Fries, Tilman; Barron, Kai

Conference Paper

Narrative Persuasion

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage"

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Fries, Tilman; Barron, Kai (2023) : Narrative Persuasion, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage", ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/277691

Narrative Persuasion*

Kai Barron
WZB Berlin

Tilman Fries
WZB Berlin

March 1, 2023

For the current version, click here.

Abstract

Modern life offers nearly unbridled access to information; it is the harnessing of this information to guide decision-making that presents a challenge. We study how one individual may try to shape the way another person interprets objective information by proposing a causal explanation (or narrative) that makes sense of this objective information. Using an experiment, we examine the use of narratives as a persuasive tool in the context of financial advice where advisors may hold incentives that differ from those of the individuals they are advising. Our results reveal several insights about the underlying mechanisms that govern narrative persuasion. First, we show that advisors construct self-interested narratives and make them persuasive by tailoring them to fit the objective information. Second, we demonstrate that advisors are able to shift investors' beliefs about the future performance of a company. Third, we identify the types of narratives that investors find convincing, namely those that fit the objective information well. 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, financial advice, conflicts of interest, behavioral finance.

*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, Valeria Burdea, Daniele Caliari, Constantin Charles, Felix Chopra, Dirk Engelmann, Nicola Gennaioli, Katrin Gödker, Thomas Graeber, Jeanne Hagenbach, Luca Henkel, Emeric Henry, Alessandro Ispano, Agne Kajackaite, Chad Kendall, Anita Kopányi-Peuker, Dorothea Kübler, Christine Laudenbach, Yves Le Yaouanq, Yiming Liu, George Loewenstein, Salvatore Nunnari, Davide Pace, Chris Roth, Joshua Schwartzstein, Paul Seabright, Alice Solda, Heidi Thysen, Joël van der Weele 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.[1] 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 used for transmitting ideas about how the world works. Individuals share them using simple stories, metaphors or anecdotes via word-of-mouth or on social media (Shiller, 2017). Importantly, narratives may also be used by individuals with a vested interest to try to shape how others interpret events.[2] This implies that narratives may be used as a persuasive tool where one individual tries to influence how another person draws inference from objective information.

By focusing on the interpretation of objective information, narrative persuasion differs from other much-studied forms of persuasion, such as disclosure games (e.g., Milgrom, 1981), where one individual may disclose truthful information to another, cheap-talk (e.g., Crawford & Sobel, 1982), where an informed party can send a non-verifiable message to an uninformed party, and Bayesian persuasion (e.g., Kamenica & Gentzkow, 2011), where one individual can construct a data-generating process for another. While examples of individuals using narratives to persuade are ubiquitous in everyday life, empirical evidence on the mechanisms that govern the construction and effectiveness of such narratives is scarce.[3]

This paper studies such narrative persuasion in the context of financial advice. We consider a setting in which financial advisors may try to influence investors' beliefs by proposing narratives to explain the available objective data. A major challenge for studying narratives is that, in field settings, narratives may take a diverse array of forms and typically interact with an individual's existing information set (which is typically endogenous and not fully observable to the analyst). We circumvent these issues by designing a financial advice experiment that

[1] 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 show how most of them use the term *narrative* to describe *a causal explanation that makes sense of a collection of events*. This is the working definition that we use in this paper.

[2] 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.

[3] Examples of 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; *Immigration policy:* people circulate stories about the impact of immigrants on crime rates and unemployment levels; *Academia:* different academics propose models to organize the available empirical evidence. The importance of narratives also extends to the *law*, where both sides generally build their case around the same body of evidence (see Pennington & 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).

allows us to study the underlying mechanisms governing narrative construction and adoption. The experimental design provides us with a setting in which we have full control over the decision environments of both advisors and investors. We use this control to exogenously vary the content of these decision environments (e.g., the information sets) and study how this influences the narratives that advisors send to investors. We then analyze how these narratives affect the beliefs that investors form. We are therefore able to identify the causal effect of narratives on investors' beliefs. We are also able to exploit the exogenous variation generated in our experiment to better understand how advisors construct their narratives and which types of narratives are most persuasive.

Drawing inspiration from the theoretical work of Schwartzstein & 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 a historical data set representing a hypothetical company's performance over a span of ten years. 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 historical performance 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 for the company's performance during the observed historical period. This narrative may guide how the investor interprets the past data and influence the beliefs he forms about the future. 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.

What exactly do advisors and investors know? For each of the ten years included in the historical company data, both individuals observe the outcome of a binary variable that reflects the performance of the company. In years where the company was profitable, the binary variable outcome is "success", and in years where the company was unprofitable, the binary variable outcome is "failure". Both advisors and investors know that a simple model generated this data. Specifically, they know that the company had exactly one major structural change during this period—in the experiment, we frame this as a change of the company's CEO. Prior to the CEO change, the probability that the company was successful in each year was constant. When the CEO changed, the company shifted to a new probability of being successful in each year. The investor does not know exactly in which year the CEO changed nor does he know the company's probability of success under the old CEO or the new CEO; he only knows that the probability was constant before and after the CEO change.

In contrast to the investor, the advisor knows all of this information—she knows when the CEO changed and the probability of success under the old and new CEO. The advisor's task is to send a message to the investor which consists of three parameters; the company's success probability under the old CEO, the company's success probability under the new CEO, and the year in which the CEO changed. She is not restricted to telling the truth in this message. The

message provides the investor with an explanation that makes sense of the company's past performance. Specifically, by providing a statement about the year of the CEO change and the success probability under the old CEO, the message not only predicts how the company will fare in the future, but also explains why the company was either more or less successful in certain years in the past— it is therefore a narrative.

After receiving the advisor's 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 perfectly aligned with investors and are incentivized to induce accurate beliefs in their matched investors. When making the assessment, the investor does not know what type of advisor sent him the message.

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

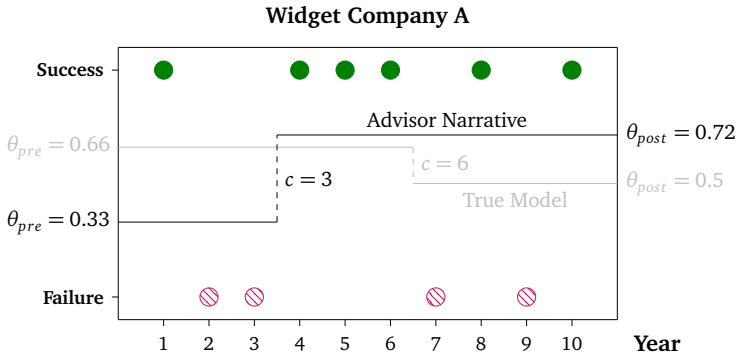


Figure 1 illustrates the basic intuition of narrative persuasion in our experiment. The solid green and the hatched red dots show an example of the historical company data. The grey line indicates an example of a possible true underlying model observed by the advisor, and the black line illustrates a potential narrative that an up-advisor might use to try to persuade an investor to hold an upward biased belief about the probability of success under the new CEO, θ_{post}. Importantly, this example highlights a central feature of narrative persuasion. While the up-advisor only cares about moving the investor's belief about θ_{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, c, from year 6 to year 3 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. To improve the fit of her narrative to the data, the advisor, therefore, correspondingly shifts her assessment of the company's probability of success under the old CEO downwards.

We present four sets of results. 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, the average advisor engages in a fairly sophisticated form of self-interested narrative persuasion in which they do not 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 make their narrative more convincing. We also provide results from an individual-level analysis which suggests that one third of misaligned advisors can be classified as "frequent opportunists" who exploit the opportunities that are presented to them in the form of particular realizations of the data to construct narratives that promote their self-interest. In order to evaluate the properties of narratives sent by different advisors, we construct an index that orders narratives according to their empirical fit, conditional on the historical data. This index provides a metric of how likely it is that the narrative under consideration generated the observed data. Using this index, we find that advisors classified as frequent opportunists construct narratives which achieve an empirical fit that is close to that of the true data generating process. This is striking since they manage to achieve this apparent fidelity of the narrative to the data even though, amongst all advisor types, they provide the most misleading predictions about the company's future success.

Second, turning to investors, we show that misaligned advisors are successful in shifting the beliefs of investors—investors that meet a misaligned 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. We also show that the distance between the investor's assessment and the truth increases in the advisor's degree of opportunism. This means that investors are most mislead by those advisors who frequently tailor their narrative to the data in order to achieve the twin goals of sending a narrative that looks plausible and whose prediction is in line with the advisor's objective.

Third, we investigate the properties of narratives that make them more convincing and find that narratives with a higher empirical fit are more persuasive. To show this, we first document that there is a correlation between the fit of the advisor's narrative and how close the investor's assessment is to the advisor's suggested narrative. To establish causality, we use additional data on investors' prior beliefs about the company that they report before receiving an advisor's narrative but after seeing the company data. Controlling for these prior beliefs, we find that investors update more towards an advisor's narrative when the advisor's narrative fits better. This suggests that investors do indeed try to assess the veracity of a narrative by comparing it to the objective data. They update more when they view a narrative as credible, given the data. It also provides an explanation for why investors are most mislead by frequent opportunist advisors; even though they provide misleading predictions, the empirical fit of their narratives are comparable to those of the true data generating process.

Fourth, we evaluate the impact of three policy interventions that could potentially protect individuals from being harmed by narrative persuasion. We introduce three treatment conditions, with each treatment corresponding to one candidate policy intervention 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 intervention raises the salience of possible alternative explanations of the data, and investors who have conducted a careful assessment of the objective data themselves may be less willing to believe received 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.

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 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.[4] 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 meaning that they disregard highly informative messages. In these scenarios, skepticism harms the investors. On average, these more skeptical investors, therefore, do no better in the DISCLOSURE treatment. Overall, the findings from the policy interventions indicate that narrative persuasion is difficult to protect against. Being able to compare a message to objective data can add credibility to the message—when a narrative fits objective data well, it can seem compelling.

To obtain a deeper understanding about how investors update their beliefs in response to different narratives, we present additional structural estimates. Our estimation results suggest that advisors can shift the probability of the investor adopting the narrative by around 25 percentage points by increasing the narrative fit. The estimates also highlight that, holding narrative fit fixed, investors in INVESTORPRIOR and PRIVATEDATA are more skeptical to narratives,

[4]It is well-documented that a significant proportion of individuals have preferences for truth-telling. For an early experiment using a lab sample, see Fischbacher & Föllmi-Heusi (2013). Abeler, Becker, & Falk (2014) provide related evidence in a representative sample of the German population. Theoretical models that combine an intrinsic preference for telling the truth and an extrinsic preference for being perceived as being truthful can describe much of the experimental data provided by lying experiments (Gneezy, Kajackaite, & Sobel, 2018; Abeler, Nosenzo, & Raymond, 2019).

compared to BASELINE.

Taken together these results are broadly in line with the predictions and assumptions of S&S. Our finding that investors take the empirical fit of a narrative into account when judging whether to believe the narrative supports the idea developed in S&S that investor's narrative adoption is guided by features of the message relative to the data. This is in contrast to two natural alternative benchmarks for investor behavior. First, investors may simply ignore the messages they receive from advisors, and instead rely on their own introspection. Second, investors could engage in sophisticated strategic thinking when interpreting the messages they receive from advisors. In relation to the first benchmark, our results show clearly that investors do not ignore the messages they receive; rather, they have their beliefs meaningfully shifted in the direction the advisor wishes to bias them. In relation to the second, the evidence that we provide in DISCLOSURE on the relative insensitivity of investors to the revelation of their advisor's incentives suggests that investors only engage in limited strategic thinking when evaluating narratives. However, to provide more rigorous evidence regarding the strategic reasoning benchmark, we consider the pattern of behavior one might expect under the Nash Equilibrium.

To do this, we take advantage of the fact that 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, provide a theoretical analysis of our setting using a cheap talk model. This enables us to contrast the the predictions of this more standard model with those prescribed by the S&S framework. The empirical data that we collect in our experiment are difficult to explain with the cheap talk model predictions. 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 is in line with the idea that individuals sometimes neglect to fully engage in sophisticated strategic thinking. To reduce the complexity of the problem, they instead reformulate it as a slightly different problem that they can solve more easily to arrive at their decision-relevant beliefs. This is consistent with an array of well-documented behavioral biases including cursedness (Eyster & Rabin, 2005), selection neglect (Jehiel, 2018; Barron, Huck, & Jehiel, 2019; Enke, 2020), and correlation neglect (Eyster & Weizsacker, 2016; Enke & Zimmermann, 2019; Laudenbach, Ungeheuer, & 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 in favor of 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, & Motes, 2011), and narrative analysis in literary and cultural studies (Koschorke, 2018; Herman & 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 argued 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, & Strack, 2018; Bénabou, Falk, & Tirole, 2020; Eliaz & Spiegler, 2020; Spiegler, 2020a,b; Mailath & Samuelson, 2020; Schwartzstein & Sunderam, 2021; Aina, 2021; Olea, Ortoleva, Pai, & Prat, 2021; Schumacher & 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.[5]

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. To generate predictions for our experiment, we draw on a framework developed by Schwartzstein & Sunderam (2021), who model a narrative as a likelihood function sent by a persuader and which a receiver evaluates by comparing it to data. The key assumption in S&S is that the receiver will adopt a narrative if it explains the data sufficiently well. This assumption then gives rise to a tradeoff for the persuader, who, when choosing a narrative, must strike a balance between *fit*, the narrative's coherence with the data, and *movement*, and the degree to which adopting the narrative will move the receiver's belief. A model with a similar recipient decision rule has been studied previously by Froeb, Ganglmair, & Tschantz (2016) in a legal context. The authors conduct a theoretical investigation of a setting in which a court has to decide whether to rule in favor of a plaintiff or defendant. A key component of their analysis is that the court's final ruling can be swayed by the plaintiff's and defendant's interpretation of evidence.

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

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 payoffs of both (Crawford & 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, & Sprinkle, 1998; Blume, DeJong, Neumann, & Savin, 2002; Wang, Spezio, & Camerer, 2010) and also on communication with evidence in the form of disclosure games (see, e.g., King & Wallin, 1991; Hagenbach & Perez-Richet, 2018; Jin, Luca, & 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 about the payoff-relevant parameter but also about the non payoff-relevant parameters. Importantly, advisors may design their messages such that they choose to communicate payoff-irrelevant parameters that justify the communicated payoff-relevant parameters. Specifically, advisors may use the payoff-irrelevant parameters to construct a better overall fit of the message to the data to make the message more convincing. In contrast, in a cheap talk framework, sending these additional non payoff relevant parameters does not matter since strategic considerations make it impossible to achieve informative communication on non payoff-relevant domains. We discuss the relationship between the narrative persuasion theoretical framework and the sender-receiver theoretical approach in detail in Section 3.4. 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, & Moore, 2005; Malmendier & Shanthikumar, 2007; Loewenstein, Sah, & Cain, 2012; Sah, Loewenstein, & Cain, 2013).[6] This literature delineates several mechanisms through which disclosure may by undermined. 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. Receivers then 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 by reduced by the fact that they make investors equally skeptical of advice received from both honest and dishonest advisors.

Finally, our work is relates to a very recent empirical literature in economics that explores how narratives (broadly construed) shape behavior. For example, Andre, Pizzinelli, et al. (2022) study households' subjective beliefs about the responsiveness of key economic variables

[6]It is worth noting that disclosure is not always ineffective. One example of a context in which conflict of interest disclosures may be effective is provided by Sah & 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.

to macroeconomic shocks and Andre, Haaland, 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, Harrs 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 & Verrina (2022) also show that stories can be used to influence prosocial behavior. Graeber, Zimmermann, & Roth (2022) explore the relationship between stories and memory, showing that information embedded in a story has a slower memory decay rate than statistics presented in the absence of a story-context. Finally, Morag & 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. Second, while some contributions in this literature conceptualize narratives in a broad sense, including stories and informal models, we focus on a particular conceptualization of a narrative as a subjective model explaining a particular process (Andre, Haaland, et al., 2022, and Charles & Kendall, 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 when an individual is deciding whether to adopt a particular narrative as an explanation for a given set of events, they may evaluate the narrative based on its veracity (fit) 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 provide a comparison of the two theoretical approaches.

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 θ_{pre}^T denotes the company's success probability in each of the years between 1 and c^T. Finally, the parameter θ_{post}^T denotes the company's success probability in the years $c^T + 1$ to 10. The true underlying model is thus given by $m^T = (c, \theta_{pre}, \theta_{post}) \in \mathcal{M} \equiv \{2, \ldots, 8\} \times [0,1]^2$. 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, m^I, about it. Based on that belief, the investor will make an assessment, θ_{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 is not informed by 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 some 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 vector that is distributed according to a density function $f(m)$ which has full support on

$\mathcal{M}$.[7]

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 θ_{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 $\theta^I_{post} = \theta^A_{post}$.

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;

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

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 is *data-optimal (DO)* in the sense that it is the narrative that best explains the data. An investor who adopts or rejects narratives according to Equation (1) will always adopt m^{DO} if he receives it as a message. For most histories, the data-optimal narrative is unique.[8] Holding c fixed, the θ_{pre} and θ_{post} parameter values of the data-optimal narrative are equal to the empirical proportion of successes in their respective period. Therefore, to find the global data-optimal narrative when we allow c to vary, we can simply compare the the log likelihood values obtained from $(c, \theta^{DO}_{pre}|c, \theta^{DO}_{post}|c)$ for each possible value of c, where $\theta^{DO}_{pre}|c$ and $\theta^{DO}_{post}|c$ are the data optimal values associated with a particular c. When comparing these different narratives, those with θ_{pre} and θ_{post} parameter values closer to either 0 or 1 dominate the other narratives in terms of empirical fit.[9] This nicely captures the following intuition: A narrative that partitions *pre* and *post* in such a way that, within both periods, the company is either very successful or very unsuccessful will usually have a high log likelihood value. In this sense, narratives that more coherently explain success and failure in the data are more likely to be adopted.

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. The advisor's utility

[7] 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.

[8] It is not unique for degenerate histories like $h = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$, where the narrative $(c, 1, 1)$ is data-optimal for any c.

[9] Denote by $\pi_{pre}(c)$ and $\pi_{post}(c)$ the proportion of successes and failures in *pre* and *post* when the structural break is in year c. When comparing two narratives $m' = (c', \pi_{pre}(c'), \pi_{post}(c'))$ and $m'' = (c'', \pi_{pre}(c''), \pi_{post}(c''))$, c' will have a lower fit than c'' if $\min\{\pi_{pre}(c'), 1-\pi_{pre}(c')\} < \min\{\pi_{pre}(c''), 1-\pi_{pre}(c'')\}$ and $\min\{\pi_{post}(c'), 1-\pi_{post}(c')\} < \min\{\pi_{post}(c''), 1-\pi_{post}(c'')\}$.

depends on the investor's assessment, θ^I_{post}, and the advisor's bliss point, φ;

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

This utility is maximized if $\theta^I_{post} = \varphi$. The exact value of φ 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 $\varphi = \theta^T_{post}$. 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 θ^A_{post} if there is a close enough fit of m^A to the objective data. The probability of the investor adopting the advisor's model is then given by a c.d.f. $G(\ell(m^A))$ which is strictly increasing and continuous on $(-\infty, \ell(m^{DO})]$, with $G(\ell(m^{DO}) = 1$. We can derive this function directly from the investor's default model distribution $f(m)$.[10] The advisor chooses m^A to solve:

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

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

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 $\mathbb{E}[U(\theta^{I,0}_{post}, \varphi) | \ell(m^A) < \ell(m^{I,0})]$.

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. Within this framework, the advisor will generally face a tension between these two motives (fit and movement) and, as we will discuss in the next section, 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.

[10] In particular, $G(l) \equiv \int_{-\infty}^{l} g(s)\,\mathrm{d}s$ where $g(l) \equiv \int_{m \in M} \mathrm{I}(\ell(m) = l) f(m)\,\mathrm{d}m$. To establish that G is strictly increasing on $(-\infty, \ell(m^{DO})]$, note that $\ell((c^{DO}, \theta^{DO}_{pre}, \theta_{post}))$ 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.

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 kinds of messages that the advisor will send to the investor.

The investor's decision rule. The investor adopts a message whenever it provides a better 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 than his existing default narrative. As we will explain in more detail 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 θ_{pre}, construct a narrative that induces the investor to adopt the θ_{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.[11] 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 information case studied by S&S where the advisor knows exactly which messages the investor will adopt and therefore can choose her preferred message from that set. 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. When doing this, the advisor will trade off message fit and belief movement. This is shown in the following proposition, which says that the advisor will never send the data-optimal model (unless it coincides with her bliss point). Instead, the advisor has a strict incentive to move away from the data-optimum towards her bliss point, thereby trading off message fit for a

[11] 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.

potentially beneficial belief movement.

Proposition 1. *In the advisor's optimal message,* $(c^*, \theta^*_{pre}, \theta^*_{post})$*, we will have* $(\varphi - \theta^{DO}_{post})^2 \geq (\varphi - \theta^*_{post})^2$*, with a strict inequality whenever* $\varphi \neq \theta^{DO}_{post}$*.*

The proposition says that, unless the advisor's bliss point is exactly equal to the θ_{post} in the data-optimal model, the advisor is always willing to sacrifice some probability of the investor adopting the model in order to move the investor's belief.

