentive to deviate from her equilibrium strategy once we introduce second-order uncertainty about types.

## F Instructions for the investor in BASELINE

### Welcome to our experiment!

This experiment will take approximately **25 minutes** to complete. It is divided into **10 rounds**. In each round, the computer will randomly match you with one other player. You and the other players that you are matched with will remain anonymous.

You will receive a show-up fee of **£3.50** for participating in the experiment. You can also earn a **bonus** payment of **£3.75** during the experiment. The amount that you earn will depend on the decisions made by you and other participants during the experiment. It is therefore important that you read the instructions carefully as this will help you to make better choices. In addition, there will be a set of understanding questions to check that you read and understood the instructions properly. You will need to answer these questions correctly in order to complete the experiment.

At the end of the experiment, one of the 10 rounds will be randomly chosen to be relevant for your payment. The decisions made by you and your matched partner in the chosen round will determine your payment.

Next



## Your task as an investor

There are two types of roles in this experiment: **investors** and **advisors**. You have been randomly chosen to be an **investor** throughout the whole experiment.

In each round, you will be randomly matched with an **advisor**, who is another participant in the experiment. The advisor will send you a message. After receiving the message, you will make a decision.

As an investor, in each round your task is to evaluate a hypothetical company and to assess how likely it is that the company will be successful (i.e., profitable) in the coming year. Since there are ten rounds, you will evaluate ten companies labelled Company A, Company B, Company C, ..., Company J.

Each of the companies that you will evaluate produces a fictitious product called a Widget. Widgets are a type of technological product. As the built-in technology advances very quickly, the product is outdated after about a year. At the start of each year, the company, therefore, discontinues the previous year's model and releases a new model of the Widget. In every year, the company's success depends on the model produced in that year.

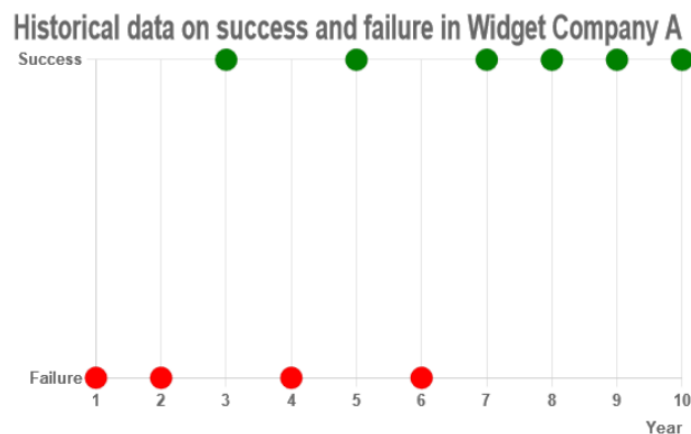
To evaluate how likely it is that the Widget company will be successful with the model it will produce in the coming year, you are given several pieces of information.

### THE PIECES OF INFORMATION ARE:

#### 1. Historical data

First, you will be shown the historical data for the company, showing whether it was "successful" or "not successful" with the models it produced in each of the past ten years (i.e., from Year 1 to Year 10).

For example, the data could look like this:



Here, you can see that Widget Company A was successful with the model it produced in six out of the past ten years. This historical data is **public**. So both you and your matched **advisor** have access to it.

#### 2. The Widget company's probability of success depends on its CEO

Second, you know that the CEO of a Widget company determines the probability of success in each year. You also know that the CEO of each company that you're evaluating changed exactly once during the ten years. This occurred at the **start of Year 3, 4, 5, 6, 7, 8 or 9**, but you do not know exactly which year. Under a specific CEO the probability of success is the same in every year.

(screen continues on next page)

This means that when evaluating each company, there are three important things to consider:

- a. The **year** in which the CEO changed.
- b. What the company's **initial** probability of success was, in each year *before the change of the CEO*.
- c. What the company's **current** probability of success is, in each year *after the change of the CEO*.

All companies are completely independent of one another. So these features differ between companies.

### 3. Advice received from your matched advisor

Third, in each round, you are matched with an **advisor** who knows the truth about the following information: the **year** in which the CEO changed, what the company's **initial** percentage probability of success (Initial PoS%) was and what its **current** percentage probability of success (Current PoS%) is. In other words, the advisor is fully informed about (a), (b) and (c).

The advisor's task is to send you a **message** before you submit your evaluation of the company's **current** percentage probability of success. The message will contain the advisor's assessment of the **year** in which the CEO changed, as well as their assessment of the company's **initial** percentage probability of success and **current** percentage probability of success. For example, an advisor's message to you could look like this:

Your advisor in this round says that the CEO of Widget Company A changed at the **start of Year 5**. They say that **71** was Widget Company A's **initial** percentage probability of success. They say that **45** is Widget Company A's **current** percentage probability of success.

