Narrative Persuasion*

Kai Barron
WZB Berlin

Tilman Fries
LMU Munich

May 27, 2024

For the current version, click here.

Abstract

We study how one person may shape the way another person interprets objective information. They do this by proposing a sense-making explanation (or narrative). Using a theory-driven experiment, we investigate the mechanics of such narrative persuasion and document four main findings. First, narratives are persuasive: We find that they systematically shift beliefs. Second, *narrative fit* (coherence with the facts) is a key determinant of persuasiveness. Third, this fit-heuristic is anticipated by narrative-senders, who systematically tailor their narratives to the facts. Fourth, the features of a competing narrative predictably influence both narrative construction and adoption.

JEL Codes: D83, G40, G50, C90.
Keywords: Narratives, beliefs, explanations, mental models, experiment, financial advice.

*We are grateful 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, Philippe d'Astous, Dirk Engelmann, Nicola Gennaioli, Katrin Gödker, Thomas Graeber, Jeanne Hagenbach, Luca Henkel, Emeric Henry, Steffen Huck, Alessandro Ispano, Agne Kajackaite, Chad Kendall, Anita Kopányi-Peuker, Dorothea Kübler, Christine Laudenbach, Yves Le Yaouanq, Yiming Liu, George Loewenstein, Frieder Neunhoeffer, Salvatore Nunnari, Davide Pace, Collin Raymond, Chris Roth, Klaus Schmidt, Christoph Semken, Joshua Schwartzstein, Paul Seabright, Alice Solda, Adi Sunderam, Heidi Thysen, Joël van der Weele, Georg Weizsäcker, and Florian Zimmermann for many interesting discussions and helpful suggestions. We thank the WZB for generously funding this project by means of its "seed money" programme and gratefully acknowledge financial support from the *Deutsche Forschungsgemeinschaft* through CRC TRR 190 (project number 280092119). The experiments reported in this study were preregistered in the AEA registry with the unique identifiers: AEARCTR-0009103 and AEARCTR-0011565.

1 Introduction

Narratives are sense-making devices; they provide causal explanations for how events are interconnected. Recent work has argued that narratives play a key role in economic thinking and behavior, with Shiller (2019) asserting that narratives are a major driver of economic fluctuations, Spiegler (2020a) developing a formal toolbox that places causal misperceptions at the heart of nonrational expectations, and Andre, Haaland, Roth, and Wohlfart (2023) demonstrating that individuals display substantial heterogeneity in their causal accounts of macroeconomic events (e.g., inflation). Importantly, individuals do not make sense of the world on their own. Therefore, narratives are often communicated; individuals share them using simple stories, metaphors, or anecdotes via word-of-mouth or on social media. As they may also be used as a persuasive tool, it is crucial to understand how the communication of narratives operates. Yet, empirical work is scarce. One reason for this is that it is challenging to study the transmission of narratives in field settings. Narratives, for example, are difficult to measure and the analyst rarely observes the incentives and information sets of the narrative sender and receiver.

In this paper, we circumvent these issues by designing an experiment that allows us to study the construction and persuasiveness of narratives in a controlled strategic setting. Our experiment is framed as a financial advice task, with participants assigned to being either advisors or investors. Both players receive identical historical performance data from a hypothetical company. The investor wishes to evaluate the company's future prospects, but, crucially, the advisor may try to influence the investor's *interpretation of the historical performance data* and, therefore, his beliefs about the company. The advisor does this by proposing a narrative that makes sense of the data. A key attribute of our study is that we can study both sides of the strategic interaction with full knowledge of (and tight control over) both players' information sets. Using the control provided by our design, we can analyze how advisors with different incentives construct narratives and how these narratives causally influence investors' beliefs. Importantly, we are also able to measure a central feature of narratives, namely narrative fit—how well the narrative explains the historical data—which allows us to test key assumptions and predictions of the theoretical narrative persuasion framework provided by Schwartzstein and Sunderam (2021) (henceforth S&S).