We will now hold the message-fit motive fixed to isolate the impact that the belief-movement motive has on message construction. To do this, we will focus on a set of messages which all have the same fit $\bar{\ell}$. Denote this set by $\mathcal{M}(\bar{\ell})$. Our result states that, among all messages in $\mathcal{M}(\bar{\ell})$, the advisor will choose the message $m^*(\bar{\ell})$ whose θ_{post} value *moves* the investor closest to the advisor's objective. We can see how this result follows by inspecting the expected utility function in Equation (2). Fixing the message fit at $\bar{\ell}$, the only moving part in the expected utility function is payoff the advisor will receive if the investor adopts the message (since all messages with the same fit have the same probability of being adopted). Therefore, the advisor will prefer the message which maximizes her payoff in the case of adoption:

Proposition 2. *Among all possible messages with the same message fit* $\bar{\ell}$*, the advisor chooses a message which minimizes the distance between* θ^A_{post} *and* φ*;*

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

Proposition 2 constrains the set of possible messages the advisor will consider; for any message fit $\bar{\ell}$, the advisor will choose a message which, if adopted, will move the investor's assessment closest to the advisor's bliss point. We collect the messages that survive Proposition 2 in a set $\tilde{\mathcal{M}}$. The next result will constrain the set of possible messages that the advisor sends even further. It shows that, among all messages in $\tilde{\mathcal{M}}$, the advisor only considers those whose c^A and θ^A_{pre} parameters maximize the message fit *conditional* on θ^A_{post}. This occurs because the advisor is only directly incentivized to move the investor's θ_{post}-assessment in a certain direction; she does not have any incentive to shift the investor's beliefs about the other two parameters. Therefore, if we hold θ^A_{post} 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 θ_{post} component of the message appear more plausible in view of the historical data.

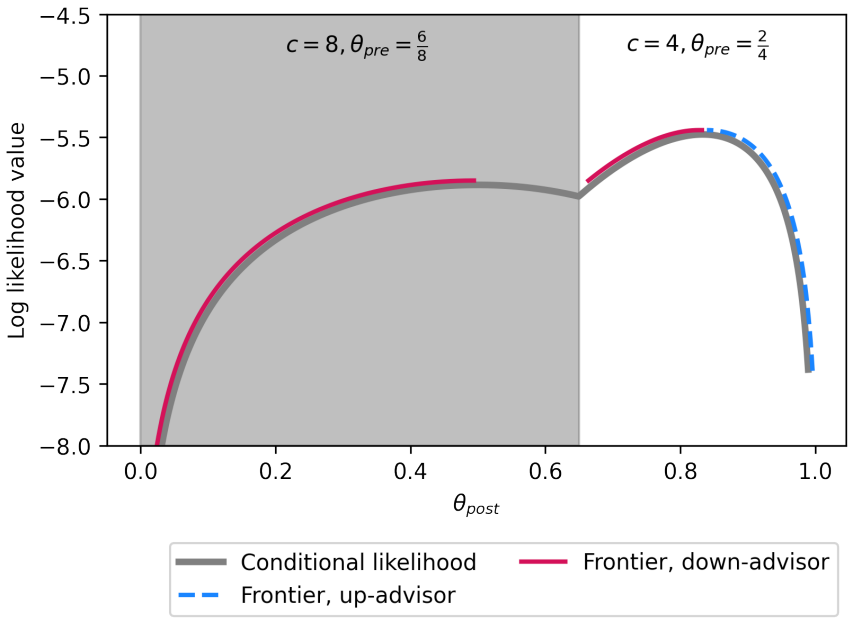
Proposition 3. *Among all messages in* $\tilde{\mathcal{M}}$*, the advisor will only consider sending those whose c and* θ_{pre} *maximize the log likelihood function conditional on* θ_{post}*. This implies that*

$$(c^*, \theta^*_{pre}) \in \underset{(c, \theta_{pre}) \in \{2,\ldots,8\} \times [0,1]}{arg\ max} \ell(c, \theta_{pre}, \theta^*_{post}).$$

Since each θ^A_{post} will yield a pair, (c^A, θ^A_{pre}), that maximizes the fit, the advisor can then compare these messages along the continuum of possible θ^A_{post}s. 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", which is the set of all messages that are not dominated on both their fit and movement by any other message. 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 grey line 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 θ_{post}. It takes on its maximum value when the message equals the data-optimal model, $(c^{DO} = 4, \theta^{DO}_{pre} = 2/4, \theta^{DO}_{post} = 5/6)$.[12] At this point, the up- and the down-advisor's likelihood frontiers, as illustrated by the red and blue lines, almost meet. The up-advisor's likelihood frontier includes all θ_{post} values larger than the data-optimum, as each of these messages can be rationalized under some intensity of the tradeoff between movement and model fit. The likelihood frontier of the down-advisor is instead discontinuous because a range of messages with intermediate θ_{post} parameter values around 0.6 is dominated by messages with lower θ_{post} values which are both closer to the down-advisor's objective of 0 and provide a better fit.

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 θ_{post} and example history $h = (0, 1, 1, 0, 1, 1, 1, 1, 0, 1)$. The c and θ_{pre} values at the top of the figure maximize the conditional maximum likelihood in the respective range of θ_{post} values.

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 is likely to hold a default model which is close to the data-optimal model, then she will send a message where θ^{A}_{post} is close to θ^{DO}_{post}. 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

[12]Note that, conditional on $c = 4$, the data-optimal θ_{pre} and θ_{post} are simply equal to the proportion of successes in their respective periods.

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 θ^A_{post} while the down advisor will decrease it.[13]

Figure 2 also illustrates how belief-movement motive induces the advisor to systematically deviate from the data-optimal structural change parameter in her message construction. To see this, consider the likelihood frontier of the down-advisor. As the down-advisor lowers θ_{post} from the data-optimal value (i.e., when she wishes to move from inducing a θ_{post} that is in the lighter shaded region to one in the darker shaded region), it becomes optimal for her to move the structural change from year 4 towards year 8. A later structural change better justifies a low θ_{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 $c = 8$.

Turning to a systematic study of the advisor's choice of c^A, the following results describe the circumstances under which an advisor deviates from sending the data-optimal structural break parameter when facing an arbitrary history. Taken together, the results suggest that the above example of a down-advisor adjusting c^A to better justify θ^A_{post} as she moves along the likelihood frontier can be generalized. When presenting the results, we differentiate between up- and down-advisors and between cases where the advisor deviates to a c^A that is either smaller or larger than c^{DO}.

Proposition 4. *When constructing the optimal message:*

(i) An up-advisor will send a message with $c^A < c^{DO}$ only if the fraction of success-years in the post period is higher under c^A than c^{DO}, i.e.,

$$\frac{\sum_{t=c^A+1}^{10} s_t}{10-c^A} > \frac{\sum_{t=c^{DO}+1}^{10} s_t}{10-c^{DO}}.$$

(ii) A down-advisor will send a message with $c^A < c^{DO}$ only if the fraction of success-years in the post period is lower under c^A than c^{DO}, i.e.,

$$\frac{\sum_{t=c^A+1}^{10} s_t}{10-c^A} < \frac{\sum_{t=c^{DO}+1}^{10} s_t}{10-c^{DO}}.$$

(iii) An up-advisor will send a message with $c^A > c^{DO}$ only if the number of failure-years in the post period is lower under c^A than c^{DO}, i.e.,

$$\sum_{t=c^A+1}^{10} (1-s_t) < \sum_{t=c^{DO}+1}^{10} (1-s_t).$$

(iv) A down-advisor will send a message with $c^A > c^{DO}$ only if the number of success-years in the

[13] Similarly, the optimal θ^A_{post} of the aligned advisor lies between θ^{DO}_{post} and the true parameter, θ^T_{post}. The θ_{post} sent by the aligned advisor will move closer to the data-optimal model as the trade-off between movement and fit becomes sharper.

post period is lower under c^A than c^{DO}, i.e.,

$$\sum_{t=c^A+1}^{10} s_t < \sum_{t=c^{DO}+1}^{10} s_t.$$

Suppose that the advisor considers lowering c. When doing this, she is essentially shifting years from the *pre* period into the *post* period within the narrative under consideration. Parts (i) and (ii) of the proposition above say that the advisor will only do this if doing so 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). Intuitively, this will allow an up-advisor to use the data to justify a higher θ_{post}.

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, shifts years from the *post* period into the *pre* period. Parts (iii) and (iv) of the statement above state that 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.

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 θ^*_{post} she sends. Since θ^*_{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 parts (i) and (ii) of Proposition 4. Second, when θ^*_{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 θ^*_{post} close to 0 wishes to minimize the number of successes. This follows from the non-linearity of the log likelihood function in θ_{post} which implies that, e.g., for a large θ_{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 parts (iii) and (iv) of Proposition 4: When the up-advisor considers increasing c, she will always do so if her θ^*_{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.

In the discussion above, we have provided 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 θ^*_{post} and θ^{DO}_{post}. 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-maximizing 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 θ^*_{post} and θ^{DO}_{post} increases, it becomes more likely that coupling θ^*_{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 θ^A_{post} she sends is that it will push θ^A_{pre} into the opposite direction. For example, as the up-advisor shifts c to increase the proportion of successes or minimize the number of failures in the *post* period, this will tend to have the opposite effect on the *pre* period. This dynamic will lead to a negative correlation between θ_{pre} and θ_{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 θ_{pre} and θ_{post} parameters in opposite directions.

3.4 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 theories are consistent with the behavior we observe. In Appendix F, we dispense with the assumption that the investor can always be persuaded by any message which provides 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 different 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 θ_{post} is common knowledge.

While it is well established 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), our 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 θ_{post} around the investor's prior belief about θ_{post}. In equilibrium, all messages sent by the advisor include a θ_{post} parameter inside this interval and the investor will always adopt. This result for the strategic framework thus suggests that sophisticated investors will not adopt "extreme" messages, i.e., messages that either contain a high or a low θ_{post} value outside the interval.

One key difference from the S&S framework is that in the strategic framework the θ_{post} parameter is never part of a broader narrative where the advisor uses the c and θ_{pre} components of their message to make their communicated θ_{post} seem more compelling. The reason for this is that, if everyone is strategic, talk about c and θ_{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 θ_{post} values is uniquely determined for any historical data set

and invariant to changes in the non payoff-relevant parameter equilibrium strategies.[14]

4 Experimental Design

To gather empirical evidence on the mechanics of narrative persuasion, we conduct an experiment. The experiment is framed as a financial advice game, with participants either taking on the role of an investor or a financial advisor.

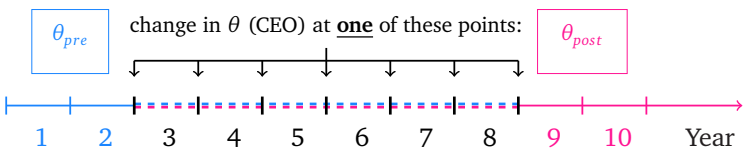
4.1 Baseline Treatment

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 pieces of information being presented side-by-side on the same screen. Based on this information, the investor reports their assessment of 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, θ. The investor and advisor both know that θ can take on values between 0 and 1 and that 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 (θ_{pre}) and the current probability of success (θ_{post}) were each 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



[14]Of course, there are multiple equilibria where the investor does not follow the message off the equilibrium path for certain values of θ_{pre} and c. However, any equilibrium induces the same equilibrium allocations.

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 before the end of 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. Specifically, this means that the advisor has to choose a value for c which lies between 2 and 8.[15] 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 θ_{post}.

Participants in the role of the investor are incentivized to provide an estimate for θ_{post} that is as close as possible to θ_{post}^T. We use the binarized scoring rule (BSR; Hossain & Okui, 2013) to ensure that investors will maximize their expected payment when reporting their expectation about θ_{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 θ_{post}-assessment. Under the three conditions, the advisor is either: (a) an up-advisor whose payoff increases in the investor's estimate of θ_{post}, (b) a down-advisor whose payoff decreases in the investor's estimate of θ_{post}, or (c) an aligned advisor whose payoff increases in the accuracy of the investor's estimate of θ_{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 advisor receiving

[15] In the instructions for the experiment, the year of the structural change 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 to be consistent with this "last year under the old CEO" convention.

the bonus payment is made by the investor, not the advisor (i.e., the belief of the investor determines the advisor's payment). Second the BSR of up- and down-advisors compares the θ_{post} reported by the investor to extreme θ_{post} values, namely $\theta_{post} = 1$ or $\theta_{post} = 0$, to determine the advisor's payment. This incentivizes these advisors to want investors to hold either high or low beliefs and, therefore, differs from the standard BSR which typically compares the reported belief to the truth (i.e., θ^T_{post}) rather than a fixed value. The aligned advisor's BSR incentives are identical 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.[16]

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 $(c^T, \theta^T_{pre}, \theta^T_{post})$ of each company. This serves to provide an opportunity for gains from communication for the investor, since the advisor is better informed. Depending on the advisor's incentives, she might sometimes try to deceive the investor into reporting an overly optimistic or pessimistic belief about θ_{post}. Specifically, the advisor can use the other dimensions of the report, c and θ_{pre}, as supporting evidence in trying to shift this belief about θ_{post}. Informing the advisor about the true model also creates a clear normative distinction between messages that are truth-telling and those that are lies. This aims to replicate an essential feature of advisor-investor relations in real life, namely that advisors are typically better informed. Even though it would be unrealistic to assume that financial advisors know the true 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. In real world scenarios, advisors are also typically morally expected or legally required to

[16]This design feature allows us to keep the advisor's instructions completely constant between the BASELINE, DISCLOSURE, and INVESTORPRIOR treatments. We discuss the additional treatments further below.

provide advice that is accurate to the best of their knowledge. Informing the advisor about the true model allows us to control 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.[17]

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 of the data before receiving the advisor's message. The third reason is that omitting this initial elicitation stage from most treatments helps to simplify the experiment, which should facilitate better participant understanding.

4.1.4 Hypotheses for the BASELINE Treatment

The experimental design allows us to test several implications of the theoretical framework. These hypotheses were preregistered in the AEA registry (AEARCTR-0009103).[18] To test these hypotheses, we rely on the exogenous variation generated within the BASELINE treatment. Below, we provide a discussion of the three intervention treatments and our hypotheses regarding the effect of these interventions on behavior.

First, we ask how advisors react to different incentives or, specifically, how introducing a conflict of interest influences the messages that advisors send. We hypothesize that advisors send self-interested narratives, which implies that the messages of misaligned advisors will be further from the truth. We test this hypothesis by comparing the distance between the advisor's message and the truth for aligned and misaligned advisors. The hypothesis corresponds to Hypothesis 6a from our preregistration.

[17]This feature of the design serves to avoid introducing an additional layer of endogeneity to the experiment that would be present if advisors first formed their own assessment of the data and only then constructed a message to the investor. If advisors are not perfectly informed about the true model, it would be more challenging for the analyst to distinguish advisor mistakes and self-deception from intentional attempts to deceive the investor. In this study, we aim to cleanly identify intentional deception, but we think that studying advisor self-deception as a mechanism for deceiving investors is a promising avenue for future research and discuss this further in our conclusion section.

[18]To keep the main text more focused, we relegate the discussion of three of the preregistered hypotheses to Appendix C. These hypotheses investigate the role that lying aversion of advisors plays and how features of the historical data influence the effectiveness of persuasion. Our empirical results offer support for these omitted hypotheses.

Hypothesis 1a (Belief movement motive, corresponds to PR.6a)**.** *The distance between the advisor's θ^A_{post} and the true value θ^T_{post} is larger for misaligned than for aligned advisors.*

A closely related test of how advisors' message construction is influenced by facing a conflict of interest concerns the narrative part of the advisor problem: An advisor may adjust their choices of c and θ_{pre} to make their report of θ_{post} more convincing. As described in Section 3, to do this, up-advisors should *decrease* their θ^A_{pre} while down-advisors should *increase* it. This suggests that, compared to aligned advisors, the θ^A_{pre} of misaligned advisors should be further from the truth.

Hypothesis 1b (Empirical fit motive, corresponds to PR.6b)**.** *The distance between the advisor's θ^A_{pre} and the true value θ^T_{pre} is larger for misaligned than for aligned advisors.*

Second, we study whether advisors are successful in using narratives to persuade investors. If they are, investor assessments should be further from the truth when they face an advisor with misaligned incentives than when facing an advisor with aligned incentives.

Hypothesis 2 (Persuasiveness of narratives, corresponds to PR.1)**.** *The distance between the investor's assessment and the truth is larger when advisor incentives are misaligned than when advisor incentives are aligned.*

Third, a key assumption of the theoretical framework is that investors decide whether or not to adopt a narrative based on the empirical fit of the narratives. We test this assumption empirically by evaluating whether the distance between the advisor's θ^A_{post} and the investor's assessment, $\theta^{I,1}_{post}$ decreases in the narrative's empirical fit. Essentially, this says that an investor will be more willing to follow an advisor's message if it fits the data they observe well.

Hypothesis 3 (Investors believe plausible narratives, corresponds to PR.5a)**.** *The distance between the advisor's message and the investor's assessment decreases in the empirical fit of the narrative.*

To test this hypothesis, we construct a (pre-registered) index, which we refer to as the Empirical Plausibility Index (EPI), that reflects the empirical fit of every feasible narrative, conditional on a particular historical company data set. The index value for a particular narrative is proportional to the likelihood value of the narrative when evaluated against the data. Section 5 contains a more detailed discussion of the construction of this index.

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.

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 environment which do not influence the data or narrative received from the advisor. We can, however, augment the narrative persuasion framework so that it provides a lens for examining which elements of the decision environment may protect the investor from adopting narratives from conflicted advisors. In scenarios where we think of persuasion as being potentially harmful, interventions which leverage these factors might thus protect the investor from harmful persuasion. In the following section, we describe the design of the intervention treatments as well as the associated hypotheses. As in the BASELINE case, the hypotheses pertaining to the intervention treatment effects were preregistered.

DISCLOSURE. 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 (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 he receives may be biased. 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.

To investigate whether knowing their specific matched advisor's incentives makes investors skeptical, we introduce the DISCLOSURE treatment. In this treatment, the advisor's incentives are fully disclosed to the investor. In each round of the experiment, on the decision screen, investors in DISCLOSURE learn whether they have received a message from an up-, down- or aligned advisor. We hypothesize that investors who are matched with a misaligned advisor will form assessments that are closer to the truth in this DISCLOSURE treatment than they do in BASELINE.[19]

[19]Our hypotheses regarding the treatment interventions focus on interactions involving misaligned senders because we want to study mechanisms that protect investors from harmful persuasion. It is not necessary to protect individuals from persuasion when interests are fully aligned.

Hypothesis 4 (Corresponds to 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*.*

InvestorPrior. 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 prefers. If the investor were instead to always adopt the data-optimal narrative as a default, then no better-fitting alternative narrative would exist 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 might make the investor more immune to adopting the advisor's narrative.[20]

We introduce the InvestorPrior treatment to 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 first receive only the data. We then ask them to report their prior belief about the data generating process (i.e., c, θ_{pre}, and θ_{post}). Thereafter, investors receive the advisor's message, and we elicit their final assessment of θ_{post}.

This treatment allows us to evaluate whether being encouraged to try to make sense of the data oneself first serves a protective function against persuasion using models.[21] We hypothesize that, when matched to a misaligned advisor, investors' assessments are closer to the truth in InvestorPrior than in Baseline.

Hypothesis 5 (Corresponds to PR.3)**.** *When matched with an advisor with misaligned incentives, the distance between the investor's assessment and the truth is smaller in* InvestorPrior *than in* Baseline*.*

PrivateData. 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

[20] 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 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.

[21] An additional benefit of this treatment is that the reported prior beliefs provide us with 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.

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, we would expect that on average the 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.

To investigate whether having access to private data serves a protective role against persuasion, we introduce the PrivateData treatment that varies whether the advisor observes the historical performance data. In particular, both the investor and advisor in this treatment know that the advisor does not observe the historical performance data when choosing her message.[22] 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. Our hypothesis is that, when matched to a misaligned advisor, investors' assessments are closer to the truth in PrivateData than in Baseline.

Hypothesis 6 (Corresponds to PR.4)**.** *When matched with an advisor with misaligned incentives, the distance between the investor's assessment and the truth is smaller in PrivateData than in Baseline.*

Since we are mainly interested in how the interventions may influence investors' narrative adoption, we try to hold the advisor instructions constant wherever possible and only make changes between treatments where they are unavoidable. In particular, advisors in Baseline, Disclosure, and InvestorPrior see exactly the same instructions.[23] This makes it possible for us to hold advisor behavior constant across these treatments, and therefore we are able to attribute any potential treatment effects to changes in investor behavior. It also provides us with a large quantity of data on how advisors craft messages for many different combinations of the true underlying model and the observed historical data set which enables us to study message formation in detail. Since the PrivateData treatment studies an intervention which constrains advisors in their ability to tailor narratives to the data, this treatment necessarily changes the advisor's instructions in addition to introducing changes in the investor's instructions.

[22]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 in personal data, the persuader may not have access to the information that the receiver has at all. For example, a firm may consider only sharing a subset of their proprietary data with a consultancy and then use the other part for a later validation exercise which tests for the out of sample fit of the consultancy's suggestions. In addition, 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.

[23]To keep advisor instructions the same between Disclosure and the remaining two treatments, advisors in all treatments are told that investors "may or may not" know the advisor's incentives. We also fix second-order beliefs of investors across treatments by informing them that advisors know that investors may or may not know their incentives.

4.3 Procedures

The experiment was conducted via the Prolific platform. We recruited 360 participants (180 advisors and 180 investors) for BASELINE and 180 participants (90 advisors and 90 investors) for each of the intervention treatments. Participants in the experiment were balanced by gender.[24] In designing the experiment, we devoted substantial attention towards ensuring that we explained the experiment 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. The Appendix 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 was 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 time-of-the-day effects.[25]

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 underlying model in each round of the experiment. Specifically, we drew ten triplets of fundamentals, $(c^T, \theta^T_{pre}, \theta^T_{post})$, 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 received a bonus payment for one randomly chosen round of the experiment. This additional bonus that the investors and advisors could earn was £3.75. For each participant, the probability of earning the bonus depended on the relevant binarized scoring rule described above, which was evaluated in relation to 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.

[24] See Appendix B.1 for summary statistics of participant demographics by treatment.