**Important information:** Each advisor is free to choose the content of the message that they send to you; they do not need to truthfully report the information that they have to you.

Next

## Details about how your payment is calculated

### *Your incentives*

In each round, your task as an investor is to estimate the company's **current** percentage probability of success (Current PoS%) **as accurately as possible**. Your payment will, therefore, depend on how close your estimate of the Current PoS% is to the true Current PoS%. The closer your estimate is to the truth, the more likely it is that you will win the **bonus of £3.75**. Therefore, it is in your best interest to estimate the Current PoS% as accurately as possible in each round.

If you would like to see the formula that explains exactly how your payment is calculated, you can click on the following button:

[Click here to see the formula](#)

### *Advisor's incentives*

Your advisor's earnings will also depend on your estimate of the company's **Current** PoS%. They know that you want to estimate the likely success of the company as accurately as possible.

You will face advisors with various incentives.

Recall that the experiment consists of 10 rounds. In the course of these 10 rounds, you can meet three different types of advisor:

- ↑ • The ↑ advisor wants their matched investor to believe that the company has a **high Current PoS%**—i.e., the ↑ advisor's payment is **higher** when the investor estimates that the company is *more likely to succeed*.
- ↓ • The ↓ advisor wants their matched investor to believe that the company has a **low Current PoS%**—i.e., the ↓ advisor's payment is **higher** when the investor estimates that the company is *less likely to succeed*.
- • The → advisor wants their matched investor to hold **accurate** beliefs about the **Current PoS%** of the company—i.e., the → advisor's payment is **higher** when the investor's estimate of the company's likely success *becomes more accurate*.

Advisors are told that you, the investor, "may or may not know the advisor's incentives".

[Previous](#)

[Next](#)

Details about how you will be incentivized

Your incentives

In each round, your task as an investor is to estimate the company's current percentage probability of success (Current PoS%) **accurately as possible**. Your payoff is determined by how close your estimate is to the true Current PoS%. The closer your estimate is to the true Current PoS%, the higher your payoff. It is in your best interest to estimate the Current PoS% as accurately as possible.

If you would like to see the formula for calculating your payoff, click on the following button:

Advisor's incentives

Your advisor's earnings will also depend on the company's current percentage probability of success as accurately as possible.

You will face advisors with various levels of expertise. Recall that the experiment consists of two parts:

↑

- The ↑ advisor will estimate the company's current percentage probability of success as accurately as possible. The ↑ advisor's payment is **high** if the company's current percentage probability of success is **high**.

↓

- The ↓ advisor will estimate the company's current percentage probability of success as accurately as possible. The ↓ advisor's payment is **high** if the company's current percentage probability of success is **low**.

→

- The → advisor will estimate the company's current percentage probability of success as accurately as possible. The → advisor's payment is **high** if the company's current percentage probability of success is **close to the true value**.

Advisors are told that you, the investor, may or may not know the advisor's incentives.

Explanation of the formula for calculating your payment



Your estimate of the company's current percentage probability of success (Current PoS%) is a number between 0 and 100. This number is used to determine your payoff according to the following formula:

$$\text{Probability of winning the bonus (in percent)} = 100 - \frac{d^2}{100},$$

where  $d$  is the **difference** between your estimate and the company's true current percentage probability of success. The formula squares this difference and divides it by a constant. This number is then subtracted from 100. Therefore, if you estimate 50 and the true value is 70, then you win the bonus 96 percent of the time, because the difference is 20 and  $100 - \frac{20^2}{100} = 96$ .

The principle underlying the above formula is simple: the closer your estimate is to the true value, the higher the percentage probability that you win the bonus of **£3.75**. Note that your estimate can be wrong by at most 100. In this case, the formula shows that your probability of winning is 0 percent.

At the end of the experiment, the computer will randomly choose one round of the experiment to determine whether or not you will win the bonus.

success (Current PoS%) **as accurately as possible**. Your payoff is determined by how close your estimate is to the true Current PoS%. The closer your estimate is to the true Current PoS%, the higher your payoff. It is in your best interest to estimate the Current PoS% as accurately as possible.

on the following button:

that you want to estimate the likely

different types of advisor:

**↑ advisor's** current PoS%—i.e., the ↑ advisor's current percentage probability of success **succeed**.

**↓ advisor's** current PoS%—i.e., the ↓ advisor's current percentage probability of success **succeed**.

**→ advisor's** current PoS% of the company—i.e., the → advisor's current percentage probability of success **becomes more accurate**.