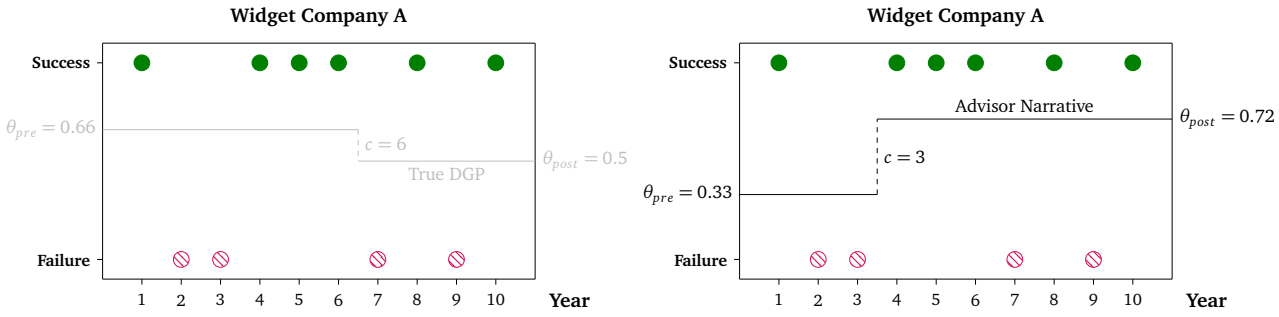
Our study makes four primary contributions to the literature. First, we show that humans are susceptible to narrative persuasion. Narratives are persuasive even if investors know the narrative is constructed by an advisor *(i)* who can tailor it to the public information about the company *ex-post*, *(ii)* who has no private information, *(iii)* with misaligned incentives. Second, we show that narrative fit (coherence with the facts) is a key determinant of their persuasiveness. Third, advisors anticipate the importance of narrative fit. When constructing their narratives, they balance the tension between making an ambitious *claim* about the company's future prospects and establishing a narrative that fits the facts well. This yields narratives with distinctive, predictable features. Fourth, when facing a competing narrative, advisors adapt their own narratives based on the characteristics of the alternative. If the competing narrative fits the data well, the advisor will increase the fit and lower the persuasive ambition of her own narrative. Finally, we introduce a versatile experimental framework that can be used as a workhorse for future inquiry. Furthermore,

since our framework constitutes a fully specified strategic setting, it permits comparisons with cheap talk.[1]

In the experiment we consider a setting with a financial advisor ("she") and an investor ("he"). Both individuals observe the same historical data from a company. This takes the form of the company's past performance and is represented visually, similar to the representation in each of the panels of Figure 1: The solid green dots denote years where the company experienced "success" and the hatched red dots denote years where the company experienced "failure". In each year, the company's probability of success depends on an underlying parameter—in the experiment, we tell participants that the parameter captures the quality of the company's CEO. Importantly, the CEO changed once during the ten years of the company's history. Therefore, the data-generating process (DGP) can be described by three parameters; the probability of success under the previous CEO (θ_{pre}), the probability of success under the current CEO (θ_{post}), and the year of CEO change (c). The grey line in Panel (i) provides an example of what such a true DGP might look like.

The investor in our experiment knows that the data follows this basic structure, but does not know the exact parameter values of the true DGP. His task is to estimate the company's probability of success under the current CEO. Therefore, he only cares directly about θ_{post}.

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



(i) Data and true DGP (ii) Data and advisor narrative

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 period covered by the historical data. This narrative may guide how the investor interprets the data and influence the beliefs he forms about the company's future performance. The advisor's narrative has the same structure as the true DGP; it consists of three parameter values that describe a possible underlying DGP. Specifically, the advisor sends the investor a message that consists of a statement about the company's success probability under the previous CEO, the company's success probability under the current CEO,

[1]The key distinguishing feature of narrative persuasion is that it operates by influencing the *interpretation of information* (Schwartzstein and Sunderam, 2021). This differs from other much-studied forms of persuasion, such as disclosure games (e.g., Milgrom, 1981), cheap-talk (e.g., Crawford and Sobel, 1982), and Bayesian persuasion (e.g., Kamenica and Gentzkow, 2011). In these scenarios, persuasion typically involves *information transmission* between a more informed and a less informed individual. Essentially, standard communication games assume that different individuals interpret information in the same way (e.g., by applying Bayes' rule) but differ in the information they possess, while narrative persuasion considers cases where everyone possesses the same information but may interpret it differently. Therefore, while in standard communication games, persuasion functions by providing the receiver with *new information*, narrative persuasion operates by providing the receiver with *a new interpretation* of commonly known information.

and the year in which the CEO changed. The message thus provides the investor with a *claim*, θ_{post}, about the company's future performance and an *explanation*, θ_{pre} and c, that makes sense of the company's past performance. Focusing on this well-defined set of narratives enables us to precisely and quantitatively capture core features of narratives (i.e., bias and fit), facilitating a direct mapping between the theoretical frameworks we consider and our empirical analysis.