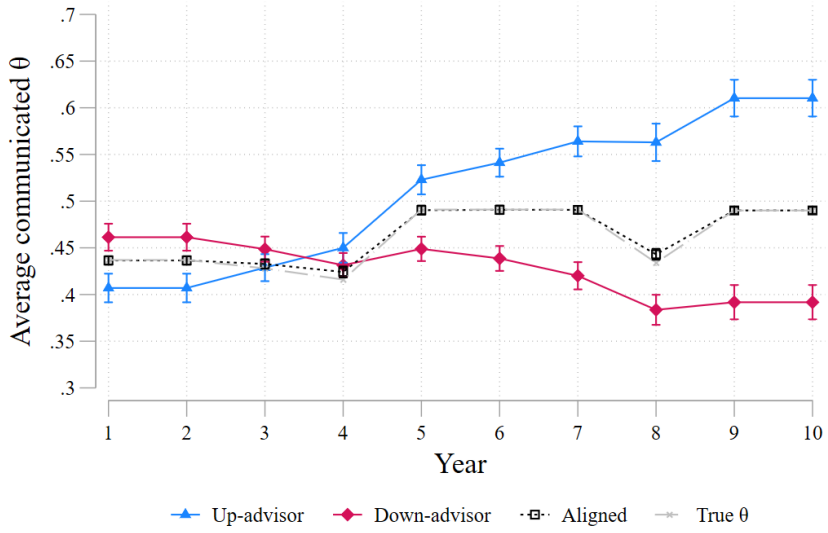
[25] We collected the advisor data for each group one day before we collected the investor data. Therefore, participants were not randomly allocated to a particular role conditional on the session. We did this so that participants did not have to wait for their group members to finish their assessments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

5 Results

5.1 Advisor Narrative Construction

We start by examining how advisors construct narratives. Within our theoretical framework, advisors face a tradeoff between belief *movement* and empirical *fit* when constructing narratives. An implication of this is that Hypotheses 1a and 1b posit that misaligned advisors should bias θ^A_{pre} and θ^A_{post} away from the truth. They should, however, bias θ^A_{pre} and θ^A_{post} for different reasons. While misaligned advisors should bias θ^A_{post} towards their persuasion target, they should bias θ^A_{pre} in the *opposite* direction.

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



Notes: The figure includes data from advisors who received the BASELINE instructions. Error bars represent 95% confidence intervals that were derived from regressions which cluster standard errors at the advisor level.

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, θ, in each of the ten years—this is given by the θ_{pre} from the period before the CEO change and the θ_{post} from the period after the CEO change. To obtain Figure 4, we take the average θ for each year across all messages sent by advisors of each type. We can see that up-advisors (denoted by the blue line) construct messages that imply a *higher* θ in year 10 than down-advisors (denoted by the red line). Conversely, up-advisors send messages with a *lower* θ in year 1 than down-advisors. As one might expect, the messages sent by the aligned advisors imply θs 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 regressions that test Hypotheses 1a and 1b—i.e., we test whether

the θ_{post} and θ_{pre} parameters sent by misaligned advisors are further from the truth.

Column (1) shows that the average misaligned advisor reports a θ_{post} that is 13pp further from the truth than the average aligned advisor. Similarly, column (2) shows that advisors also shift the θ_{pre} component of the narrative 6pp further from the truth when they hold misaligned incentives. Thus, in addition to finding statistical evidence that advisors adjust θ_{post} in response to the incentives they face, we 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, θ_{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) $\lvert\theta^A_{post} - \theta^T_{post}\rvert$	(2) $\lvert\theta^A_{pre} - \theta^T_{pre}\rvert$
Misaligned advisor = 1	12.72***	6.492***
	(0.702)	(0.660)
Dep. var. aligned adv. mean	1.478	1.929
Round FE	Yes	Yes
Observations	3600	3600

Notes: (i) The dependent variable is the distance between the true θ parameter and the corresponding θ parameter of the advisor's message, (ii) The sample contains data from all advisors who received the BASELINE instructions, (iii) For each advisor we have 10 observations—one for each round, (iv) Standard errors are clustered at the advisor level, implying that there are 360 clusters, and are reported in parentheses; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

This evidence is further corroborated when we compare the θ^A sent by advisors to the θ^A that would have been data-optimal, given the advisor's choice of c^A.[26] Figure B.2 in Appendix B.2 shows that, while up-advisors exaggerate θ_{post} relative to the data-optimum and down-advisors attenuate it, all advisor types choose a θ^A_{pre} that is close to the data-optimum on average. This suggests that advisors do engage in a trade off between movement and fit when choosing θ_{post}, while the fit motive guides them exclusively in choosing θ_{pre}. These results are in line with the theoretical predictions. The Appendix also presents further results on the advisor's choice of c^A which suggest that misaligned advisors systematically choose their c^A in ways that better justify their choice of θ^A_{post}. Together, these results indicate that advisors distort their θ^A_{post} report in a self-interested way and use the other components of the narrative to try to make their message taken as a whole convincing when compared to the data.

Result 1 (Related to Hypotheses 1a and 1b)**.** *Relative to aligned advisors, misaligned advisors*

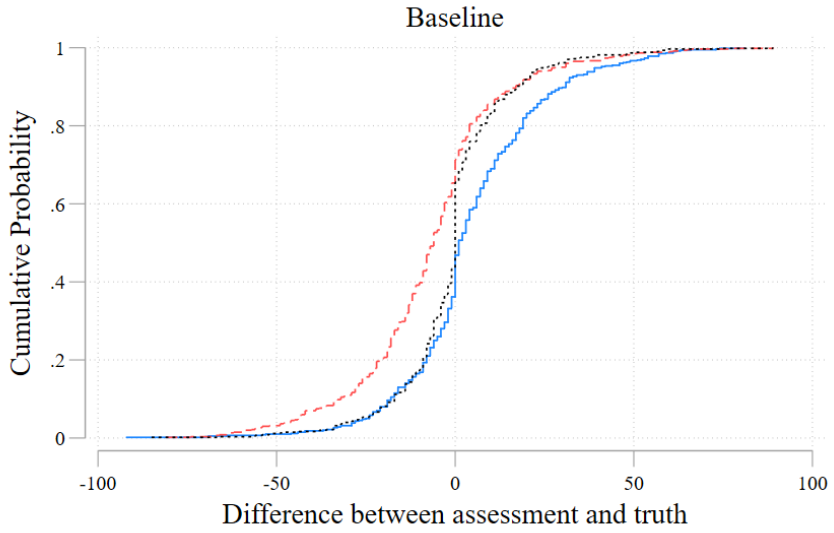
[26] As discussed in the theory section, data-optimal refers to parameters that maximize the fit of the narrative to the data—here, we consider the θ^A_{pre} and θ^A_{post} that best fit the data, conditional on the choice of c^A. For example, if an advisor were to choose $c^A = 5$ for a data set where there are 3 successes between years 6 and 10, then the data-optimal θ^A_{post} given c^A would be 3/5.

bias θ^A_{post} and θ^A_{pre} away from the truth. They bias θ^A_{post} towards their persuasion target and θ^A_{pre} in the opposite direction.

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 5 provides an initial visual answer to this question by plotting the cdf of the distance between investors' beliefs, θ^I_{post}, and the truth, θ^T_{post}. 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.

Figure 5: Distance between investor belief, θ^I_{post}, and the truth, θ^T_{post} (by advisor type)



Notes: (i) The figure uses data from the *Baseline* treatment, (ii) The figure plots the cdf of the measure $\theta^I_{post} - \theta^T_{post}$ for all investor-rounds where the investor is matched with a particular advisor type, (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.

Are advisors successful in using narratives to persuade investors? We test Hypothesis 2 explicitly in Table 2 by asking whether advisors are able to shift the beliefs of the average investor through the narratives they send. This table reports the results from a regression with

the distance between the investor's assessment and the truth as the dependent variable and a misaligned advisor dummy as the dependent variable. 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, θ^T_{post}. 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 2: Movement of investor beliefs when matched with a misaligned advisor

	$\lvert\theta^{I,1}_{post} - \theta^T_{post}\rvert$
Misaligned advisor = 1	5.111***
	(0.679)
Aligned adv. dep. var. mean	10.163
Round FE	Yes
Observations	1800

Notes: (i) The dependent variable is the absolute distance between the investor's belief, θ^I_{post}, and the true value θ^T_{post}, (ii) The sample contains data from all investors in BASELINE, (iii) For each of the investors, we have 10 observations—one for each round (iv) Standard errors are clustered at the Interaction Group level (i.e., the matching group of 3 investors and 3 advisors), implying that there are 60 clusters, and are reported in parentheses; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

Result 2 (Related to Hypothesis 2)**.** *The average investor's assessment is further away from the truth when their advisor has misaligned incentives in comparison to when their advisor has aligned incentives.*

What types of narratives do investors follow? According to the theoretical framework, investors will be more willing to follow advice when the proposed narrative fits the data well. Specifically, Hypothesis 3 posits that the average investor's assessment will move closer to the advisor's message as the empirical fit of the advisor's narrative increases. We measure the narrative fit with what we call the Empirical Plausibility Index (EPI). To derive the EPI, 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.[27] Therefore, the EPI takes on values between 0 and 1. A value of 1 can be obtained if the advisor sends the the data-optimal (best-fitting) narrative and a minimum value of 0 is obtained if the advisor sends the worst-fitting narrative.[28] We use the EPI to test the hypothesis, by regressing the distance between the advisor's message, θ^A_{post}, and the investor's report on the EPI of the

[27]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.

[28]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—i.e., a narrative containing either $\theta_{pre} = \theta_{post} = 0$ or $\theta_{pre} = \theta_{post} = 1$ will have a likelihood value of zero.

advisor's narrative, controlling for round fixed effects and clustering at the Interaction Group level (i.e., the matched group of 3 advisors and 3 investors). The results are reported in Table 3. 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, θ^A_{post}. This suggests that investors find narratives that fit the data well to be more compelling.

Table 3: Investor conformity and the fit of the advisor's narrative

	$\lvert\theta^{I,1}_{post} - \theta^A_{post}\rvert$
Advisor message fit (EPI)	-14.59***
	(1.892)
Misaligned advisor = 1	0.691
	(0.668)
Dependent variable mean	11.085
Round FE	Yes
Observations	1800

Notes: (i) The dependent variable is the absolute distance between the investor assessment and the advisor narrative (ii) The sample contains data from all investors in BASELINE, (iii) For each of the investors, we have 10 observations—one for each round (iv) Standard errors are clustered at the Interaction Group level (i.e., the matching group of 3 investors and 3 advisors), implying that there are 60 clusters, and are reported in parentheses; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

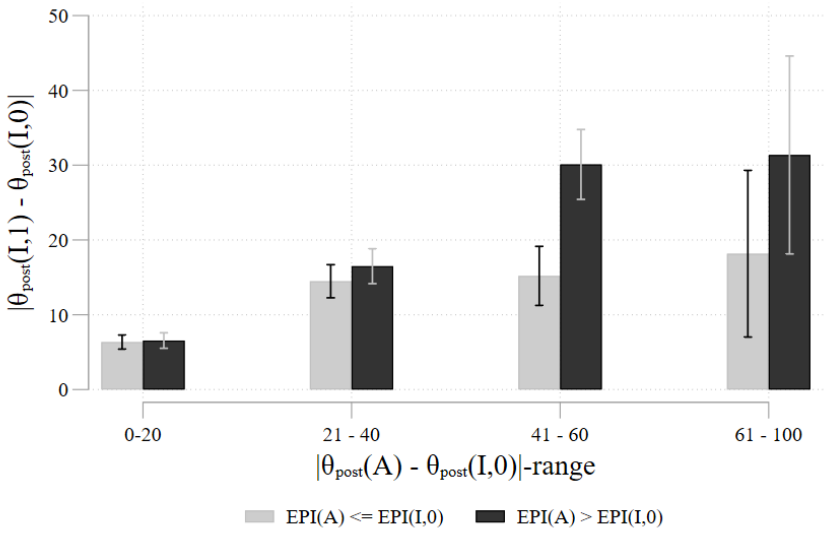
Result 3a (Related to Hypothesis 3)**.** *The average investor's assessment moves closer to their advisor's narrative as the empirical fit of the narrative increase.*

Is this "investor conformity-narrative fit" relationship due to belief updating? One concern that may be raised regarding the relationship documented in Table 3 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 $|\theta^{I,1}_{post} - \theta^A_{post}|$ and EPI(m^A) that we find in Table 3. The first potential explanation is about updating—when an advisor proposes a better-fitting narrative (with a higher EPI), the investor shifts their beliefs more, resulting in a smaller gap between the investor's posterior belief, $\theta^{I,1}_{post}$, and advisor's message, θ^A_{post}. 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's 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 θ^{A}_{post} will likely be closer to the investor's default $\theta^{I,0}_{post}$ than if the advisor sends a model with a low fit. The reason is simply that both the investor's prior has a high fit and therefore the high-fit advisor's message is likely to be closer to it than the low-fit advisor's message. 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, could lead to a spurious correlation between $|\theta^{I,1}_{post} - \theta^{A}_{post}|$ 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 more 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 6: 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 distance between the advisor's θ^{A}_{post} and the investor's prior $\theta^{I,0}_{post}$ and the difference between the fit of the advisor's message and the investor's default model, (iv) Error bars are 95% confidence intervals derived from regressions which cluster standard errors at the interaction-group level.

Figure 6 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, $|\theta^{I,1}_{post} - \theta^{I,0}_{post}|$, and the x-axis disaggregates the data into categories according to the distance between the advisor's message and the investor's prior belief, $|\theta^{A}_{post} - \theta^{I,1}_{post}|$. 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 fits 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 θ^{A}_{post} and the investor's prior $\theta^{I,0}_{post}$ 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 to be less risky than in the case where the difference is large.

Table 4: Belief updating and narrative fit

	(1) $\lvert\theta^{I,1}_{post} - \theta^{I,0}_{post}\rvert$	(2) $\lvert\theta^{I,1}_{post} - \theta^{I,0}_{post}\rvert$	(3) $\lvert\theta^{I,1}_{post} - \theta^{I,0}_{post}\rvert$	(4) $\lvert\theta^{I,1}_{post} - \theta^{I,0}_{post}\rvert$
I($EPI^{A} > EPI^{I,0}$)	3.465***	3.350***	-2.203*	-1.393
	(0.835)	(0.852)	(1.172)	(1.190)
Misaligned advisor	0.0117	-0.165	-0.733	-0.681
	(1.090)	(1.204)	(0.747)	(0.810)
$\lvert\theta^{I,0}_{post} - \theta^{A}_{post}\rvert$			0.266***	0.363***
			(0.0530)	(0.0547)
I($EPI^{A} > EPI^{I,0}$) × $\lvert\theta^{I,0}_{post} - \theta^{A}_{post}\rvert$			0.238***	0.173**
			(0.0729)	(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
Observations	900	779	900	779

Notes: (i) The outcome variable in the regressions in this table is the absolute distance that investors update, $\lvert\theta^{I,1}_{post} - \theta^{I,0}_{post}\rvert$, (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) The sample contains data from investors in InvestorPrior, (iv) 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, (v) For each of the investors, we have 10 observations—one for each round, (vi) Standard errors are clustered at the Interaction Group level (i.e., the matching group of 3 investors and 3 advisors), implying that there are 30 clusters, and are reported in parentheses; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

Table 4 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 updates their θ_{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 θ^{A}_{post} and the investor's prior, $\theta^{I,0}_{post}$, 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.[29] 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

[29]Table B.2 in Appendix B.3 reports the results from an additional set of robustness exercises that either use (i) the EPI difference or (ii) only the advisor's narrative's EPI as alternative, continuous measures of relative narrative fit. The results are robust to these alternative measures.

of the first channel discussed above.

Result 3b (Related to Hypothesis 3)**.** *The better the empirical fit of a narrative, the more investors update their beliefs towards this narrative.*

5.3 Additional Results from the Baseline Treatment

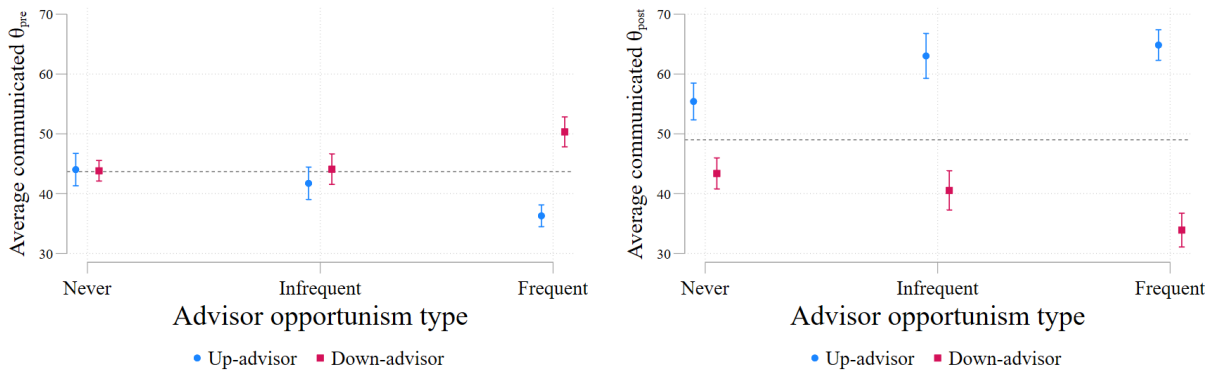
In this section, we provide a discussion of several additional results on heterogeneity in advisor narrative construction and on how narrative persuasion interacts with the historical data.

5.3.1 Advisor Heterogeneity

The results on narrative construction reported so far outline aggregate patterns in the narrative 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 θ_{pre} and c to construct a convincing narrative. We use the repeated nature of the experiment to classify advisors according to the different strategies they might use when constructing their narratives. One measure of an advisor's skill in narrative construction is whether and how frequently she adjusts the non payoff-relevant parameters in ways that support her 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 θ_{post} is using the malleability of the structural break parameter to their advantage.

Figure 7: Average θ^As, by opportunism type



(a) Average values of θ^A_{pre}

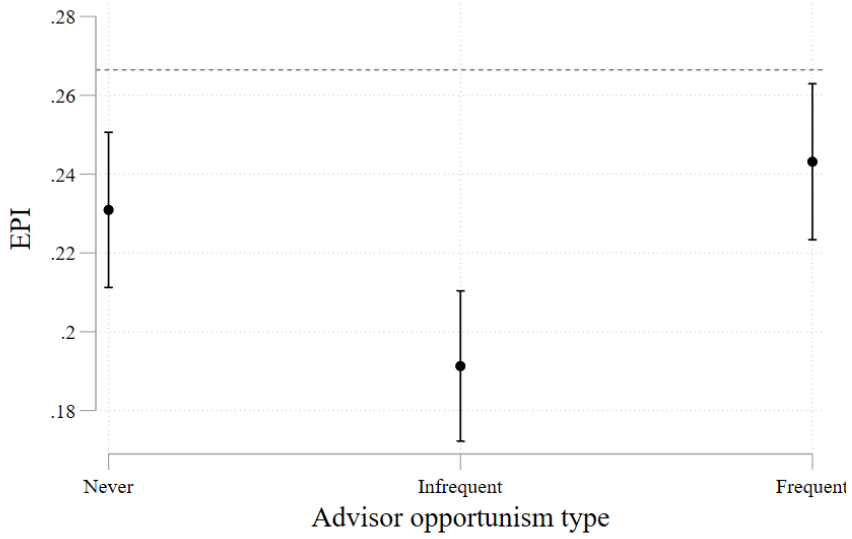
(b) Average values of θ^A_{post}

Notes: The figure includes data from advisors who received the Baseline instructions. The dashed line denotes the average true θ_{pre} and θ_{post} encountered by advisors in Baseline. Error bars are 95% confidence intervals derived from regressions which cluster standard errors at the advisor level.

To examine heterogeneity in advisor's narrative construction skills, 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 approximately 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.[30]

Based on this classification, we can ask what kinds of messages advisors of each opportunism type send. Figure 7 shows the average θ_{pre} and θ_{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 θ_{post}-reports towards their persuasion goal. However, more opportunistic types bias their θ_{post}-reports by more. Second, never-opportunists do not on average bias their θ_{pre}-report away from the true parameter value, which is in line with the idea that they are not constructing narratives in a sophisticated way. In contrast, we can see that the difference in average θ_{pre}-reports is driven by the frequent opportunists, who are the only type which systematically bias θ_{pre} in an opposite direction to θ_{post}. This indicates that these frequent-opportunists are engaging in a sophisticated form of narrative construction.

Figure 8: Average EPI of advisor messages, by opportunism type



Notes: 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.

[30]Among the 240 misaligned advisors who received the BASELINE instructions, 79 are never-, 80 are infrequent-, and 81 are frequent-opportunists.

These different strategies also affect narrative quality or message fit, which we proxy using the EPI measure defined in the previous subsection. Figure 8 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 fit similar to that of the true model. Frequent-opportunists on the other hand introduce a high bias in θ_{post}, but, by adjusting the non payoff-relevant parameters, they achieve empirical fits comparable to the never-opportunists (and not too far off the average fit of the true model, denoted by the dashed horizontal line). Since the infrequent-opportunists tend to bias θ_{post} by relatively large amounts without adjusting the non payoff-relevant parameter values, their narratives have comparatively worse fits.

Flexibly adjusting narratives to the data makes them more persuasive. In the previous section, we asked which features make individual narratives convincing and showed that the empirical fit matters. In this section, we ask what types of advisors have more persuasive success. Specifically, we examine whether 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 results on the persuasiveness of narratives sent by different opportunism types. Table 5 reports the results from regressing the investor's assessment of θ_{post} on the advisor's sent θ_{post}-value, her opportunism type as well as the interaction of the two. The table reveals several insights. First, we find that the relationship between $\theta^{I,1}_{post}$ and θ^{A}_{post} is positive. This is consistent with the findings discussed above, namely that investors, to an extent, follow the narratives of their advisors. Second, the coefficients associated with the opportunism type interactions, however, suggest that the strength of the relationship between narrative and assessment depends on the advisor's opportunism type. The interaction coefficient for the infrequent type is negative, which suggests that infrequent-opportunists are less successful than never-opportunists at persuading investors. In contrast, the interaction coefficient for the frequent type is insignificantly different from zero, which suggests that frequent-opportunists are not significantly less persuasive than never-opportunists (furthermore, the difference between the interaction term coefficients of infrequent- and frequent-opportunists is marginally significant ($p = 0.087$)). These results are consistent with previous results on how frequent-opportunists send narratives with empirical fits which are indistinguishable from those of never-opportunists, who often tell the truth. They suggest that narratives—if frequently adjusted to fit the data—are an effective tool for persuasion.[31]

[31]This conclusion is also supported by evidence on the expected payoffs received by the different opportunism types: The expected payoff is increasing in the "score" achieved by the different advisor incentive-types, which is equal to $\theta^{I,1}_{post}$ for up-advisors and equal to $1-\theta^{I,1}_{post}$ for down-advisors. Averaged over the different opportunism types, the average scores are 53.52 (never opportunists), 54.68 (infrequent opportunists), and 58.96 (frequent opportunists). The score of frequent opportunists is also significantly higher than the score of never and infrequent opportunists ($p < 0.005$). This suggests that, among the different opportunism types, frequent opportunists realized the highest expected payoffs in the experiment.