Close

Previous

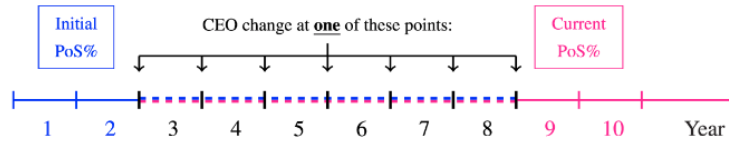
Next

## More details on what determines success and failure in each company

The details below will help you to estimate each company's current percentage probability of success as accurately as possible.

### (a) When did the CEO change?

Recall that the CEO of each company changed once and for all at the **start of Year 3, 4, 5, 6, 7, 8 or 9**. For each company, the year in which the CEO changed will be **randomly determined** by the computer. So each of these seven years has an equal probability of being chosen.



**Example** Suppose that the start of Year 3 is randomly chosen by the computer for a particular company as the moment when the company's CEO changed. This means that the Initial PoS% is relevant for Years 1 and 2 while the Current PoS% is relevant for Years 3 to 10.



**Hint** Success in Years 1 and 2 is always determined by the Initial PoS%. Success in Years 9 and 10 is always determined by the Current PoS%.

### (b) How is a company's Initial PoS% determined?

For each company, the computer will **randomly** draw a whole number between 0 and 100. Each whole number is equally likely to be drawn. This number determines the company's Initial PoS%.

### (c) How is a company's Current PoS% determined?

Similarly, to determine each company's Current PoS%, the computer will **randomly** draw a second whole number between 0 and 100 (i.e., each whole number is equally likely to be drawn).

**Important information:** The company's Current PoS% is completely **independent** of its Initial PoS%. This means that, no matter what the company's initial percentage probability of success was, any number between 0 and 100 is equally likely to be drawn as its current percentage probability of success. Intuitively, the quality of the company's initial CEO does not tell you anything about how good its current CEO is.

Also, there is no relationship between companies. So, the Initial PoS% and Current PoS% of each of the companies is completely unrelated to all other companies.

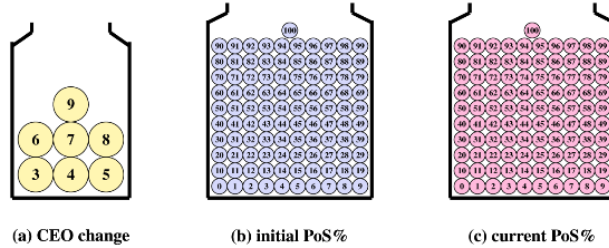
Previous

Next

## How does the CEO affect the company's success in every year?

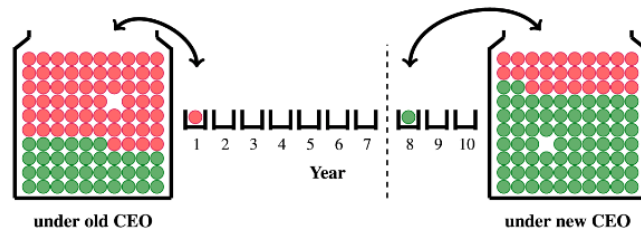
You can think of the computer going through a **two-step process** prior to each round.

In **Step 1**, the computer draws one ball at random from each of the following three urns (i.e., three balls in total). These three ball draws determine the year in which the CEO changed, the **initial** percentage probability of success (Initial PoS%), and the **current** percentage probability of success (Current PoS%).



Say, for example, that 8, 36, and 72 are drawn. This means that the CEO changed at the **start of Year 8**. The new CEO is in charge in Years 8, 9 and 10. Compared to the old CEO, this new CEO also turns out to be quite good: compared to a **36%** probability of success in Years 1 to 7 under the old CEO, the company has a **72%** probability of success in every year when the new CEO is in charge.

In **Step 2**, the computer determines success and failure for each single year. To do so, the computer draws from an urn with 100 balls, which are either green or red. The number **36** determines the quantity of green balls in a company's urn under the old CEO; the number **72** determines the quantity of green balls in the urn under the new CEO. The computer draws a ball at random from the relevant urn for each of the years that a CEO is in charge. If the ball drawn is **green**, then the company is **successful** in that year. If the ball drawn is **red**, then the company **fails** in that year. After each draw, the computer places the ball back into the urn before making a draw for the next year. This means that, in each period, success and failure are only determined by the percentage probability of success and do not depend on success or failure in earlier periods.



Previous

Next

## How does the matching of investors and advisors work?

At the beginning of the experiment, you will be randomly allocated to a group containing six participants – three investors and three advisors. In every round of the experiment, the advisors and investors are randomly re-matched into three pairs. This means that in each round you could be matched with any one of the three advisors in your group.

Your group includes one advisor of each type. In particular:

- One advisor in your group is the ↑ advisor, who has an interest in their matched investor believing that the company has a **high Current PoS%**.
- One advisor in your group is the ↓ advisor, who has an interest in their matched investor believing that the company has a **low Current PoS%**.
- One advisor in your group is the → advisor, who has an interest in their matched investor to hold **accurate** beliefs about the **Current PoS%** of the company.