Our identification of the mechanics of narrative persuasion relies primarily on exogenous variation in advisor knowledge and advisor incentives. To induce variation in advisor knowledge, we consider two scenarios. In the first knowledge scenario, SYMMETRIC, the information sets of the investor and advisor are the same—both only observe the historical company performance data. In the second knowledge scenario, ASYMMETRIC, the advisor additionally learns the parameter values of the true underlying process that generated the historical performance data. To induce variation in advisor incentives, we assign each advisor to one of three incentive-types: They can be up-advisors, who are incentivized to persuade investors that the company's current probability of success is high, down-advisors, who are incentivized to persuade investors that the company's current probability of success is low, or aligned advisors, who are incentivized to induce accurate beliefs in their matched investors. When making the assessment, the investor does not know which type of advisor sent him the narrative.

Figure 1 illustrates the core intuition of narrative persuasion in our experiment. The grey line in Panel (i) indicates an example of one possible true underlying DGP (observed by the advisor in the ASYMMETRIC scenario). The black line in Panel (ii) visualizes 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 current CEO, θ_{post}. This example highlights a central feature of narrative persuasion. While the advisor only cares about moving the investor's belief about θ_{post}, she can choose the other two components of the message (the "explanation") in a way that improves the fit of the narrative to the data. She might do this if she believes that investors evaluate the plausibility of narratives by assessing how well they fit the public 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 current CEO has had more successful years than they actually have had. Consequently, according to her narrative, there are fewer successful years during the tenure of the previous CEO. Therefore, to improve the fit of her narrative to the data, the advisor also shifts her assessment of the company's probability of success under the previous CEO downwards.

We present four sets of main results. First, focusing on investors, we find that narratives are persuasive: they systematically shift investors' beliefs. Specifically, investors who meet an up-advisor form more optimistic beliefs about the company than those who meet a down-advisor. Furthermore, investors who meet a misaligned advisor form beliefs that are further from the truth than those who meet an aligned advisor. Importantly, these results hold in both ASYMMETRIC and SYMMETRIC. This is significant because it implies that narratives are persuasive even in *pure interpretation* scenarios where the advisor has no private information and the investor knows this. In scenarios like this, persuasion operates solely via influencing the interpretation of the public information.

Second, we investigate which properties of narratives make them more convincing. We find that a key determinant of a narrative's persuasiveness is its coherence with the facts (as measured by the empirical fit).[2] To show this, we document that when advisors construct narratives with a better fit, investors form beliefs that are closer to the advisor's claim about the company's future success probability. To establish causality, in an additional treatment, we exogenously vary the fit of narratives that investors are exposed to. We find that when the fit of a narrative is exogenously increased, investors are more likely to believe the narrative. Together, these results indicate that investors try to assess the veracity of a narrative by comparing it to the available facts. They shift their beliefs more when the narrative achieves a high empirical fit.

Third, turning to advisors, we find that they do try to use the narratives they send as a tool for persuading investors. They do this by transmitting narratives that contain a claim about the company that is biased by their private incentives. Crucially, advisors also anticipate the key role of empirical fit. This means that they do not only bias their claim about the company's future success, but also systematically adjust the explanation they provide to ensure that the narrative fits the public facts well. This constitutes a fairly sophisticated strategy that balances the desire to *move* investors' beliefs against the recognition that a high *fit* is key to ensuring that investors find the narrative convincing. Our results demonstrate that the average advisor in our experiment constructs narratives that are consistent with balancing this fit-movement tradeoff. This is informative as it shows that not only do investors evaluate narratives based primarily on their empirical fit, but advisors anticipate this and design their narratives accordingly. Consequently, advisors are able to construct narratives that investors find almost as convincing as the truth, on average. In one decision in our COMPETITION treatment, investors are presented with two narratives—one that is the true underlying data-generating process and one that is constructed by a human advisor. They are asked to choose which they believe is closer to the truth. Investors choose the narrative constructed by the human advisor over the true process 49% of the time.

Fourth, we examine how the presence of a competing narrative influences advisors. According to the S&S framework, increasing the fit of a competing narrative will lead advisors to construct narratives that *(i)* fit better, and *(ii)* are less biased in the claims they make about the company's future success. The reasoning behind the first prediction is that, in order for her narrative to be more compelling than the alternative, it needs to fit better than the alternative. As the fit of the competing narrative improves, the advisor will increase the fit of her own narrative to remain competitive. The second prediction is an implication of the tension between movement and fit: The more the advisor biases the claim she makes about the company, the more she reduces the fit of her narrative. When she faces a tighter fit-constraint, she is more limited in the extent to which she can bias her claim while still sending a narrative that is more convincing than the alternative. This causes her to reduce the bias. To examine whether these predictions are borne out in observed behavior, in our COMPETITION treatment, we exogenously vary the fit of the narrative that an advisor competes with. The advisor is shown the historical company data and the details of this