Table 5: Relation between advisor's narrative and investor's assessment, for different opportunism types

	$\theta^{I,1}_{post}$
β_1: θ^{A}_{post}	0.550***
	(0.0340)
β_2: θ^{A}_{post}×Opportunism: Infreq.	-0.0897***
	(0.0324)
β_3: θ^{A}_{post}×Opportunism: Freq.	-0.0222
	(0.0304)
Opportunism: Infrequ.	5.679***
	(1.623)
Opportunism: Frequent	4.159**
	(1.595)
H_0: $\beta_2 = \beta_3$ p-value	.087
Round FE	Yes
Observations	1200

Notes: (i) The outcome variable in the regressions in this table is the investor's assessment (ii) The sample contains data from investors in BASELINE, (iii) For each of the investors, we have 10 observations—one for each round, (iv) Standard errors are clustered at the Interaction Group level (i.e., the matching group of 3 investors and 3 advisors), implying that there are 60 clusters, and are reported in parentheses; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

5.3.2 Disentangling the Influence of Data and Narrative on the Investors' Assessments

While much of the analysis above has focused on evaluating the impact of the advisor's narrative on the investor's assessment, it is also informative to examine how the investor uses the historical data directly to form his assessment. In particular, we can ask whether more recent successful years in the company's history have a larger effect on his assessment than years further in the past. And, importantly, we can ask whether the narrative proposed by his advisor mediates how he draws inference from the data. To analyze the relationship between the investors' assessments and history, we estimate the following regression equation:

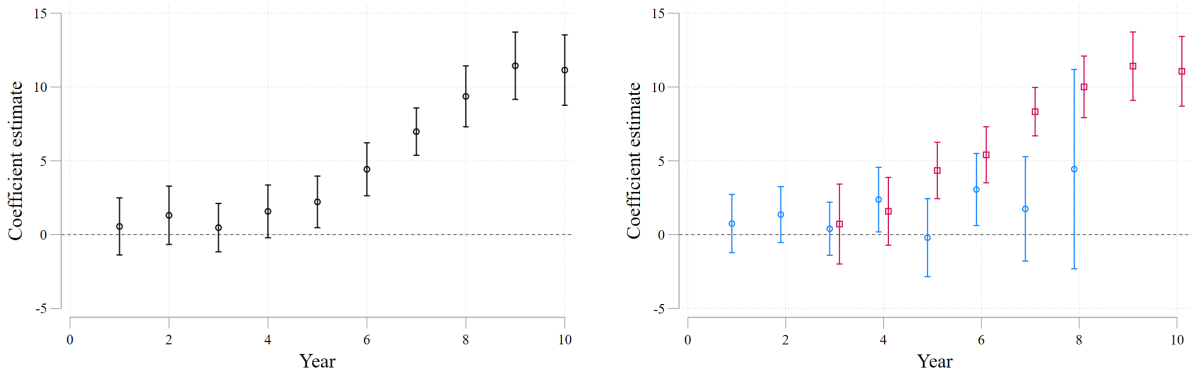
$$\theta^{I,1}_{post} = \sum_{t=1}^{10} \beta_t s_t + \rho + \varepsilon.$$

In the equation above, s_t indicates a success in year t and ρ are round fixed effects. The left panel of Figure 9 plots the β-coefficient estimates. The qualitative patterns of the coefficient estimates imply that investors interpret the data in reasonable ways. Successes in year 9 or 10—where the investor is sure that they belong to the *post* period—have the largest effect on investors' assessments (as they should). The effect of a success between years 3 to 8, where the investor is uncertain whether any individual year belongs to the *post* period, is gradually increasing. Finally, the coefficient estimates are not significantly different from zero in years 1

and 2, which always belong to the *pre* period.

By sending a narrative, the advisor can potentially change how the investor interprets the data. In particular, by providing a suggestion regarding the year in which the CEO changed, the advisor essentially tells the investor which years to focus on to assess the company's future probability of success. The right panel of Figure 9 plots coefficient estimates from regressions which interact success and failure with dummy variables that indicate whether a year belongs to the company's *pre* or *post* period, *according to the advisor's narrative.* The figure gives an insight into the interaction between data and narrative. After receiving a narrative, the investor places more weight on evidence from years between 3 and 8 if those years are in the post period (red) relative to those in the pre period (blue) according to the advisor's narrative ($p < 0.001$, see section B.4 for regression outputs and formal tests). This result is consistent with the idea that the advisor influences which years in the data the investor deems relevant when making his assessment.

Figure 9: Effect of company success on assessments, by year



(a) Average effect of a success

(b) Effect of a success in *pre* and *post*

Notes: The left panel plots coefficient estimates of the marginal effect of a success in year t in the data on the investor's assessment, using data from the BASELINE. The right panel plots the same coefficient estimates interacted with whether the advisor suggested that the year belongs to the *pre* period (blue) or to the *post* period (red). Error bars are 95% confidence intervals derived from regressions which cluster standard errors at the matching group level.

Mechanisms. Some of the observed difference between the effect of a success in *pre* and *post* might be driven endogenously by how advisors construct narratives. Specifically, there are two mechanisms that could generate the pattern of behavior observed in the right panel of Figure 9. First, the advisor's choice of c^A could mediate the inference drawn by the investor from the successes and failures in the data, as discussed above. Second, advisors themselves adjust the θ^A_{post} that they send to fit the data, given their choice of c^A. Therefore, if the investor sometimes adopts the advisor's θ^A_{post}, they will appear to be weighting the suggested *post* periods more strongly in their assessment. We can partially account for this second mechanism by controlling for the direct effect of θ^A_{post} on the assessment in the regression. Our results suggest that, while there is a direct, significantly positive effect of θ^A_{post} on the investor's assessment, the company history and the narrative still have a lasting impact. In particular, the gap between the *pre* and

post coefficient estimates as displayed in Figure 9 (and reported in Section B.4) remains even after controlling for θ^A_{post}. This suggests that the advisor's narrative has a lasting impact on how the investor interprets the data that goes beyond the direct effect of θ^A_{post}. This suggests that, even if investors distrust the θ^A_{post} they receive, they might still take the c^A that they receive from the advisor at face value and allow it to influence their assessment.

Robustness. One potential concern with the results discussed above is that there might be a positive correlation between the advisor's c^A and the investor's prior belief, $c^{I,0}$, about the structural break before receiving the advisor's narrative. Then, the gap that we observe in the influence of successes on the investor's assessment between years attributed to *pre* and *post* might exist even when before the investor is exposed to the advisor's narrative. We address this concern in two ways in Section B.4. First, we conduct a placebo exercise where we replace the investor's assessment with the data-optimal assessment and run the same regressions. We do not find that the split of the company history into *pre* and *post*, as suggested by the advisor, has a measurable effect in this placebo regression ($p = 0.283$). This suggests that even if investors were very adept at using the data directly to estimate the structural break, it would not generate the relationship that we observe in the right panel of Figure 9. We also conduct a similar second exercise using the investor's prior belief, $\theta^{I,0}_{post}$, that we observe in InvestorPrior as the dependent variable in the regression. In this exercise we also do not find a statistically significant effect of the *pre-post* split ($p = 0.606$). We also run a diff-in-diff regression and show that the difference in the updating gap between prior beliefs and final assessments is statistically significant ($p = 0.002$).

5.4 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 the three treatment interventions by comparing investor behavior in these treatments 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. Disclosure asks what happens when advisors' incentives are fully disclosed, InvestorPrior essentially nudges investors to carefully evaluate the data themselves prior to meeting their advisor, and PrivateData considers a scenario where advisors do not have full knowledge of the data that investor's see. Our hypotheses 4, 5, and 6 state that, relative to Baseline, each of the interventions will bring the investor's assessment closer to the truth.

Table 6 tests these hypotheses by examining whether the interventions do indeed 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 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.

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

	Disclosure $\lvert\theta^I_{post} - \theta^T_{post}\rvert$		InvestorPrior $\lvert\theta^I_{post} - \theta^T_{post}\rvert$		PrivateData $\lvert\theta^I_{post} - \theta^T_{post}\rvert$	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Intervention = 1	-0.713	2.403	0.454	1.241	-0.124	-0.0775
	(1.001)	(1.549)	(0.924)	(1.117)	(0.750)	(1.192)
Advisor lied=1		9.340***		9.200***		9.419***
		(1.012)		(1.024)		(1.018)
Intervention × Advisor lied		-3.974**		-0.764		0.116
		(1.633)		(1.425)		(1.558)
Dep. var. 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 θ^T_{post} parameter and the corresponding belief held by the investor θ^I_{post}, (ii) Each column uses data from the Baseline treatment as well as the relevant treatment mentioned in the column header, (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) Standard errors are clustered at the Interaction Group level, reported in parentheses, (v) there are 90 clusters (v) The results in columns (*a) relate to Hypotheses 2, 3 and 4 from the pre-registration, (vi) ** $p < 0.05$, *** $p < 0.01$.

Surprisingly, we see that none of the three interventions has a statistically significant protective effect for the average investor.

Result 4 (Related to hypotheses 4, 5, and 6)**.** *Relative to Baseline, none of the treatment interventions brings investors significantly closer to the truth.*

For the InvestorPrior and PrivateData treatments, this result can also be seen visually in Appendix Figure B.6, 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.

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. Out of all narratives sent by misaligned advisors, 30,083% are actually truthful.[32] However, since investors only know the

[32]See Appendix Figure C.1 for a histogram showing the number of lies by sender type in the experiment.

advisor's incentives but not whether they are truthful, they cannot easily distinguish advisors who are being honest from those that are being dishonest. This implies that in DISCLOSURE they also become skeptical of narratives received from misaligned 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 in DISCLOSURE 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, even when investors are explicitly told that they face an advisor with a conflict of interest, on average, they are not better off than when they were uncertain about the advisor's interests.

5.5 Quantifying How Investors Update Their Beliefs

The results so far suggest that investors react to the fit of the message they receive. In addition, we saw that the intervention treatments failed to move the average investor's belief closer to the truth, despite changing investor behavior (e.g., making investors more skeptical in the DISCLOSURE treatment). To better understand the underlying mechanics that govern the investor's narrative adoption decision, it is informative to structurally estimate the underlying decision-relevant parameters. This will allow us to quantify the degree to which investors are willing to adopt particular types of narratives. For example, the baseline S&S framework suggests that investors will fully adopt a narrative (i.e., completely replace their prior narrative with the received narrative) if the empirical fit of the received narrative is better than the prior. This full-adoption rule is captured by Equation (1). It is easy to imagine alternative partial adoption rules which instead suggest that investors are not fully credulous, but can only assess the relative empirical fit with noise. Our aim in this section is to estimate a narrative adoption, or updating, rule which maps the investor's default narrative and the advisor's narrative into the investor's final assessment. To do this, we adopt a more flexible version of the S&S framework that allows the data to determine which factors are most relevant for the investors' decisions. We consider the following specification::

$$\begin{aligned}\hat{\theta}^{I,1}_{post}(\kappa,\lambda;m^{I,0},m^{A},h) &= p(\kappa,\lambda;m^{A},m^{I,0},h)\cdot\theta^{A}_{post}+(1-p(\kappa,\lambda;m^{A},m^{I,0},h))\cdot\theta^{I,0}_{post},\\ \text{where } p(\kappa,\lambda;m^{A},m^{I,0},h) &\equiv \frac{\exp\{\kappa+\lambda\Delta\text{EPI}(m^{I,0},m^{A},h)\}}{1+\exp\{\kappa+\lambda\Delta\text{EPI}(m^{I,0},m^{A},h)\}},\end{aligned} \tag{3}$$

where $\Delta\text{EPI}(m^{I,0},m^{A},h)$ is the difference between the EPIs of the advisor's narrative and investor's default model when both are evaluated against history h. Therefore, equation 3 specifies the investor's assessment as a function of the advisor's narrative, the investor's default model, the observed historical data, and two parameters, κ and λ. This updating rule can be derived from a model which extends the investor's narrative adoption rule as specified in

Equation 1 in two ways. The first extension introduces a parameter that measures the extent of the investor's credulity. In the equation above, credulity is denoted by κ. As κ increases, the investor will put more weight on the received narrative independently of the narrative's empirical fit. The second extension introduces noise to the narrative selection rule, in a spirit similar to Froeb et al. (2016). The idea here is that investors perceive the empirical fit of both their default model and the received narrative with noise, such that their perception of the fit equals the actual empirical fit plus a noise term. In the updating rule, the parameter λ measures the precision of the noise term. As λ increases—which indicates that the investor can more accurately detect differences in the true empirical fit—p becomes more responsive to the relative narrative fit. We formally derive this general updating rule in Appendix D.

The specification above nests various particular updating rules. For example, the parameter combination $(\kappa = 0, \lambda \to \infty)$ captures the updating rule suggested by S&S—investors adopt the advisor's narrative if and only if it provides a better empirical fit than the default model. The cases $(\kappa \to \infty, \lambda = 0)$ and $(\kappa \to -\infty, \lambda = 0)$ capture complete credulity (the investor always adopts the advisor's narrative regardless of fit) and complete skepticism (the investor never adopts the advisor's narrative). Parameter combinations with intermediate values of κ and λ describe updating rules involving a compromise, where the investor's expected assessment is a weighted average of the narrative and the default.

The challenge in estimating κ and λ is that, we only elicit the investor's default model in the InvestorPrior treatment. We address this issue in the estimation by replacing the default model in Equation (3) with the expected default model given the history. This requires approximating the distribution of default models. To achieve this, we assume that the distribution of the default model's EPI (but not of the default model parameters) is independent of the history and the treatment. Under these assumptions, we derive a non-parametric estimate of the distribution of default models based on the sample distribution in InvestorPrior. We then use this distribution to take the expectation over the default model in Equation (3). Thereafter, we derive estimates for κ and λ by minimizing the squared distance between the observed assessment and the assessment predicted by Equation (3).

Appendix D includes the estimation details. It also examines whether our approach of taking the expectation over $m^{I,0}$ is reasonable. We do this by comparing the estimates that we obtain from using our approach with the InvestorPrior data with estimates obtained from exploiting the additional individual-level information on default models available for the InvestorPrior treatment. The point estimates of both approaches are similar and not significantly different. This suggests that assuming that the distribution of default model EPIs is invariant to histories is reasonable. We also compare the explanatory power obtained when using the additional individual-level information to the case where it is not used and generally find that it remains relatively high even if we take the expectation over default models. While the model that uses the individual-specific default model data has a Mean Squared Error (MSE) of 122.73, this increases to 184.81 when we estimate the model taking the expectation over default models. The MSE value remains at comparable levels if the model parameters are

estimated using data from Baseline. This suggests that the model can reasonably explain investors' assessments in treatments different than InvestorPrior.

Table 7: Updating parameters estimation results

	(1)	(2)	$H_0 : \kappa = \kappa_{\text{Baseline}}$ p-value
κ_{Baseline}	0.232	0.201	
	(0.145)	(0.148)	
$\kappa_{\text{InvestorPrior}}$		-0.25	0.086
		(0.217)	
$\kappa_{\text{PrivateData}}$		-0.311	0.052
		(0.219)	
$\kappa_{\text{DisclosureAligned}}$		1.301***	0.022
		(0.458)	
$\kappa_{\text{DisclosureMisaligned}}$		-0.005	0.44
		(0.222)	
λ	1.272***	0.904***	
	(0.461)	(0.342)	
MSE	198.315	205.171	
Sample	Baseline	All treatments	
Observations	1800	4500	

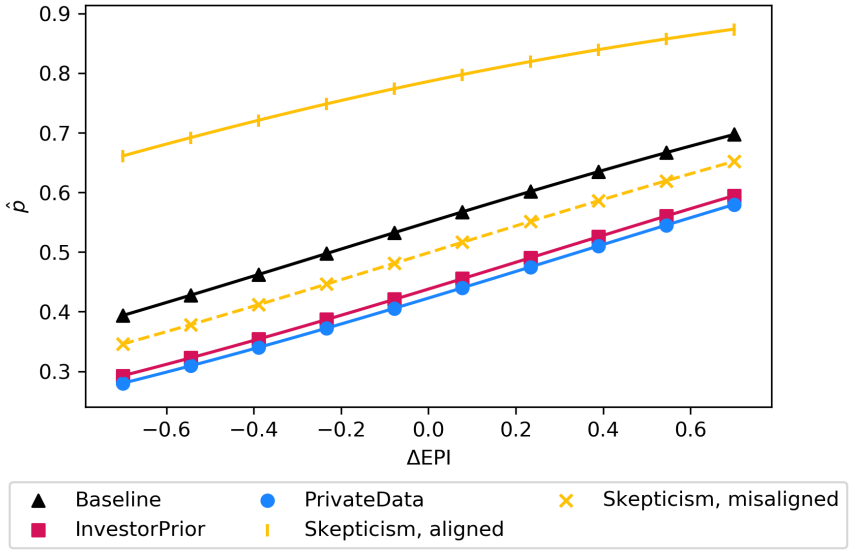
Notes: Bootstrapped standard errors in parentheses. *** $p < 0.01$.

Table 7 presents estimation results. Column (1) only uses data from Baseline. The estimation results show that the precision parameter, λ, is significantly larger than zero, which suggests that investors can with some precision distinguish between the empirical fits of different models, but do so with noise. Column (2) estimates the model using observations from all treatments and allows for the credulity parameter to vary at the treatment level. Figure 10 plots the corresponding weighting functions p that are implied by the model estimates. The plotted lines suggest that, for example, advisors in Baseline can shift the expected weight that investors put on the narrative from ca. 40% to more than 60% by increasing the fit of their proposed narrative (relative to the default narrative fit).[33] The results also suggest that the credulity parameter, κ, in most conditions is distributed around zero, which implies that investors weigh the default and the narrative approximately equally if their model fits are similar.[34] The exception is the DisclosureAligned condition, where investors meet an aligned advisor in Disclosure. Here, the credulity parameter is estimated to be significantly larger than zero and also larger than the credulity parameter in Baseline. This makes sense as investors in this treatment know that they are meeting and advisor with aligned incentives and therefore should be more willing to accept their advice.

[33]Around 95% of all messages induce an EPI difference between -0.6 and 0.6, which is the range over which the weighting function is plotted in the figure.

[34]The ratio κ/λ measures investor credulity in units of the EPI difference. For example, in the Column (1) specification, $\kappa/\beta \approx 0.182$, which implies that investors weigh the default and the narrative equally if the narrative's EPI is .182 points below the fit of the default model.

Figure 10: Expected weight that investors put on the narrative (by treatment)



Notes: The figure plots the p function, as specified in Equation (3), using the parameter estimates reported in Column (2) of Table 7.

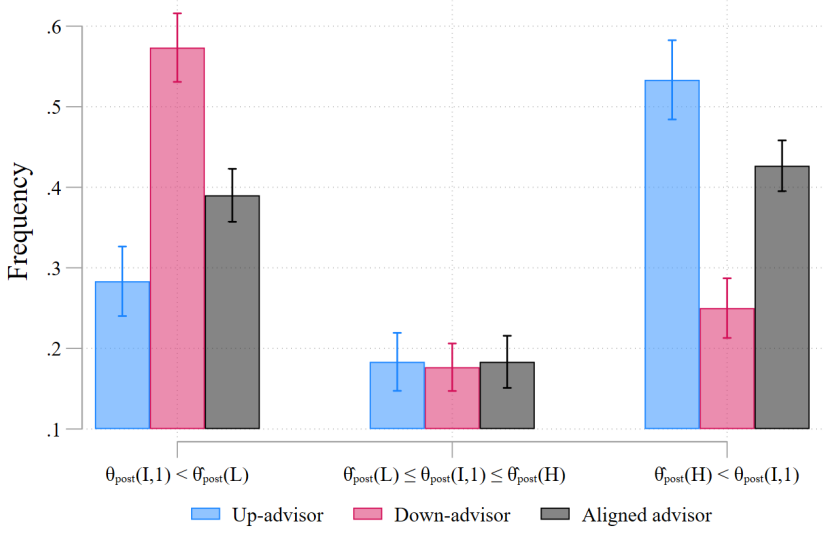
5.6 Consistency of the Experimental Data with Nash Equilibrium

While all of the analysis above has examined our data through the lens of an augmented version of the S&S framework, we can also test for the existence of patterns in the data that are predicted by the most informative Nash equilibrium of the underlying cheap talk game. In this way, we can ask whether the behavior we observe is also consistent with the sophisticated strategic thinking typically assumed by Nash equilibrium analyses. As we discuss in Section 3.4 and Appendix F, in the most informative equilibrium of the game underlying the BASELINE treatment, some persuasive communication around the investor's prior expectation of θ_{post} is possible; the Nash equilibrium predicts the existence of a lower threshold, θ^L_{post}, and an upper threshold, θ^H_{post}, with $\theta^L_{post} < \mathbb{E}[\theta_{post}] < \theta^H_{post}$. The investor adopts a message only if it includes a θ_{post} parameter on the interval between the two thresholds.

We can solve numerically for these thresholds, which are uniquely determined for every historical data set observed by participants in the experiment. With these thresholds in hand, 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 θ^L_{post}, (ii) between θ^L_{post} and θ^H_{post}, or (iii) larger than θ^H_{post}, 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: Relative to being matched with an aligned advisor, being matched with a down-advisor significantly increases the proportion of assessments lower than θ^L_{post} while being matched with an up-advisor significantly increases the proportion of

assessments above θ^H_{post}. 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.