[Previous](#)[Next](#)

## Overview of the sequence followed in the experiment

The experiment will consist of ten rounds.

Each round consists of the same five steps:

1. You are matched randomly with an advisor.
2. The advisor receives accurate information about the year in which the CEO changed, the initial PoS% and the current PoS%. The advisor also observes the public dataset that shows the past performance of the company (i.e., whether the company succeeded or failed in each year).
3. The advisor chooses the message that they send to you. The message will contain an assessment about the year in which the CEO changed, the initial PoS% and the current PoS%.
4. You see the following pieces of information:
  - i. the public dataset that shows the past performance of the company, and
  - ii. the advisor's message.
5. You submit your estimate of the Current PoS%.

[Previous](#)[Next](#)

## Understanding questions


Please answer the following questions to make sure that you understand the experimental instructions.

You can use the navigation bar above to quickly access specific screens of the instructions.


How many advisors are there in your group?


How many investors are there in your group?

Are you an advisor or an investor?

How many different companies will each investor evaluate?

Will a specific investor be matched with the same advisor in every round?

Is it possible that in one of the companies the CEO did not change during the last ten years?

Relation between the initial CEO on the current CEO

If a company was very successful before the CEO changed, does this mean that it is more likely that it was successful after the CEO changed?

After the CEO changed in a company, is the probability of success the same in every year after that?

What is the probability that a company is successful in Year 10?

(screen continues on next page)



Advisor knowledge

Do advisors always know the true date at which the CEO changed and the true probability of the company succeeding in every year?

Which of the following statements is correct?

Which of the following statements is correct?

# Make your assessment — Round 1

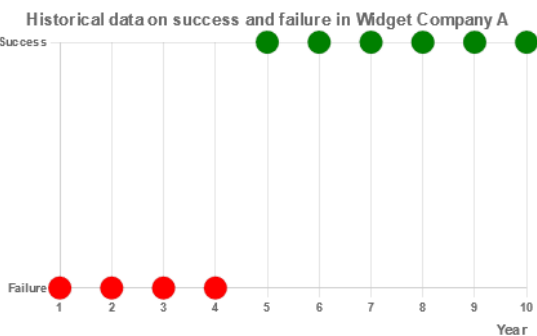
In this round, you will assess Widget Company A. When making the assessment, you can refer to a message from your advisor for this round.

When composing the message, your advisor had access to:

- The historical data of success and failure in Widget Company A and
- Information about the **year** in which the CEO changed, the company's **Initial PoS%**, and the company's **Current PoS%**.

You can also use the historical data to inform your assessment.

## YOUR INFORMATION



### Message from advisor:

Your advisor in this round says that the CEO of Widget Company A changed at the **start of Year 5**. They say that **36** was Widget Company A's **initial** percentage probability of success. They say that **95** is Widget Company A's **current** percentage probability of success.

Year of change	Initial PoS%	Current PoS%
5	36	95

What is your assessment of the Current PoS% of Widget Company A?

Current PoS%

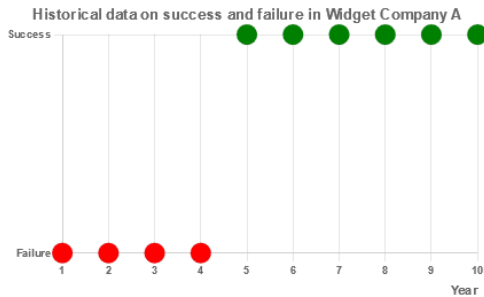
91

Next

## Confirm your decision



Based on the following information, you evaluate Widget Company A to have a Current PoS% of **91**.



### Message from advisor:

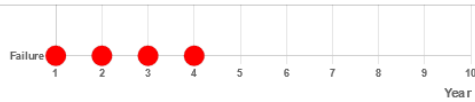
Your advisor in this round says that the CEO of Widget Company A changed at the **start of Year 5**. They say that **36** was Widget Company A's **initial** percentage probability of success. They say that **95** is Widget Company A's **current** percentage probability of success.

Year of change	Initial PoS%	Current PoS%
5	36	95

Please confirm your decision below.

I confirm

Close



Year of change	Initial PoS%	Current PoS%
5	36	95

What is your assessment of the Current PoS% of Widget Company A?

Current PoS%

91

Next

# Survey

Before you are finished with today's experiment, we ask that you please answer a few more questions:

Age

How old are you?

Gender

Please select the gender that you identify with.

-----

▼

Education

What is the highest level of education you have completed?

-----

▼

How did you decide?

Please briefly explain how you took your decisions in the study.

Comments

Do you have any comments about the study? Or anything you'd like to tell us?

Next