Figure 11: Frequency of investor assessments that fall below, within, and above the interval consistent with Nash equilibrium (by advisor type)



Notes: The figure includes data of investors who participated in the BASELINE treatment. Error bars are 95% confidence intervals derived from regressions which cluster standard errors at the matching group level.

We provide further results in Appendix B.6, by investigating whether the distance between the investor's assessment and the advisor's message is discontinuous around the Nash equilibrium thresholds. This can be seen to be a weaker test of Nash-related behavior, because it merely tests whether investors grow more skeptical to messages that are just outside the threshold interval. The results provide some limited support for a modest increase in the distance to the message if the message is just outside the interval, though effect size and significance of this result is not consistent across different specifications. Therefore, the evidence that investors grow more skeptical to messages around the Nash equilibrium threshold is weak.

Taken together, these results indicate that behavior is not very consistent with sophisticated strategic thinking.

6 Concluding Discussion

The discussion above has provided empirical evidence showing how narratives can be used as a tool for persuasion, with one individual shaping how another interprets objective data. The results are largely in line with the persuasion mechanics outlined in the S&S theoretical framework. Specifically, 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—they distort the target parameter in the direction of their private self-interest and use the auxiliary parameters to make their deception seem more plausible by improving the overall fit of the narrative. This behavior is consistent with a narrative construction approach that trades off *fit* and *movement*. Advisors in the experiment manage to construct narratives which influence the assessments of investors; this is especially the case for narratives 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 distort 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.

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 & 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.[35] 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 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

[35] 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.

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 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.[36] 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 poten-

[36] See Andre, Haaland, 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.

tially 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.

References

Abeler, J., Becker, A., & Falk, A. (2014). Representative evidence on lying costs. *Journal of Public Economics*, *113*, 96–104. doi: 10.1016/j.jpubeco.2014.01.005

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for Truth-Telling. *Econometrica*, *87*(4), 1115–1153. doi: 10.3982/ECTA14673

Aina, C. (2021). Tailored Stories. *Mimeo*.

Akerlof, G. A., & Snower, D. J. (2016). Bread and Bullets. *Journal of Economic Behavior & Organization*, *126*, 58–71.

Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M., & Yang, D. (2020). Polarization and Public Health: Partisan Differences in Social Distancing During the Coronavirus Pandemic. *Journal of Public Economics*, *191*, 104254.

Andre, P., Haaland, I., Roth, C., & Wohlfart, J. (2022). Narratives about the macroeconomy. *Working Paper*.

Andre, P., Pizzinelli, C., Roth, C., & Wohlfart, J. (2022). Subjective Models of the Macroeconomy: Evidence from Experts and a Representative Sample. *Review of Economic Studies*, *89*(6), 2958–2991.

Barron, K., Harmgart, H., Huck, S., Schneider, S., & Sutter, M. (2022). Discrimination, Narratives and Family History: An Experiment with Jordanian Host and Syrian Refugee Children. *Review of Economics and Statistics*.

Barron, K., Huck, S., & Jehiel, P. (2019). Everyday econometricians: Selection neglect and overoptimism when learning from others. *WZB Discussion Paper*.

Bénabou, R., Falk, A., & Tirole, J. (2020). Narratives, Imperatives, and Moral Persuasion. *Working Paper*.

Blume, A., DeJong, D. V., Kim, Y.-G., & Sprinkle, G. B. (1998). Experimental evidence on the evolution of meaning of messages in sender-receiver games. *The American Economic Review*, *88*(5), 1323–1340.

Blume, A., DeJong, D. V., Neumann, G. R., & Savin, N. (2002). Learning and communication in sender-receiver games: an econometric investigation. *Journal of Applied Econometrics*, *17*(3), 225–247.

Braghieri, L., Levy, R., & Makarin, A. (2022). Social media and mental health.

Bruner, J. (1991). The Narrative Construction of Reality. *Critical Inquiry*, *18*(1), 1–21. Retrieved from http://www.jstor.org/stable/1343711

Cai, H., & Wang, J. T.-Y. (2006, July). Overcommunication in strategic information transmission games. *Games and Economic Behavior*, *56*(1), 7–36. Retrieved 2019-02-18, from https://linkinghub.elsevier.com/retrieve/pii/S0899825605000692 doi: 10.1016/j.geb.2005.04.001

Cain, D. M., Loewenstein, G., & Moore, D. A. (2005). The dirt on coming clean: Perverse effects of disclosing conflicts of interest. *The Journal of Legal Studies*, *34*(1), 1–25.

Charles, C., & Kendall, C. (2022). Causal narratives. *Mimeo*.

Charnysh, V. (2021). Remembering past atrocities—good or bad for attitudes toward minorities? *Mimeo*.

Chater, N., & Loewenstein, G. (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., & Sobel, J. (1982). Strategic Information Transmission. *Econometrica*, *50*(6), 1431-1451. Retrieved 2020-04-15, from https://www.jstor.org/stable/1913390?origin=crossref doi: 10.2307/1913390

Eliaz, K., & Spiegler, R. (2020). A Model of Competing Narratives. *American Economic Review*, *110*(12), 3786–3816. Retrieved 2020-11-30, from https://pubs.aeaweb.org/doi/10.1257/aer.20191099 doi: 10.1257/aer.20191099

Eliaz, K., Spiegler, R., & Weiss, Y. (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., & Zimmermann, F. (2019). Correlation neglect in belief formation. *The Review of Economic Studies*, *86*(1), 313–332.

Eyster, E., & Rabin, M. (2005). Cursed Equilibrium. *Econometrica*, *73*(5), 1623–1672.

Eyster, E., & Weizsacker, G. (2016). Correlation neglect in portfolio choice: Lab evidence. *Available at SSRN 2914526*.

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise-an experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525–547. doi: 10.1111/jeea.12014

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.

Froeb, L. M., Ganglmair, B., & Tschantz, S. (2016). Adversarial Decision Making: Choosing between Models Constructed by Interested Parties. *The Journal of Law and Economics*, *59*(3), 527–548. doi: 10.1086/689283

Gennaioli, N., & Shleifer, A. (2018). *A crisis of beliefs*. Princeton University Press.

Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying Aversion and the Size of the Lie. *American Economic Review*, *108*(2), 419–453. doi: 10.1257/aer.20161553

Graeber, T., Zimmermann, F., & Roth, C. (2022). Stories, statistics, and memory. *CESifo Working Paper*.

Hagenbach, J., & Perez-Richet, E. (2018). Communication with evidence in the lab. *Games and Economic Behavior*, *112*, 139–165.

Hagmann, D., Minson, C., Tinsley, C., 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., & Rasmusen, E. (2018). Coarse Grades: Informing the Public by Withholding Information. *American Economic Journal: Microeconomics*, *10*(1), 210–235. doi: 10.1257/mic.20130078

Harrs, S., Müller, L. M., & Rockenbach, B. (2021). How optimistic and pessimistic narratives about covid-19 impact economic behavior.

Heidhues, P., Kőszegi, B., & Strack, P. (2018). Unrealistic Expectations and Misguided Learning. *Econometrica*, *86*(4), 1159–1214. doi: 10.3982/ECTA14084

Herman, L., & Vervaeck, B. (2019). *Handbook of Narrative Analysis*. University of Nebraska Press.

Hillenbrand, A., & Verrina, E. (2022). The Differential Effect of Narratives on Prosocial Behavior. *Games and Economic Behavior*, *135*, 241–270.

Hossain, T., & Okui, R. (2013). The Binarized Scoring Rule. *Review of Economic Studies*, *80*(3), 984–1001. doi: 10.1093/restud/rdt006

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., Luca, M., & Martin, D. (2021). Is no news (perceived as) bad news? an experimental investigation of information disclosure. *American Economic Journal: Microeconomics*, *13*(2), 141–73.

Kamenica, E., & Gentzkow, M. (2011, October). Bayesian Persuasion. *American Economic Review*, *101*(6), 2590–2615. Retrieved 2019-04-26, from http://pubs.aeaweb.org/doi/10.1257/aer.101.6.2590 doi: 10.1257/aer.101.6.2590

Karlsson, N., Loewenstein, G., McCafferty, J., et al. (2004). The economics of meaning. *Nordic Journal of Political Economy*, *30*(1), 61–75.

King, R. R., & Wallin, D. E. (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., Ungeheuer, M., & Weber, M. (2019). How to alleviate correlation neglect. *CEPR Discussion Paper No. DP13737*.

Laudenbach, C., Weber, A., & Wohlfart, J. (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., Sah, S., & Cain, D. M. (2012). The unintended consequences of conflict of interest disclosure. *Jama*, *307*(7), 669–670.

Mailath, G. J., & Samuelson, L. (2020). Learning under Diverse World Views: Model-Based Inference. *American Economic Review*, *110*(5), 1464–1501. Retrieved 2020-06-21, from https://pubs.aeaweb.org/doi/10.1257/aer.20190080 doi: 10.1257/aer.20190080

Malmendier, U., & Shanthikumar, D. (2007). Are small investors naive about incentives? *Journal of Financial Economics*, *85*(2), 457–489.

Mannheim, K. (2015). *Ideology and Utopia*. USA: Martino Publishing.

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

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, *115*(2), 502.

Morag, D., & Loewenstein, G. (2021). Narratives and valuations. *Available at SSRN 3919471*.

Olea, J. L. M., Ortoleva, P., Pai, M. M., & Prat, A. (2021). Competing Models. *Working Paper*. Retrieved from http://arxiv.org/abs/1907.03809

Pennington, N., & Hastie, R. (1986). Evidence Evaluation in Complex Decision Making. *Journal of Personality and Social Psychology*, *51*(2), 242.

Pennington, N., & Hastie, R. (1988). Explanation-Based Decision Making: Effects of Memory Structure on Judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 521.

Pennington, N., & Hastie, R. (1992). Explaining the Evidence: Tests of the Story Model for Juror Decision Making. *Journal of Personality and Social Psychology*, *62*(2), 189.

Polletta, F., Chen, P. C. B., Gardner, B. G., & Motes, A. (2011). The Sociology of Storytelling. *Annual Review of Sociology*, *37*, 109–130.

Roos, M., & Reccius, M. (2021). Narratives in economics. *arXiv preprint arXiv:2109.02331*.

Sah, S., & Loewenstein, G. (2014). Nothing to declare: Mandatory and voluntary disclosure leads advisors to avoid conflicts of interest. *Psychological science*, *25*(2), 575–584.

Sah, S., Loewenstein, G., & Cain, D. M. (2013). The burden of disclosure: increased compliance with distrusted advice. *Journal of personality and social psychology*, *104*(2), 289.

Schumacher, H., & Thysen, H. C. (2022). Equilibrium contracts and boundedly rational expectations. *Theoretical Economics*, *17*(1), 371–414.

Schwartzstein, J., & Sunderam, A. (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., & Lagnado, D. (2015). Causality in Thought. *Annual Review of Psychology*, *66*(1), 223–247. Retrieved 2021-02-09, from http://www.annualreviews.org/doi/10.1146/annurev-psych-010814-015135 doi: 10.1146/annurev-psych-010814-015135

Spiegler, R. (2016). Bayesian Networks and Boundedly Rational Expectations. *Quarterly Journal of Economics*, *131*(3), 1243–1290. Retrieved 2019-02-20, from https://academic.oup.com/qje/article-lookup/doi/10.1093/qje/qjw011 doi: 10.1093/qje/qjw011

Spiegler, R. (2020a). Behavioral Implications of Causal Misperceptions. *Annual Review of Economics*, *12*(1), 81–106. Retrieved 2021-01-22, from https://www.annualreviews.org/doi/10.1146/annurev-economics-072219-111921 doi: 10.1146/annurev-economics-072219-111921

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., Spezio, M., & Camerer, C. F. (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 & Reccius (2021) show that the number of economics publications containing the term "narrative" in the title or abstract has been growing sharply for at least the past ten years. Nevertheless, as noted by Roos & 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 & Loewenstein (2021) view a narrative as "a story [that] places selected events on a timeline and establishes causal links between them" (p. 2).[37] 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 and coauthors draw 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 & Spiegler, 2020; Eliaz, Spiegler, & 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.[38] Similar to Bayesian Networks approach, Schwartzstein & Sunderam (2021) consider a setting in which individuals may inter-

[37]This view is reminiscient of that proposed by Akerlof & Snower (2016), who characterizes narratives as "a sequence of causally linked events and their underlying sources" (p. 58). This conceptualization adopted by Morag & Loewenstein (2021) and Akerlof & Snower (2016) draws on the perspective commonly adopted in psychology (see, e.g., Bruner, 1991; Pennington & Hastie, 1992; Sloman & Lagnado, 2015). Roos & 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).

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

pret an existing data set in different ways. However, while the line of research by Spiegler and coauthors focuses predominantly on the implications of subjectivity in the causal structures (DAGs) that are used to interpret data, Schwartzstein & Sunderam (2021) instead formalize subjective models as likelihood functions and focus on how individuals select amongst models given the data. To illustrate some of the key differences between the two approaches, in Appendix Section A.4 we provide a simple example that shows how individuals end up adopting certain narratives in the setup studied in our experiment according to each of these two theoretical frameworks. More generally, one key difference is that Schwartzstein & Sunderam (2021) study model selection when there is finite data. In such scenarios, the true model may sometimes appear less compelling in the data than an incorrect model. S&S focus on model fit as the criterion for selecting between models (i.e., S&S propose a fit-based decision rule for explaining how individuals select between models). In contrast, the Bayesian Networks approach of, e.g. Eliaz & Spiegler (2020), considers inference from an infinitely large data set where the true model will maximize the likelihood. Therefore, using model fit to select between models in this setting would select the true model (if it is considered). The Bayesian Networks literature, therefore, focuses less on analyzing how individuals select between models and more on the implications of interpreting a data set through the lens of an incorrect causal model (DAG).

Given the fluidity of the concept of narratives in everyday usage and the relative infancy of it's 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.[39] 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

[39]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 & 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".[40] Specifically, in the examples, the concept is used to refer to providing an explanation for a collection of events.[41] 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 & 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, & 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 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.[42] 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-constructor 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

[40]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.

[41]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 & Loewenstein (2016).

[42]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.[43]

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).

[43]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.[44] 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 the 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, Haaland, 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 & 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, Haaland, et al., 2022; Charles & 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,

[44]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 & Spiegler (2020), Andre, Haaland, et al. (2022), and Charles & Kendall (2022) focuses predominantly on the first level of narrative formation (i.e., *narrative construction*), Schwartzstein & 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, Haaland, et al., 2022; Charles & Kendall, 2022). However, there are also several other features in which we differ First, we follow Schwartzstein & 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 & Spiegler (2020) and empirically by Charles & 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, Haaland, et al., 2022), and the adoption of hopeful narratives and narrative transmission (Charles & 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 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.

A.4 An illustrative comparison of the Spiegler and S & S frameworks.

To illustrate some of the key differences between the Spiegler and S&S approaches, we provide a simple example that shows how individuals end up adopting certain narratives in the setup studied in our experiment according to each of the theoretical frameworks.

Our experiment studies a setup with one outcome variable $y \in \{0, 1\}$ (failure or success). We also have a time variable $t \in \{1, \ldots, 10\}$. The probability of the company being successful, $\Pr(y = 1)$, depends on the company's CEO. This CEO changed in one of seven different years, leading to a structural break in the success probability. Suppose now that we want to use the Bayesian Networks approach to analyze how narratives influence the interpretation of the available data. To do this, let us represent the seven different possible structural breaks with seven additional variables $c_2, c_3, \ldots, c_8$. These variables are defined as follows:

$$c_i \equiv \mathbb{I}(t > i) \text{ for } i \in \{2, 8\},$$

where $\mathbb{I}$ is the indicator function. These latent variables allow us to represent each of the possible structural change candidates as a simple DAG, which then permits analysis using the Bayesian Networks approach.[45] Now, consider an example history, $h = (0, 1, 1, 1, 0, 0, 1, 0, 1, 1)$, which contains a sequence of realizations of the outcome variable, y. Together with the seven latent variables, c_i, this yields the following data set:

Table A.1: Data set for the example history $h = (0, 1, 1, 1, 0, 0, 1, 0, 1, 1)$

t	y	c_2	c_3	c_4	c_5	c_6	c_7	c_8
1	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0
3	1	1	0	0	0	0	0	0
4	1	1	1	0	0	0	0	0
5	0	1	1	1	0	0	0	0
6	0	1	1	1	1	0	0	0
7	1	1	1	1	1	1	0	0
8	0	1	1	1	1	1	1	0
9	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1

Suppose that an individual believes that the CEO changed after year 6. We can represent this as the agent believing in the DAG $c_6 \to y$. Then, to form a belief about the quality of the old CEO, the individual would calculate:

$$\Pr(y = 1|\text{old CEO}) = \Pr(y = 1|c_6 = 0) = \frac{1}{2}.$$

This is done simply by looking at the c_6 column, selecting the rows in which $c_6 = 0$, and calculating the average value of y corresponding to these rows.

Analogously, under the new CEO,

$$\Pr(y = 1|\text{new CEO}) = \Pr(y = 1|c_6 = 1) = \frac{3}{4}.$$

In a similar way we can calculate the success probability beliefs under the old and new CEO of individuals who believe that the structural break took place in any given year t using $\Pr(y = 1|c_t = 0)$ and $\Pr(y = 1|c_t = 1)$: Essentially, a narrative in this framework is a lens for interpreting the data—here, holding a particular narrative guides the individual's attention to a specific c-column of the data set, which they then use to form their belief. When forming

[45] One way to think of this is the following. The individual knows that the CEO determines the probability of success of the company (i.e., $CEO \to y$). If we define a variable z, which takes a value of 1 in years in which the new CEO is in charge and 0 in which the old CEO is in charge, then we can write down the true DAG as $z \to y$. The issue is that individuals in our experiment don't know when the structural change occurred. This implies that z is an unobserved variable—they do not know the values of z in each year. Instead, they know that there are seven potential versions of z. These are the c_i variables—one is correct and six are not, but the individual does not know which one. This implies that there are seven potential DAGs, $c_i \to y$, that could explain the data, but the individual does not know which is the true DAG.

this belief, the individual will choose DAG-parameters (i.e., the two success probabilities) that are most consistent with the data. Therefore, coherency of the adopted narrative is an important feature of narratives highlighted by Eliaz & Spiegler (2020). Importantly, this coherency with the data pertains to the *selection of the DAG-parameters*, conditional on the selected DAG; coherency is not the criterion used for the *selection of the DAG itself* (i.e., the selection of the c-column here).

While coherence between the model and the data is also of central importance in the Schwartzstein & Sunderam (2021) framework, there are several important differences between the two approaches. First, in the Schwartzstein & Sunderam (2021), the fit criterion is used to select between different candidate models that involve a proposal containing a particular c-column as well as DAG-parameters. Under the Bayesian Networks approach, the fit is not used to select between DAGs, rather it is relevant for predicting how agents select the DAG-parameters, conditional on the DAG.

Second, under Schwartzstein & Sunderam (2021)-style analysis that we conduct for our setting, individuals do not necessarily choose the most coherent DAG-parameters, conditional on their adopted DAG. Intuitively, since they are willing to adopt any model that fits better than their default, they may be willing to adopt a model where the DAG-parameters fit the data poorly, provided the overall model fit is superior to their default. For example, in the Schwartzstein & Sunderam (2021)-framework an individual who holds the default model $m^d = (c = 5, \theta_{pre} = 3/5, \theta_{post} = 3/5)$, which is the most coherent parameterization of the DAG given that $c = 5$, would be convinced to adopt an persuader's narrative $m^A = (8, 3/5, 1)$. While the persuader's narrative is not the most coherent DAG parameterization given $c = 8$, it has a higher overall empirical fit than the individual's default. Therefore, a key difference is that instead of focusing on the implications of mistakes in subjective beliefs about causal relations, the Schwartzstein & 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.

A third key difference is that the two frameworks have different implications for belief formation from a big vs. small data set. Suppose that the company in the example above produces multiple binary outputs $y_1, \ldots, y_K$ in every year, all determined by the same DGP (i.e., for each year t, we observe K rows in the data). As K grows large, the true parameterization of the DGP will almost surely maximize the likelihood function. Therefore, a persuader in Schwartzstein & Sunderam's framework will find it more difficult to convince the individual to adopt a different narrative than the truth under big data than under small data. In contrast, since internal coherence (conditional on the DAG) is the sole criterion in Eliaz & Spiegler's framework, a persuader can still suggest different structural breaks to induce different beliefs in the individual, even if under big data. Whether the individual will be willing to adopt the narrative will then be decided on the margin by the individual's preferences, i.e., whether the narrative suggested by the persuader induces beliefs that maximize the individual's anticipated utility. Therefore, while both approaches study mistakes that may arise in the interpretation of data, they each

shine a spotlight on different features of narratives that may influence the individual's narrative adoption decision. Schwartzstein & Sunderam focus on the sense-making function of narratives, which proposes that individuals intrinsically prefer one story to another if it gives them a more convincing explanation of the data. Eliaz & Spiegler assume that individuals are able to form the most coherent belief structure, conditional on their adopted DAG but do not provide a ranking of the different internally coherent DAG-structures. Instead they assume that another aspect of the narrative—whether it raises the individual's anticipated utility—is the relevant narrative selection criterion. In both of these frameworks, individuals make the same mistake of not questioning the persuader's motive for proposing a certain narrative. That is, they are strategically unsophisticated.

B Additional Results

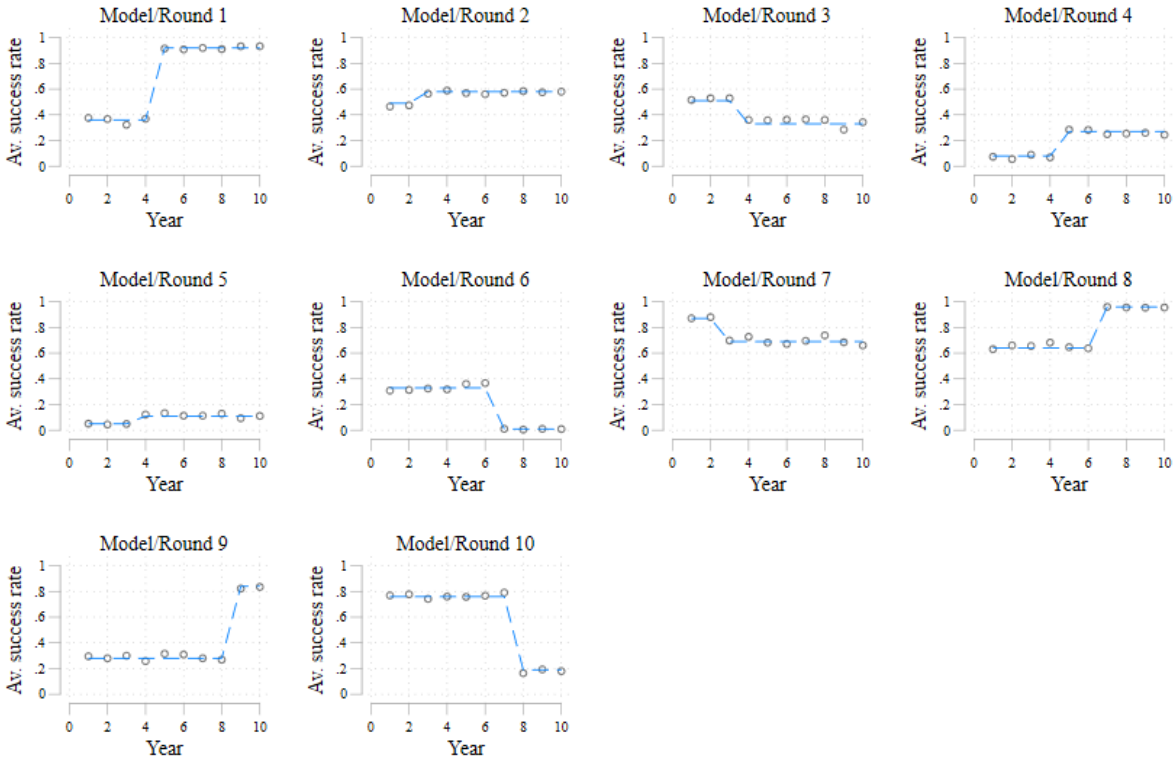
B.1 Additional Results for Section 4.3: Procedures

Table B.1: Demographic characteristics of participants (by treatment and role)

	Baseline	Skepticism	InvestorPrior	PrivateData
Investors				
Age	35.878	35.500	34.989	34.967
	(12.030)	(11.619)	(12.257)	(12.471)
Gender: Female	0.506	0.500	0.511	0.556
	(0.501)	(0.503)	(0.503)	(0.500)
Gender: Male	0.483	0.478	0.489	0.433
	(0.501)	(0.502)	(0.503)	(0.498)
Gender: Other	0.011	0.022	0.000	0.011
	(0.105)	(0.148)	(0.000)	(0.105)
Edu: Primary school	0.017	0.000	0.000	0.011
	(0.128)	(0.000)	(0.000)	(0.105)
Edu: Secondary school	0.056	0.089	0.078	0.111
	(0.230)	(0.286)	(0.269)	(0.316)
Edu: Higher secondary education	0.211	0.244	0.200	0.289
	(0.409)	(0.432)	(0.402)	(0.456)
Edu: College or university	0.467	0.389	0.489	0.389
	(0.500)	(0.490)	(0.503)	(0.490)
Edu: Post-graduate	0.250	0.267	0.233	0.189
	(0.434)	(0.445)	(0.425)	(0.394)
Edu: Prefer not to say	0.000	0.011	0.000	0.011
	(0.000)	(0.105)	(0.000)	(0.105)
Observations	180	90	90	90
Advisors				
Age	36.044	34.389	36.278	35.800
	(12.674)	(11.624)	(12.469)	(12.190)
Gender: Female	0.506	0.467	0.444	0.556
	(0.501)	(0.502)	(0.500)	(0.500)
Gender: Male	0.489	0.522	0.544	0.411
	(0.501)	(0.502)	(0.501)	(0.495)
Gender: Other	0.006	0.011	0.011	0.033
	(0.075)	(0.105)	(0.105)	(0.181)
Edu: Primary school	0.000	0.000	0.000	0.011
	(0.000)	(0.000)	(0.000)	(0.105)
Edu: Secondary school	0.078	0.089	0.067	0.111
	(0.269)	(0.286)	(0.251)	(0.316)
Edu: Higher secondary education	0.183	0.244	0.244	0.267
	(0.388)	(0.432)	(0.432)	(0.445)
Edu: College or university	0.478	0.467	0.467	0.411
	(0.501)	(0.502)	(0.502)	(0.495)
Edu: Post-graduate	0.244	0.189	0.211	0.189
	(0.431)	(0.394)	(0.410)	(0.394)
Edu: Prefer not to say	0.017	0.011	0.011	0.011
	(0.128)	(0.105)	(0.105)	(0.105)
Observations	180	90	90	90

Note: Standard deviations in parenthesis.

Figure B.1: Average observed history vs true model (by round)



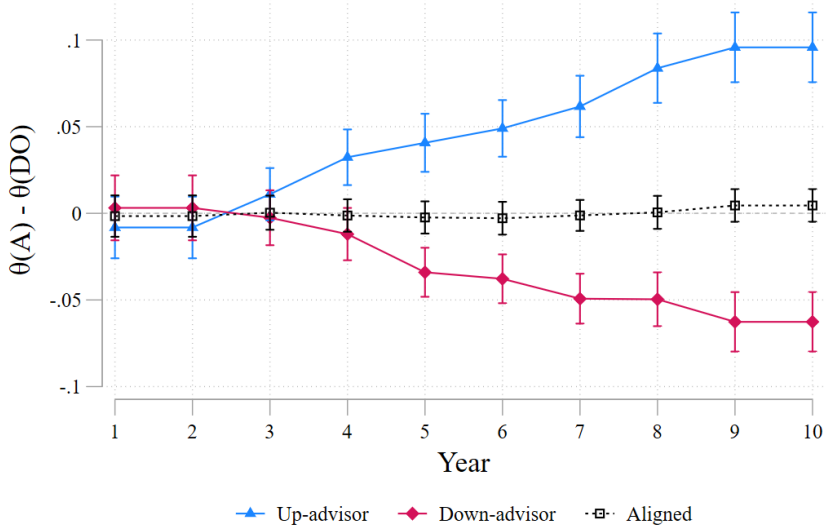
Notes: The figure shows the average history observed by investors in each of the rounds in comparison to the true underlying model generating the data. Hollow dots are the average data observed and the blue line denotes the true data-generating model.

► Back to Section 4.3 (Procedures)

B.2 Additional Results for Section 5.1: Advisor Narrative Construction

Movement-fit tradeoff. The figure below plots the average difference between θ^A and $\theta^{DO}|c^A$, the advisor's data-optimal narrative for the history and her choice of c^A. We observe that the average success probabilities suggested by aligned advisors are non-significantly different from the data-optimum in every year of the historical company data. For misaligned advisors, they are non-significantly different only in the early years of the company's history. In later years, up-advisors exaggerate and down-advisors downplay the company's success probability relative to the data-optimum This suggests that misaligned advisors reduce the narrative's empirical fit to increase the bias in θ^A_{post}, but not in θ^A_{pre}.

Figure B.2: Difference between sent and data-optimal narrative (by advisor type)

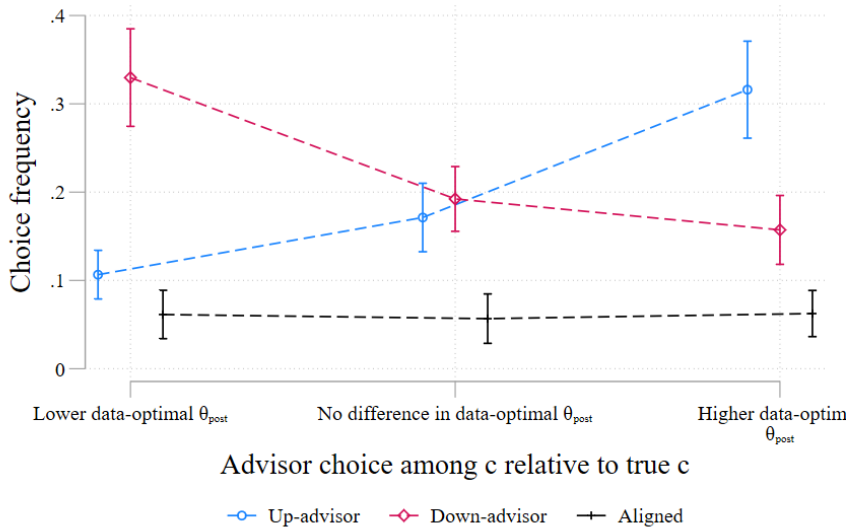


Notes: The figure includes data from advisors who received the BASELINE instructions. Error bars represent 95% confidence intervals that were derived from regressions which cluster standard errors at the advisor level.

Choice of c^A. If advisors strategically adjust c to change the empirical proportion of successes in *post* to be more in line with the communicated θ_{post}, then we should also observe that advisors choose a more advantageous c when they have the possibility to do so (advantageous in the sense that it allows them to justify a θ_{post} that is closer to their desired assessment). We test for this behavior in the following way. For each historical dataset we compare the true year of the structural break (CEO change) to each possible alternative year that advisors could propose as the year of the structural change in their narrative. For each data set and each alternative change-year, we can calculate the data-optimal θ_{post} conditional on that year (i.e., $\theta^{DO}_{post}|c$). This is simply equal to the proportion of successes among all years larger than c. We then classify each of the possible years of CEO change into three different categories: those that justify a higher data-optimal θ_{post} than under truth-telling, those that justify a lower

data-optimal θ_{post} than under truth-telling, and those that induce the same data-optimal θ_{post}.

Figure B.3: Advisor strategies in adjusting c (by advisor type)

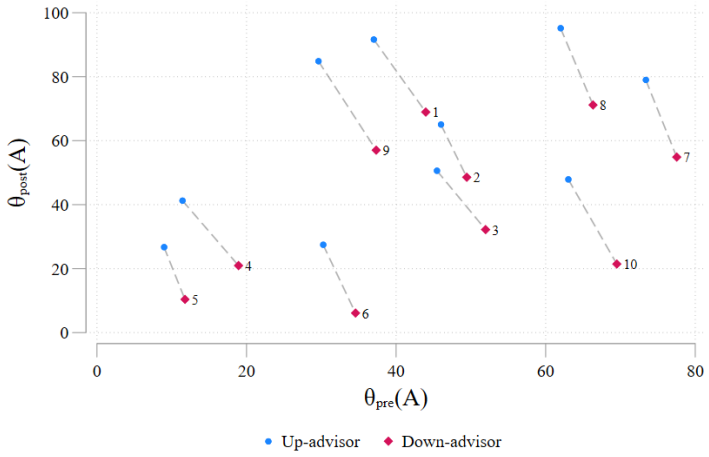


Notes: (i) The figure plots the frequency of advisors choosing to communicate a year that the CEO changed, c, that induces a data-optimal model with either a lower, higher, or the same data-optimal θ_{post} as under the true c, conditional on a c from an alternative category being available, (ii) It shows that up-advisors are more likely to choose a c that induces a higher data-optimal θ_{post}, while down-advisors are more likely to choose a c that induces a lower data-optimal θ_{post} than the c dictated by the true model, (iii) Error bars are 95% confidence intervals derived from regressions which cluster standard errors on the advisor level.

Figure B.3 plots the percentage of advisors that switch from the true year of the structural change to another year that is in one of the three categories. The resulting pattern shows that up-advisors are indeed more likely to deviate to a year that better justifies a higher θ_{post}, while down-advisors are more likely to deviate to a year that better justifies a lower θ_{post}. As expected, no systematic pattern emerges for aligned advisors, who do not have a motive to bias the investor's assessment in a systematic way. Overall, this pattern of behavior suggests that one reason why advisors shift θ_{post} and θ_{pre} in opposite directions is that they are adjusting the year of the CEO change to rationalize a particular θ_{post}, and then adjusting the θ_{pre} to improve the narrative fit. This evidence is consistent with the idea that advisors use all three narrative components at their disposal in trying to construct a convincing narrative.

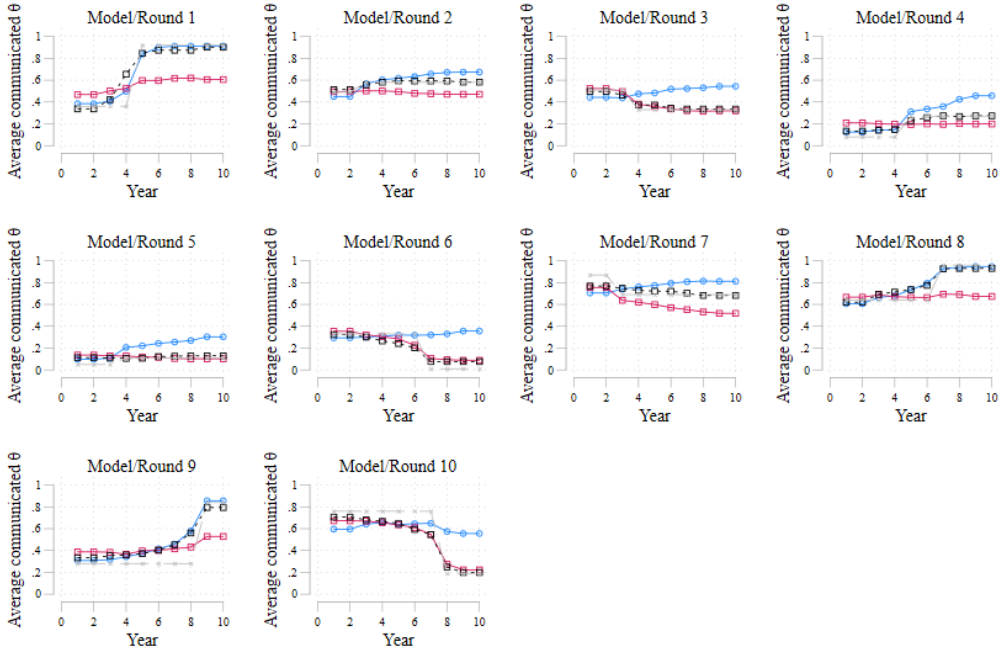
Average narrative by round and advisor type

Figure B.4: Advisor θ_{post} and θ_{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 θ_{post} and θ_{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 θ_{post} and θ_{pre} in opposing directions to construct convincing narratives.

Figure B.5: Average narrative sent (by advisor type and round)



Notes: The figure shows the average message sent by up-advisors (blue), down-advisors (red), and aligned advisors (black). The grey line plots the true model.

► Back to Section 5.1 (Advisor Narrative Construction)

B.3 Additional Results for Section 5.2: Persuasion of Investors

The table below examines the influence of the fit of the advisor's narrative when using a more continuous measure of narrative fit, compared to the one that we used in the main text. 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 4, 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.

Table B.2: Belief updating and narrative fit (with continuous EPI variables)

	(1) $\lvert\theta^{I,1}_{post} - \theta^{I,0}_{post}\rvert$	(2) $\lvert\theta^{I,1}_{post} - \theta^{I,0}_{post}\rvert$	(3) $\lvert\theta^{I,1}_{post} - \theta^{I,0}_{post}\rvert$	(4) $\lvert\theta^{I,1}_{post} - \theta^{I,0}_{post}\rvert$
$\lvert\theta^{I,0}_{post} - \theta^{A}_{post}\rvert$	0.398***	0.464***	0.382***	0.451***
	(0.0412)	(0.0443)	(0.0408)	(0.0433)
EPI^A	-1.953***	-1.556***		
	(0.507)	(0.554)		
$\lvert\theta^{I,0}_{post} - \theta^{A}_{post}\rvert \times EPI^A$	0.121***	0.0907**		
	(0.0374)	(0.0359)		
Misaligned advisor	-0.875	-0.824	-0.745	-0.607
	(0.782)	(0.864)	(0.765)	(0.837)
$(EPI^A - EPI^{I,0})$			-1.987***	-1.894**
			(0.674)	(0.730)
$\lvert\theta^{I,0}_{post} - \theta^{A}_{post}\rvert \times (EPI^A - EPI^{I,0})$			0.129***	0.118***
			(0.0281)	(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
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, $\lvert\theta^{I,1}_{post} - \theta^{I,0}_{post}\rvert$ (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 $\theta^{I,0}_{post} > \theta^{A,1}_{post}$), (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.

► Back to Section 5.2 (Persuasion of Investors)

B.4 Additional Results for Section 5.3.2: Disentangling Data and Narrative

In this section we report further results on the interaction between the company's history and the advisor's narrative. We first show estimation results from a regression that estimates the

equation

$$\theta^{I,1}_{post} = \sum_{t=1}^{10} \beta_t s_t + \sum_{t=3}^{8} \gamma_t s_t \times \mathbb{I}(t > c^A) + \rho + \varepsilon.$$

This regression estimates the investor's assessment as a function of success and failure in each year of the history. In addition it includes interactions between success in a certain year and whether the advisor suggested that this year belongs to the *post* period.[46] Column (1) in Table B.3 reports the estimates of the interaction coefficient. Most of them are positive, which suggests that, when a year is designated to be part of the *post* period, investors weight a success or failure in that year more strongly when making their final assessment. The test for the joint significance of the interaction coefficients is significantly different from zero, as reported further down in the table. Column (2) in addition controls for the advisor's suggested θ^A_{post} and shows that the effect of the suggested *post* period remains significantly different from zero. Columns (3) and (4) estimate identical specifications taking θ^{DO}_{post} of the specific history observed by the investor as the outcome variable. In this placebo specification, we do not find that the interaction coefficients are significantly different from zero.

Table B.3: The effect of history and narrative on assessments and placebo assessments

	(1) $\theta^{I,1}_{post}$	(2) $\theta^{I,1}_{post}$	(3) θ^{DO}_{post}	(4) θ^{DO}_{post}
γ_3	0.318	-1.674	-0.732	-0.693
	(1.446)	(1.235)	(1.133)	(1.140)
γ_4	-0.792	-0.539	-0.133	-0.138
	(1.505)	(1.262)	(1.078)	(1.076)
γ_5	4.551***	3.145***	0.0742	0.101
	(1.355)	(1.068)	(1.054)	(1.060)
γ_6	2.350*	1.014	0.512	0.538
	(1.305)	(1.280)	(1.102)	(1.091)
γ_7	6.586***	3.555**	2.979**	3.038**
	(1.804)	(1.680)	(1.383)	(1.400)
γ_8	5.570	5.078*	0.888	0.898
	(3.453)	(2.809)	(2.513)	(2.518)
θ^A_{post}		0.430***		-0.00832
		(0.0336)		(0.0127)
Dep. var. mean	48.002	48.002	47.727	47.727
H_0: $\gamma_3 = \ldots = \gamma_8 = 0$ p-value	0	.002	.283	.276
Round FE	Yes	Yes	Yes	Yes
Included $\beta_1 - \beta_{10}$	Yes	Yes	Yes	Yes
Observations	1800	1800	1800	1800

Note: Standard errors are clustered at the matching group level, reported in parentheses, (iii) * $p < 0.1$ ** $p < 0.05$ *** $p < 0.01$.

As a further placebo test, Column (1) in Table B.4 presents estimates of the regression

[46]Note that, since years 1-2 never and years 9-10 always belong to the *post* period, only effects of *post* in the years 3-8 are identified.

equation

$$\theta^{I,0}_{post} = \sum_{t=1}^{10} \beta_t s_t + \sum_{t=3}^{8} \gamma_t s_t \times \mathbb{I}(t > c^A) + \rho + \varepsilon$$

using data from InvestorPrior. That is, this specification only differs from the one reported in Column (1) of Table B.3 in that it uses the investor's prior belief as the dependent variable, and not the final assessment. The coefficient estimates of the interaction terms are not jointly different from zero. Column (2) reports estimates of the diff-in-diff specification

$$\theta^{I,d}_{post} = \sum_{t=1}^{10} [\beta_t s_t + \delta_t s_t \times \mathbb{I}(d = 1)] + \sum_{t=3}^{8} \left[\gamma_t s_t \times \mathbb{I}(t > c^A) + \zeta_t s_t \times \mathbb{I}(t > c^A) \times \mathbb{I}(d = 1)\right] + \rho + \rho' \times \mathbb{I}(d = 1) + \varepsilon.$$

In the equation, $d \in \{0, 1\}$ denotes whether the dependent variable is the investor's prior belief ($d = 0$) or final assessment ($d = 1$). The right-hand side includes a number of flexible interactions between the final assessment and the history.[47] Most important in this specifications are the ζ_t-coefficients. They estimate the triple interaction between (i) success in year t, (ii) the year being in *pre* or *post*, as suggested by the advisor, (iii) the investor's belief being the final assessment. The estimation includes observations of prior beliefs and final assessments from InvestorPrior and final assessments from Baseline. The table shows that these interactions are jointly significantly different from zero and mostly positive. They remain positive and marginally jointly significant after controlling for θ^A_{post}.

[47] We include Round×FinalAssessment fixed effects as final assessments are typically influenced by the round's true model, while prior beliefs are not. We include interactions between success in any given year and the final assessments because, compared to final assessments, prior beliefs typically over-extrapolate from the data. These interactions thus pick up an effect which would otherwise wrongly be picked up by the success-post period-final assessment interaction terms.

Table B.4: The effect of history and narrative on prior beliefs and assessments

	(1) $\theta^{I,0}_{post}$	(2) $\theta^{I,0}_{post}, \theta^{I,1}_{post}$	(3) $\theta^{I,0}_{post}, \theta^{I,1}_{post}$
γ_3	-0.249	-0.249	-0.262
	(2.361)	(2.319)	(2.312)
γ_4	2.296	2.296	2.288
	(2.708)	(2.659)	(2.659)
γ_5	1.468	1.468	1.458
	(2.108)	(2.070)	(2.073)
γ_6	-0.956	-0.956	-0.974
	(2.275)	(2.234)	(2.204)
γ_7	-0.0589	-0.0589	-0.0890
	(2.479)	(2.434)	(2.433)
γ_8	-2.450	-2.450	-2.462
	(2.321)	(2.279)	(2.302)
ζ_3		1.197	-0.472
		(2.241)	(2.172)
ζ_4		-1.496	-1.635
		(2.446)	(2.385)
ζ_5		1.876	0.687
		(2.007)	(1.940)
ζ_6		3.336	1.953
		(2.077)	(2.118)
ζ_7		5.285*	2.330
		(2.776)	(2.645)
ζ_8		6.746**	6.015**
		(3.123)	(2.793)
θ^{A}_{post}			0.00402
			(0.0280)
Assessment=1$\times\theta^{A}_{post}$			0.397***
			(0.0322)
Dep. var. mean	47.851	47.987	47.987
H_0: $\gamma_3 = \ldots = \gamma_8 = 0$ p-value	.606	.58	.586
H_0: $\zeta_3 = \ldots = \zeta_8 = 0$ p-value	-	.002	.061
Round FE	Yes	No	No
Round × Assessment FE	No	Yes	Yes
Included $\beta_1 - \beta_{10}$	Yes	Yes	Yes
Included $\delta_1 - \delta_{10}$	No	Yes	Yes
Observations	900	3600	3600

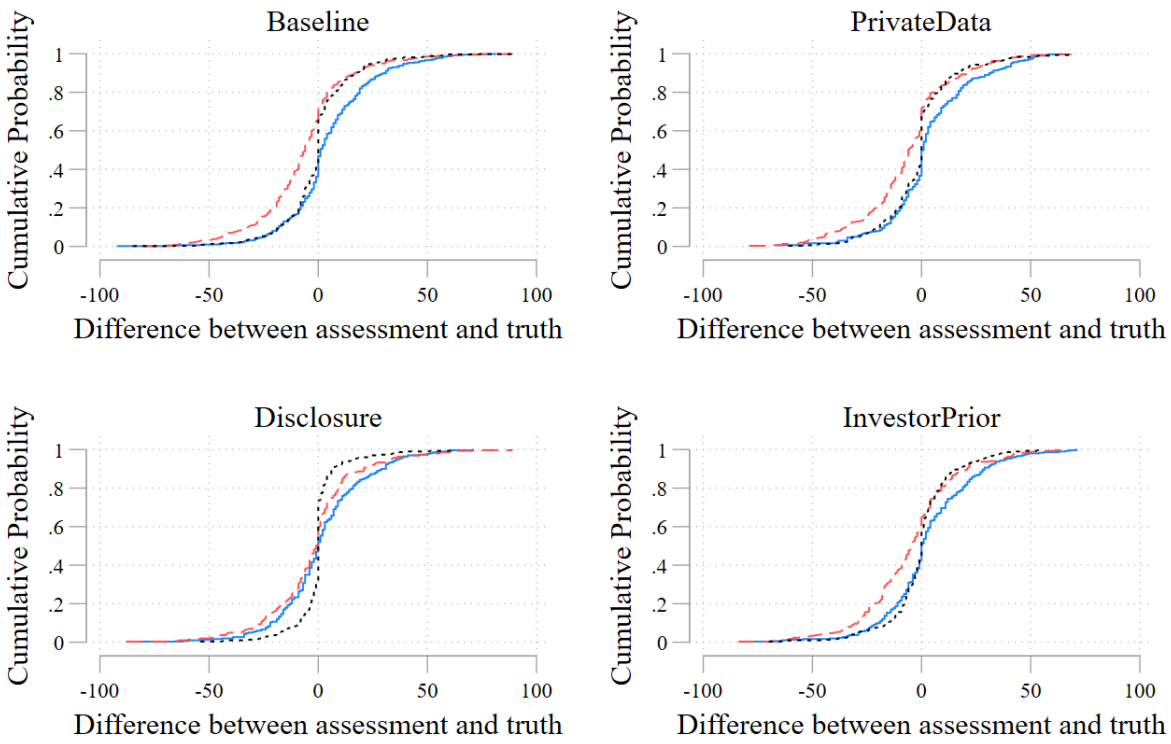
Note: Standard errors are clustered at the matching group level, reported in parentheses, (iii) * $p < 0.1$ ** $p < 0.05$ *** $p < 0.01$.

► Back to Section 5.3.2 (Disentangling Data and Narrative)

B.5 Additional Results for Section 5.4: Evaluating Potential Protective Interventions

The following figure reproduces Figure 5 for all treatments.

Figure B.6: Difference between θ^I_{post} and θ^T_{post} (by treatment and advisor type).



Notes: (i) The figure plots the cdf of the difference between the investor's belief and the truth, $\theta^I_{post} - \theta^T_{post}$, 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.

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 B.5. 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.

Table B.5: Evaluating the impact of interventions (distance to advisor message)

	DISCLOSURE $\lvert\theta^I_{post} - \theta^A_{post}\rvert$		INVESTORPRIOR $\lvert\theta^I_{post} - \theta^A_{post}\rvert$		PRIVATEDATA $\lvert\theta^I_{post} - \theta^A_{post}\rvert$	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Treatment	2.038*	2.488	1.730	0.914	3.020***	-0.442
	(1.088)	(1.619)	(1.044)	(1.158)	(1.090)	(1.179)
Advisor lied		3.855***		3.710***		3.685***
		(0.911)		(0.921)		(0.916)
Treatment × Advisor lied		-0.521		1.227		4.696***
		(1.825)		(1.681)		(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 θ^A_{post} and the corresponding belief held by the investor θ^I_{post}, (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.

Aligned advisors. Turning to the aligned advisors, Table B.6 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.

Table B.6: Evaluating the impact of interventions (aligned advisor)

	Disclosure $\lvert\theta^I_{post} - \theta^T_{post}\rvert$ (1a)	Disclosure $\lvert\theta^I_{post} - \theta^A_{post}\rvert$ (1b)	InvestorPrior $\lvert\theta^I_{post} - \theta^T_{post}\rvert$ (2a)	InvestorPrior $\lvert\theta^I_{post} - \theta^A_{post}\rvert$ (2b)	PrivateData $\lvert\theta^I_{post} - \theta^T_{post}\rvert$ (3a)	PrivateData $\lvert\theta^I_{post} - \theta^A_{post}\rvert$ (3b)
Treatment	-4.597***	-5.075***	-0.0800	-0.278	0.530	0.632
	(0.994)	(0.934)	(0.972)	(1.029)	(1.110)	(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 θ^T_{post} parameter and the corresponding belief held by the investor θ^I_{post}, (ii) The dependent variable in the (*b) columns is the distance between the advisor's message θ^A_{post} and the corresponding belief held by the investor θ^I_{post}, (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.

Comparing how investors react to messages in Baseline and Disclosure. To compare how investors in both treatments react to receiving a narrative with a certain θ_{post} value, we order θ_{post} values according to their percentile in the Bayesian prior belief conditional on a history.[48] Higher percentiles correspond to higher θ_{post} values for a given history, but, comparing two different histories, this must not be the case.

Figure B.7: Scatter plot of messages and assessments in *Baseline* and *Disclosure*

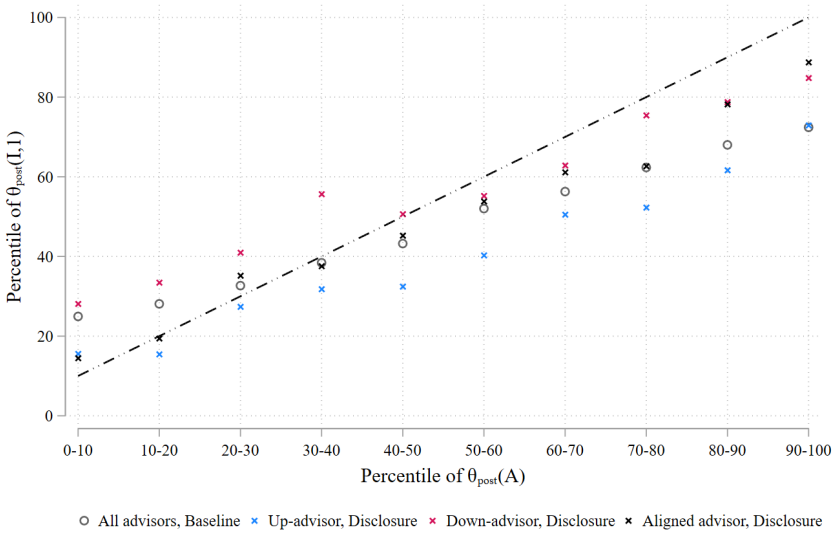


Figure B.7 plots the percentile of the advisor's θ_{post} against the investor's assessment of θ_{post}. Suppose the investor would always adopt the advisor's narrative. Then, the dots in the figure would line up on dashed, 45 degree line. We observe that investors react less strongly

[48]Given the information about the underlying parameter distribution, the Bayesian posterior about θ_{post} can be calculated as being a mixture of Beta distributions, where each Beta distribution is characterized by the number of success and failure in *post* for a given structural break c. See Appendix F for the formal derivation of this posterior and an example.

to messages than this. We also observe that, in *Skepticism*, Investors on average shade an up-advisor's message downwards and a down-advisor's message upwards. Messages of aligned advisors induce assessments that are closer to the 45 degree line than in *Baseline*.

► Back to Section 5.4 (Evaluating Potential Protective Interventions)

B.6 Additional Results for Section 5.6: Consistency of the Experimental Data with Nash Equilibrium

The Figure below plots the distance between the investor's assessment and the advisor message, $|\theta^{I,1}_{post} - \theta^{A}_{post}|$, against difference between the θ^{A}_{post} sent by the advisor and the lower (left panel) and upper (right panel) Nash equilibrium thresholds. While Nash equilibrium would predict that the distance is zero if θ^{A}_{post} is within the interval predicted by Nash equilibrium (i.e. for positive x-axis values in the left panel and for negative x-axis values in the right panel) and positive otherwise, there is only a modest discontinuity around the threshold. This is confirmed in regressions that account for round fixed-effects in a regression-discontinuity framework. As shown in the regressions reported in Table B.7, beliefs of investors are modestly closer to the message if the advisor's θ^{A}_{post} is within the Nash equilibrium interval relative to it being outside the interval. However, the size of this effect is not always significant and is sensitive to the chosen bandwidth around the threshold.

Figure B.8: Distance to the message against difference of message and Nash equilibrium thresholds

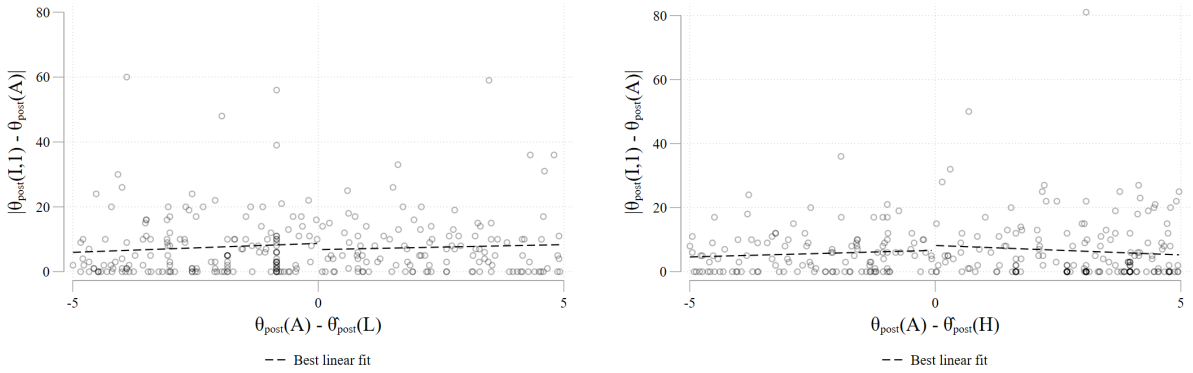


Table B.7: Regression discontinuity estimates of the change in the distance to the message around the NE thresholds

	$\lvert\theta^{I,1}_{post} - \theta^{A}_{post}\rvert$	$\lvert\theta^{I,1}_{post} - \theta^{A}_{post}\rvert$	$\lvert\theta^{I,1}_{post} - \theta^{A}_{post}\rvert$	$\lvert\theta^{I,1}_{post} - \theta^{A}_{post}\rvert$
θ^{A}_{post} is within NE interval	-2.218	-1.958**	-1.009	-1.075
	(1.397)	(0.957)	(0.716)	(0.652)
Misaligned advisor = 1	-0.994	0.111	0.414	-0.434
	(1.330)	(0.885)	(0.711)	(0.620)
Inclusion crit.: dist. to threshold	$\theta^{A}_{post} \pm 1$	$\theta^{A}_{post} \pm 3$	$\theta^{A}_{post} \pm 5$	$\theta^{A}_{post} \pm 10$
Dependent variable mean	7.692	7.134	6.885	7.012
Round FE	Yes	Yes	Yes	Yes
Observations	120	320	549	900

Note: (i) The sample contains data from all investors in BASELINE and where the θ^{A}_{post} is in a minimum distance around the threshold (ii) Standard errors are clustered at the matching group level, reported in parentheses, (iii) ** $p < 0.05$.

► Back to Section 5.6 (Consistency of the Experimental Data with Nash Equilibrium)

C Additional pre-registered hypotheses and results

C.1 Narrative construction of aligned advisors and the role of truth-telling preferences

Balancing persuasiveness against the truth (aligned advisors). We 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, θ^T_{post}, 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 θ^{DO}_{post} is far from the true value θ^T_{post}—aligned advisors will send a message that contains a θ^A_{post} value that is further from θ^{DO}_{post}. In other words, we hypothesize that the average aligned advisor will follow a strategy that involves sending a θ^A_{post} that is a weighted average of the truth, θ^T_{post}, and the data-optimal narrative, θ^{DO}_{post}.

Hypothesis 7a. *[PR.7a] The distance between the data-optimal model and the aligned advisor's message, $|\theta^A_{post} - \theta^{DO}_{post}|$, increases in the distance between the truth and the data optimal report $|\theta^T_{post} - \theta^{DO}_{post}|$.*

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 7b. *[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 $|\theta^T_{post} - \theta^{DO}_{post}|$ than in the aligned advisor's report.*

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

$$|\theta^A_{post} - \theta^{DO}_{post}| = \beta_0 + \beta_1 \mathbb{I}(\text{Misaligned}) + (\beta_2 + \beta_3 \mathbb{I}(\text{Misaligned})) \cdot |\theta^T_{post} - \theta^{DO}_{post}| + \rho_r + \varepsilon$$

In the equation above, $\mathbb{I}$(Misaligned) is an indicator function which takes a value of 1 if the advisor's incentives are misaligned, ρ_r are round fixed effects and ε is an error term.[49] Table C.1 reports the results. We test Hypothesis 7a by examining β_2. Since β_2 is statistically greater than 0, we find evidence in support of Hypothesis 7a. 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.

Table C.1: The influence of the truth on advisor narratives

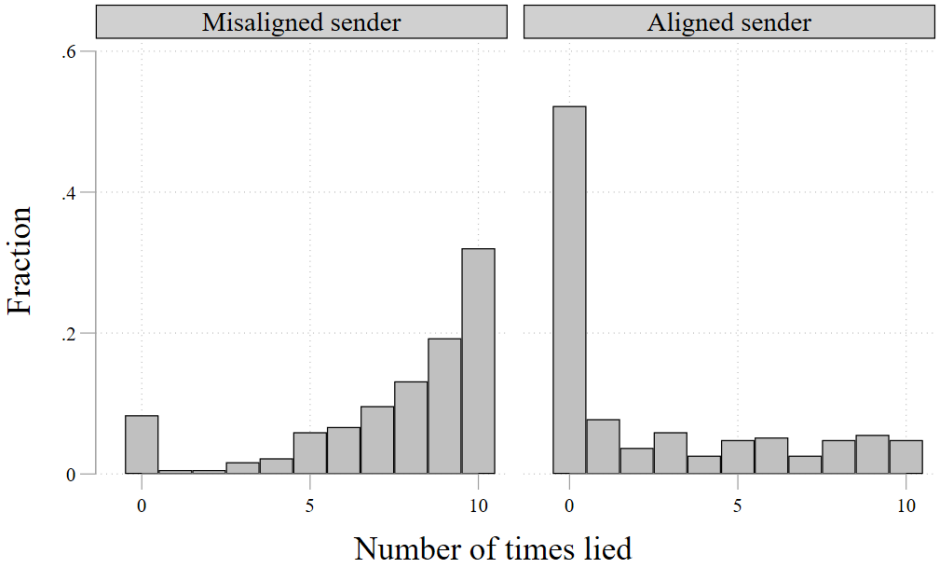
	$\lvert\theta^A_{post} - \theta^{DO}_{post}\rvert$
β_1: Misaligned advisor = 1	13.33***
	(0.864)
β_2: $\lvert\theta^T_{post} - \theta^{DO}_{post}\rvert$	0.974***
	(0.0149)
β_3: Misaligned advisor × $\lvert\theta^T_{post} - \theta^{DO}_{post}\rvert$	-0.411***
	(0.0322)
Dependent variable mean	22.169
$\beta_2 + \beta_3 = 0$	.001
Round FE	Yes
Observations	3600

(i) The dependent variable is the distance between the advisor's report, θ^A_{post}, and the true value θ^T_{post}, (ii) The sample contains data from all advisors who received the BASELINE instructions, (iii) Standard errors are clustered at the advisor level, reported in parentheses, (iv) There are 360 clusters, (v) For each advisor, we have 10 observations—one for each round, (vi) *** $p < 0.01$.

In support of Hypothesis 7b, 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, as displayed in Figure C.1. 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.

[49] Since the true model is held constant within each round of the experiment, the ρ_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.

Figure C.1: Distribution of lying across ten rounds (by advisor type)



Notes: The figure includes data from all advisors who received the BASELINE instructions. A message is defined to be a lie when at least one parameter value is not equal to the truth.

C.2 Features of the dataset that make it more difficult to persuade the investor

The influence of alternative available models on receiver trust: Here, we introduce a sub-hypothesis that checks for a potential force moderating the relationship between the message's EPI and the receiver's assessment: if there exist different models that fit the observed data comparatively well, does this make it more difficult to persuade the receiver to adapt the sender's model compared to the case where there is a single salient data-optimal model?

We study the impact of the availability of alternative models by examining whether the shape of the EPI function, taken across all possible values of θ_{post}, affects the distance between the sender's message and the receiver's assessment, $D^S(\theta^R_{post})$. The EPI function is single-peaked in cases where the data provides a relatively salient data-optimal model but has multiple peaks when the data provides room for multiple competing explanations. We hypothesize that, if the history of outcomes can be equally well explained by different models, the receiver is less easily swayed by the sender's model (assuming that the receiver has reason to believe that there is at least some chance that the sender does not have aligned incentives, as is the case in our BASELINE treatment). The rationale behind this hypothesis is that when the EPI has multiple peaks, the receiver can more easily entertain alternative models that explain the data similarly well. Therefore, we conjecture that the distance between the sender's message and the receiver's assessment is higher if, among all possible values of θ_{post}, the EPI has multiple

local optima.[50] To adjust for possible changes in the sender's message quality across different histories, we condition the hypothesis on the value of the EPI evaluated at the sender's model.

Hypothesis 8 (PR.5b)**.** *Conditional on the value of the EPI evaluated at the sender's model, the distance between the sender's message and the receiver's assessment is smaller if the EPI has a single global optimum than if it has multiple local optima.*

We test for this hypothesis by running a regression of the following form using data from receivers in the Baseline treatment:

$$D^S(\theta^R_{post}) = \beta_0 + \beta_1 \text{EPI}(c^S, \theta^S_{pre}, \theta^S_{post}|h) + \beta_2 \mathbb{I}(\text{EPI has multiple peaks}) + \alpha + \rho_r + \varepsilon.$$

The table below shows results using the Baseline data. They indicate that investors indeed report beliefs that are further away from the advisor's narrative if the EPI has multiple optima.

	(1) $\lvert\theta^{I,1}_{post} - \theta^A_{post}\rvert$
Advisor message fit (EPI)	-15.45***
	(1.968)
$\mathbb{I}$(EPI has multiple optima)	4.812***
	(1.564)
Misaligned advisor = 1	0.807
	(0.664)
Dependent variable mean	11.085
Round FE	Yes
Observations	1800

D Details on the Structural Estimation and Robustness Checks

A theoretical rationale for the updating rule. We generalize the narrative adoption rule as described in Equation (1) by two factors. First, we assume that investors' perceptions of the empirical fit is noisy. That is, they might not be perfectly accurate in assessing the empirical fit of a given narrative. Second, we assume that investors are to some extend credulous, in the sense that they do not require the narrative to explain the data better than the default model does, but simply require that it does not explain the data significantly worse than the

[50] Another way to think about this is that, if the log likelihood function of the model for a given history is relatively flat in θ_{post}, the sender is less swayed by the receiver's message, even if the communicated model has a high EPI because alternative models exist that also have a high EPI. We proxy flatness of the log likelihood function by distinguishing between flat (multiple peaked) and non-flat (single peaked) functions.

default model. We also take the EPI as an empirical fit measure to account for the fact that the absolute likelihood values differ by the underlying data set.[51] Under these assumptions, the investor adopts the advisor's narrative whenever

$$\text{EPI}(m^A) + \tilde{\kappa} + \varepsilon^A \geq \text{EPI}(m^{I,0}) + \varepsilon^{I,0}.$$

Above, $\tilde{\kappa}$ denotes the investor's degree of credulity. Note that $-\tilde{\kappa}$ can be interpreted as the degree of skepticism. The error terms ε^A and $\varepsilon^{I,0}$ measure the noise in the ways the investor perceives the empirical fit of the narrative and the default model. We assume that ε^A and $\varepsilon^{I,0}$ are independently distributed according to the Gumbel distribution, with mean 0 and scale parameter $1/\lambda$. This is very similar to how Froeb et al. (2016) model noisy model selection in their theoretical model of persuasion. Under these assumptions on the noise term, the probability of adoption becomes equal to

$$\Pr(\text{adopt narrative}|\tilde{\kappa}, \lambda, m^{I,0}, m^A) = \frac{\exp\{\lambda(\tilde{\kappa} + \text{EPI}(m^A))\}}{\exp\{\lambda(\tilde{\kappa} + \text{EPI}(m^A))\} + \exp\{\lambda \text{EPI}(m^{I,0})\}}.$$

Defining $\kappa \equiv \lambda\tilde{\kappa}$ and dividing the numerator and denominator by $\exp\{\lambda\text{EPI}(m^{I,0})\}$, we arrive at

$$\Pr(\text{adopt narrative}|\kappa, \lambda, m^{I,0}, m^A) = \frac{\exp\{\kappa + \lambda\Delta\text{EPI}(m^{I,0}, m^A))\}}{\exp\{\kappa + \lambda\Delta\text{EPI}(m^{I,0}, m^A))\} + 1}.$$

Given the default model and the narrative, the investor's expected assessment then becomes

$$\Pr(\text{adopt narrative}|\kappa, \lambda, m^{I,0}, m^A)\theta^A_{post} + (1 - \Pr(\text{adopt narrative}|\kappa, \lambda, m^{I,0}, m^A))\theta^{I,0}_{post}.$$

This is equal to the updating rule in that we estimate in the main text.

Estimation details. We can relatively straightforwardly estimate the κ and λ parameter values that best describe updating in the InvestorPrior treatment using a minimum distance estimator. In particular, we can find the parameter values that minimize

$$\sum_i^N \left(\theta^{I,1}_{post,i} - \hat{\theta}^{I,1}_{post,i}(\kappa, \lambda; m^{I,0}_i, m^A_i, h_i)\right)^2,$$

where $\hat{\theta}_{post,i}$ is as described in Equation (3). The estimation problem is slightly more challenging in the remaining treatments which did not elicit the default model. Take the expectation of Equation (3) with respect to $m^{I,0}$;

$$\mathbb{E}_{m^{I,0}}[\hat{\theta}^{I,1}_{post}(\kappa, \lambda; m^{I,0}, m^A, h, \tau)|h] =$$
$$= \int p(\kappa, \lambda; m^A, m^{I,0}, h)\theta^A_{post} + (1 - p(\kappa, \lambda; m^A, m^{I,0}, h))\theta^{I,O}_{post} f(m^{I,0}|h, \tau)\, \mathrm{d}m^{I,0}.$$

[51] For a given history, EPI(m) is proportional to Pr($h|m$).

In the equation above, τ is a treatment indicator. We can in principle estimate κ and λ if we derive an estimate for the distribution of default models, $f(m^{I,0}|h,\tau)$. In order to arrive at such an estimate, we make two observations and two assumptions. First, observe that we only need to know the joint distribution of the default model's EPI and $\theta^{I,0}_{post}$. We can rewrite the expectation term to reflect this observation as

$$\mathbb{E}_{m^{I,0}}[\hat{\theta}^{I,1}_{post}(\kappa,\lambda;m^{I,0},m^A,h,\tau)|h] = \\ = \int p(\kappa,\lambda;m^A,m^{I,0},h)\theta^A_{post} + (1-p(\kappa,\lambda;m^A,m^{I,0},h))\theta^{I,O}_{post} n(\theta^{I,0}_{post}|\text{EPI}^{I,0},h) g(\text{EPI}^{I,0}|h,\tau)\,\text{dEPI}^{I,0}.$$

In the equation above, g is the marginal distribution of the default model EPI and n is the distribution of $\theta^{I,0}_{post}$ conditional on $\text{EPI}^{I,0}$. Second, if we assume that the distribution of the default model's EPI is independent of the history and treatment, we can estimate it using the distribution of default model EPIs in InvestorPrior. Third, conditional on $\text{EPI}^{I,0}$ and h, we can directly calculate the distribution of $\theta^{I,0}_{post}$, by identifying for a given dataset which default models would induce a certain level of $\text{EPI}^{I,0}$. This suggests that we can derive estimates for κ and λ by following three steps: First, approximate $g(\text{EPI}^{I,0})$ using data from InvestorPrior. We do this by rounding all observed default model EPIs to two digits. This gives us 100 bins of default models with EPI values $0, 0.01, \ldots, 1$. For each bin, we calculate the fraction of observed default model EPIs that fall into that bin. This is our estimate $\hat{g}(\text{EPI}^{I,0})$. Second, for each bin of the default model EPI and for each history, we calculate the conditional expectation $\mathbb{E}[\theta^{I,0}|h,\text{EPI}^{I,0}]$. Essentially, we do this by calculating the EPIs of all possible combinations $(c,\theta_{pre},\theta_{post})$ and then discretize these EPIs into the same 100 bins. For each bin and history, we then take the average of θ_{post} of the models that fall into this bin. We use these estimates in a third step to minimize

$$\sum_i^N \left[\theta^{I,1}_{post,i} - \sum_{j\in\{0,0.01,\ldots,1\}} \left(p(\kappa,\lambda;m^A,\text{EPI}^{I,0}_j,h)\theta^A_{post} + (1-p(\kappa,\lambda;m^A,\text{EPI}^{I,0}_j,h))\mathbb{E}[\theta^{I,0}|h,\text{EPI}^{I,0}]\hat{g}(\text{EPI}^{I,0}_j) \right) \right]^2$$

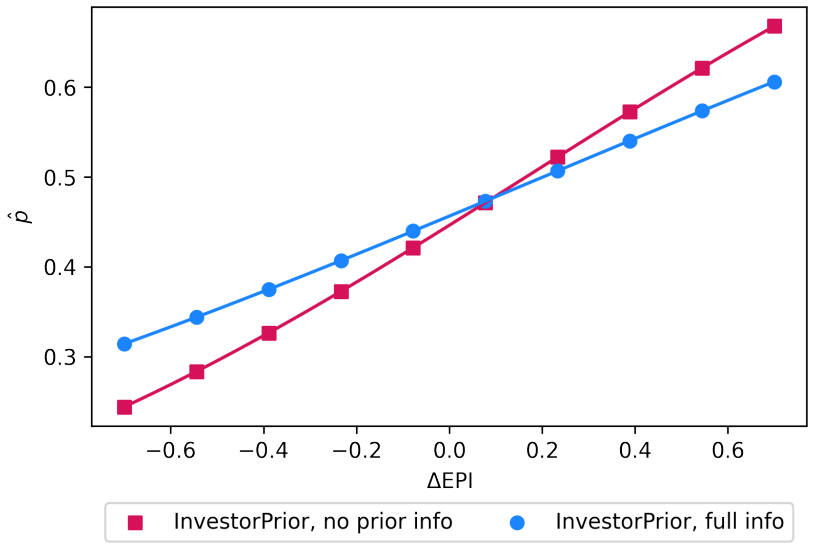
with respect to κ and λ. In all reported estimation results, we use bootstrapping to first draw a sample from InvestorPrior to approximate $\hat{g}$ and then draw a sample from one treatment to estimate κ and λ.

Comparing different estimation results for InvestorPrior. The table below presents parameter estimation results for InvestorPrior that were either estimated using the full information on individual default models or the integrating out-techinque described above. The Wald test of joint equality of the parameters cannot be rejected. Furthermore, both model estimates predict a similar shape of the weighting function. As Figure D.1 shows, while the predicted slopes of the weighting function are different, the predicted weights never differ by more than 10 p.p. and for a large part of the range they differ by less than that.

Table D.1: Comparing updating parameters for InvestorPrior

	(1)	(2)
κ	-0.175*	-0.216
	(0.096)	(0.182)
φ	0.866***	1.308**
	(0.242)	(0.611)
Approach	Full information	Expectation over $m^{I,0}$
Wald test p-value	–	0.483
MSE	131.308	192.036
Observations	900	900

Figure D.1: Predicted weights investors put on the narrative in InvestorPrior (by estimation techinque)



► Back to Section 5.5 (Quantifying How Investors Update)

E Proofs

E.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.

The log likelihood function is equal to

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

E.2 Proof of Proposition 1

We take the first-order condition of the expected utility function specified in Equation (2) with respect to θ^A_{post};

$$\begin{aligned}\frac{\partial E[U(\theta^I_{post},\varphi)|m^{DO}]}{\partial \theta^A_{post}} = g(\ell(m^{DO}))\frac{\partial \ell(m^{DO})}{\partial \theta_{post}}\Big(U(\theta^{DO}_{post},\varphi) - \mathbb{E}[U(\theta^{I,0}_{post},\varphi)|\ell(m^{DO}) < \ell(m^{I,0})]\Big) \\ + G(\ell(m^{DO}))\frac{\partial U(\theta^{DO}_{post},\varphi)}{\partial \theta^A_{post}} + (1 - G(\ell(m^{DO})))\frac{\partial \mathbb{E}[U(\theta^{I,0}_{post},\varphi)|\ell(m^A) < \ell(m^{I,0})]}{\partial \theta^A_{post}}.\end{aligned}$$

Now, because it is evaluated at the data-optimal model, $G(\ell(m^{DO})) = 1$ and $\frac{\partial \ell(m^{DO})}{\partial \theta_{post}} = 0$. Therefore, the derivative simplifies to

$$\frac{\partial E[U(\theta^I_{post},\varphi)|m^{DO}]}{\partial \theta^A_{post}} = \frac{\partial U(\theta^{DO}_{post},\varphi)}{\partial \theta^A_{post}},$$

which is nonzero whenever $\varphi \neq \theta^{DO}_{post}$. Whenever this is the case, the advisor has an incentive to marginally adjust θ^{DO}_{post} away from the data-optimal value.

E.3 Proof of Proposition 2

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

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

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, $U(\theta^A_{post}, \varphi)$, with respect to θ^A_{post}. This in turn is equal to minimizing $(\varphi - \theta^A_{post})^2$.

E.4 Proof of Proposition 3

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

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

Collect models where c, θ_{pre} are the conditional likelihood maximizers for a given θ_{post} (i.e., all models with $(\hat{c}(\theta_{post}), \hat{\theta}_{pre}(\theta_{post}), \theta_{post})$), in a set $\mathcal{C}$.

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 θ_{pre} that, if combined with any θ_{post}, lead to a message fit between minus infinity and the value of the conditional log likelihood function evaluated at θ_{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

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

Observe that if $k_{pre} > 0$, the limit $\lim_{t\to 0} \ell((\hat{c}(\theta_{post}), t, \theta_{post})) \to -\infty$ and that if $f_{pre} > 0$, the limit $\lim_{t\to 1} \ell((\hat{c}(\theta_{post}), t, \theta_{post})) \to -\infty$. 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 θ^*_{post} has to be equal to φ.

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 θ^*_{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 $\ell(m') = \ell(m^*)$ as long as $\theta^*_{post} \notin \mathcal{C}$. By Proposition 2, the advisor prefers message m' to message m^*, which contradicts the initial statement.

We proceed with Claim 3 which shows that, if θ_{post} is fixed at φ, 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 ΔG 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

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

The inequality above follows from the fact that the investor's prior has full support on $\mathcal{M}$, 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 likelihood, which implies that $U(\varphi, \varphi) > \mathbb{E}[U(\theta^{I,0}_{post}, \varphi)|\ell^{I,0} \in (\ell'', \ell')]$. Therefore, $\mathbb{E}[U(\theta^I_{post}, \varphi)|m'] > \mathbb{E}[U(\theta^I_{post}, \varphi)|m'']$, 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 θ_{post}-parameter value is equal to φ. However Claim 3 implies that, among all models in $\mathcal{M}$ with $\theta_{post} = \varphi$, the advisor most prefers the model that is also in $\mathcal{C}$, which implies Claim 4.

E.5 Proofs of proposition 4

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

E.5.1 Proof of part (i)

We will show under which conditions a cutoff $c' < c^{DO}$ can be on the up-advisor's likelihood frontier.

We will compare two potential messages $m' = (c', \theta'_{pre}, \theta_{post})$ and $m'' = (c^{DO}, \theta^{DO}_{pre}, \theta_{post})$. In message m', θ'_{pre} maximizes the likelihood conditional on c'. Therfore, both messages choose the likelihood maximizer of θ_{pre} conditional on c' or c^{DO} and hold θ_{post} fixed. For simplicity, we will use $\theta^{DO}_{pre} \equiv \theta''_{pre}$. We will also use the convention that

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

and will denote differences in the number of successes in *post* under the structural change parameters c' and c^{DO} by $\Delta k = k'_{post} - k''_{post}$ and $\Delta f = f'_{post} - f''_{post}$. Define a function that returns the log likelihood difference between messages m' and m'' for a given θ_{post} by

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

In the proof we will consider under which conditions $\Delta\ell(\theta_{post})$ can be positive. This is a necessary condition for c' to be on the likelihood frontier and therefore a necessary condition for the advisor choosing c' as part of the optimal message.

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

$$\Delta\ell'(\theta_{post}) = \frac{\Delta k}{\theta_{post}} - \frac{\Delta f}{1-\theta_{post}}. \tag{4}$$

Furthermore, as θ_{post} becomes large,

$$\begin{aligned}
\lim_{\theta_{post}\to 1} \Delta\ell(\theta_{post}) = & \Delta k(\lim_{\theta_{post}\to 1} ln(\theta_{post}) - ln(\theta'_{pre})) + \Delta f(\lim_{\theta_{post}\to 1} ln(1-\theta_{post}) - ln(1-\theta'_{pre})) + \kappa \\
= & -\Delta k ln(\theta'_{pre}) + \Delta f(\lim_{\theta_{post}\to 1} ln(1-\theta_{post}) - ln(1-\theta'_{pre})) + \kappa
\end{aligned} \tag{5}$$

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

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

When does this condition hold? Define a function

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

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

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

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

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 $\theta_{post} \to 1$, $\Delta\ell(\theta_{post}) > 0$ (see Equation (5) and the discussion afterwards). Furthermore, the derivative in Equation (4) shows that $\Delta\ell$ is strictly increasing in θ_{post}. There is thus one critical value $\theta^C_{post} > \theta^{DO}_{post}$ so that $\Delta\ell(\theta_{post}) \geq 0$ whenever $\theta_{post} \geq \theta^C_{post}$.

Case 2: $\Delta k = 0$, $\Delta f > 0$. As $\theta_{post} \to 1$, $\Delta\ell(\theta_{post}) < 0$ (see Equation (5) and the discussion afterwards). Furthermore, the derivative in Equation (4) shows that $\Delta\ell$ is strictly decreasing in θ_{post}. As $\theta^*_{post} \geq \theta^{DO}_{post}$ and $\Delta\ell(\theta^{DO}_{post}) < 0$, c' can never be on the likelihood frontier of the up-advisor.

Case 3: $\Delta k > 0$, $\Delta f > 0$. As $\theta_{post} \to 1$, $\Delta\ell(\theta_{post}) < 0$ (see Equation (5) and the discussion afterwards). Furthermore, the derivative in Equation (4) shows that $\Delta\ell$ is first increasing and then decreasing in θ_{post}. The derivative changes its sign exactly once at the point

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

Rearranging, we find that

$$\theta^0_{post} > \theta^{DO}_{post} \Longleftrightarrow \frac{k'_{post}}{1 - c'} > \frac{k''_{post}}{1 - c^{DO}}.$$

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

In summary, we find that $\Delta\ell(\theta^{DO}_{post})$ can be positive only in cases 1 or 3 and only if $k'_{post}/(1 - c') > \theta^{DO}_{post}$.

E.5.2 Proof of part (iii)

We will show under which conditions a cutoff $c' > c^{DO}$ can be on the up-advisor's likelihood frontier.

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

Case 1: $\Delta k < 0$, $\Delta f = 0$. As $\theta_{post} \to 1$, $\Delta\ell(\theta_{post}) < 0$ (see Equation (5) and the discussion afterwards). Furthermore, the derivative in Equation (4) shows that $\Delta\ell$ is strictly decreasing in θ_{post}. As $\theta^*_{post} \geq \theta^{DO}_{post}$ and $\Delta\ell(\theta^{DO}_{post}) < 0$, c' is never on the likelihood frontier.

Case 2: $\Delta k = 0$, $\Delta f < 0$. As $\theta_{post} \to 1$, $\Delta\ell(\theta_{post}) > 0$ (see Equation (5) and the discussion afterwards). Furthermore, the derivative in Equation (4) shows that $\Delta\ell$ is strictly increasing in θ_{post}. There is thus one critical value $\theta^C_{post} > \theta^{DO}_{post}$ so that $\Delta\ell(\theta_{post}) \geq 0$ whenever $\theta_{post} \geq \theta^C_{post}$.

Case 3: $\Delta k < 0$, $\Delta f < 0$. As $\theta_{post} \to 1$, $\Delta\ell(\theta_{post}) > 0$ (see Equation (5) and the discussion afterwards). Furthermore, the derivative in Equation (4) shows that $\Delta\ell$ is first decreasing and then increasing in θ_{post}. There is thus one critical value $\theta^C_{post} > \theta^{DO}_{post}$ so that $\Delta\ell(\theta_{post}) \geq 0$ whenever $\theta_{post} \geq \theta^C_{post}$.

In summary, we find that c' can be on the likelihood frontier only if $\Delta f < 0$.

F Discussion of the Nash equilibrium

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

F.1 Setup

Consider a game between a advisor and a investor. There is an unknown state of the world $\theta_{post} \in \Theta = [0,1]$. This state is distributed with full support on Θ according to a commonly known prior distribution $g(\theta_{post})$. 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 $\theta^I_{post} \in A = [0,1]$ of the state of the world. Payoffs of both advisor and investor depend on a scoring rule. The advisor's utility function is

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

where φ varies with the advisor's type. Its value is equal to θ_{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

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

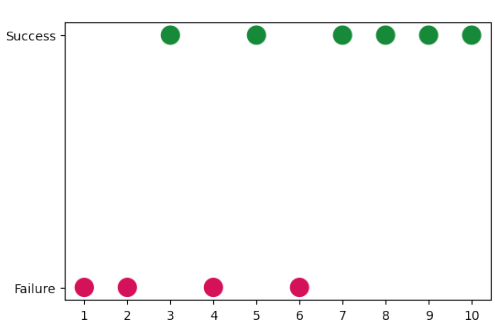
The investor maximizes utility if he makes an accurate assessment.

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 $g(\theta_{post})$. Formally, upon seeing the data, the investor can form a Bayesian posterior belief which is equal to

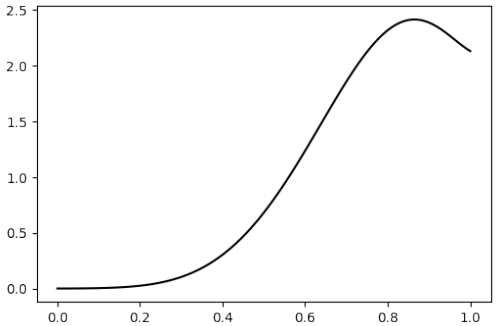
$$g(\theta_{post}) = \sum_{c=2}^{8} \frac{\int_0^1 \mathcal{L}((c, \theta_{pre}, \theta_{post}))\, \mathrm{d}\theta_{pre}}{\sum_{c=2}^{8} \int_0^1 \int_0^1 \mathcal{L}((c, \theta_{pre}, \theta_{post}))\, \mathrm{d}\theta_{pre} \mathrm{d}\theta}. \tag{6}$$

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 g by noting that $B(k+1, f+1) \equiv \int_0^1 \theta^k (1-\theta)^f \, \mathrm{d}\theta$ is the beta function and $h(\theta | k+1, f+1) \equiv \theta^k (1-\theta)^f / B(k+1, f+1)$ is the density function of the beta distribution with shape parameters $k+1$ and $f+1$. Substituting the likelihood terms out of Equation (6),

Figure F.1: Example of a history and corresponding prior belief over θ_{post}



Historical data



Belief distribution over θ_{post}

we find that

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

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

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

F.2 Equilibrium

In the described game, the advisor's strategy $m^* : \{\theta, 1\} \times \Theta \to \Delta M$ maps from the advisor's type and the state of the world into a probability distribution over messages. The investor's strategy $\theta^{I*}_{post} : M \to \Delta A$ 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 $\theta^{I*}_{post}(m) \neq \mathbb{E}(\theta_{post})$. 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

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

This implies the existence of a unique threshold $\hat{\theta} = (\theta^I_{post}(m') + \theta^I_{post}(m''))/2$ 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

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

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

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

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

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

Define a function $\Phi(\theta) \equiv \theta - 1/2\Big[\theta^{I*}_{post}(m'; \theta) + \theta^{I}_{post}(m''; \theta)\Big]$. 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. □

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 $\hat{\theta}^H_{post}$ and the down-advisor always sends $\hat{\theta}^L_{post}$. 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 θ_{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

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

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

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

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

Noting that

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

we can plug in an rearrange to arrive at

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

It is straightforward to verify that $\theta^{I*L'}_{post}(0) < 0$ and $\theta^{I*L'}_{post}(1) > 0$. 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 $g(\theta)$ is a mixture distribution of different beta distributions: For a mixture distribution where the conditional expectations of the individual components are $\mathbb{E}_1(\theta|\theta \leq \hat{\theta})$, $\mathbb{E}_2(\theta|\theta \leq \hat{\theta}), \ldots$, and where the density functions are weighted by $w_1, w_2, \ldots$ 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, $\frac{\partial \mathbb{E}_i(\theta|\theta \leq \hat{\theta})}{\partial \hat{\theta}} < 1$,[52] which implies $\frac{\partial \mathbb{E}(\theta|\theta \leq \hat{\theta})}{\partial \hat{\theta}} < 1$. Therefore, $\theta^{I*'}_{post}(\hat{\theta})$ 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

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

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

[52] See, e.g., Lemma 1 in Harbaugh & 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 θ_{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 θ^A_{post} is on the inside of the interval *and* c^A and θ^A_{pre} maximize the likelihood function conditional on θ^A_{post}. 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.