[2]To measure the coherence of narratives in relation to the available facts, we construct an index that orders narratives according to their likelihood fit—i.e., Pr(data|narrative). This index provides a metric of how likely it is that the narrative under consideration generated the observed public data.

competing narrative before she chooses her own narrative to send to the investor. The investor then has to assess which of the two narratives is closer to the truth. Our results are in line with the theoretical predictions; we find that as the fit of the competing narrative increases, advisors increase the fit of their own narrative and become less biased in their claims about the company. These results demonstrate empirically that narrative competition can exert a constraining force on advisors, limiting the extent of their persuasion when the quality of the competing narrative is high. The results also underscore the empirical relevance of the fit-movement tradeoff in narrative construction.

In addition to these four main sets of results, we also document several additional findings. First, we investigate the sensitivity of narrative persuasion to context. To do this, we consider a series of three "intervention" treatments—fully disclosing advisor incentives, a nudge, and providing the investor with private information—that aim to protect the investor from being misled by a false narrative. We find that none of these three intervention treatments moves the average investor closer to the truth. This suggests that narrative persuasion is fairly robust to contextual factors. Second, we provide direct evidence on the role played by *explanations* in our context. To do this, we exogenously vary whether an investor receives only a *claim* about the company's future performance or a *claim* accompanied by an *explanation* of the public data, describing when the CEO changed and the probability of success under the previous CEO. We find that the quality of explanations matters—claims supported by good explanations are more persuasive than claims supported by bad explanations. Third, we estimate a simple structural model to quantify the role of decision noise in our setup. Unsurprisingly, we find that there is some noise present in the decision-making of both investors and advisors. For example, when presented with two narratives, investors do not *always* adopt the better-fitting narrative over one that has a slightly worse fit. However, interestingly, our results also suggest that advisors anticipate the noise in investors' decision rules and account for it in constructing their narratives.

Taken together, these results are broadly in line with the predictions and assumptions of S&S. In scenarios where persuasion is about the *interpretation of public information*, the fit of the narrative to the information plays a key role in determining whether a proposed narrative is believed or not. This is in contrast to two natural alternative benchmarks for investor behavior. First, investors could simply ignore the messages they receive from advisors and instead rely on their own introspection to form beliefs from the public information. 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 we present from SYMMETRIC demonstrates that persuasion occurs even in scenarios where both individuals have identical information—here, communication is purely about the interpretation of public information. This indicates that narratives can be used to persuade investors beyond what is predicted by strategic communication models, where persuasion is rooted in asymmetrical knowledge of facts.

The remainder of the paper proceeds as follows. Section 2 discusses the relationship to the

literature. Section 3 develops the theoretical framework. Section 4 describes the experimental design. In Section 5, we present the results. We then consider some extensions and robustness exercises in Section 6. Section 7 contains a concluding discussion.

2 Relationship to the Literature

Many 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, 2013; Charnysh, 2023), discourse analysis and narrative analysis in sociology (Foucault, 1972; Franzosi, 1998; Polletta, Chen, Gardner, and Motes, 2011), and narrative analysis in literary and cultural studies (Koschorke, 2018; Herman and Vervaeck, 2019). While narratives have long escaped formal treatment in economics, they have not necessarily gone unnoticed. In a seminal study, Hirschman (2013) [1977] argues that a narrative held by Western monarchs in the 18th century—that commerce can serve as a vent for unruly "passions" thereby maintaining social stability—led them to actively support the rise of capitalism. He goes on to argue that, ironically, this very change in the social structure ultimately precipitated the downfall of these monarchs. Taking a similar macro-perspective, Shiller (2019) examines several case studies in which he traces the co-movement of economic outcomes and the prevalence of particular narratives.

More recent work in economic theory has begun to study narratives more formally.[3] One strand of this literature focuses on investigating the formation and consequences of (possibly incorrect) subjective models of the world in settings without persuasion (e.g. Spiegler, 2016; Heidhues, Kőszegi, and Strack, 2018; Spiegler, 2020a,b; Mailath and Samuelson, 2020; Montiel Olea, Ortoleva, Pai, and Prat, 2022; Schumacher and Thysen, 2022; Ba, 2024). These papers typically study how particular features of the environment may allow certain subjective model misspecifications to persist (e.g. Heidhues et al., 2018; Ba, 2024), or may promote the emergence of more or less complex models (Montiel Olea et al., 2022). Another strand of the literature that relates more closely to our paper analyzes the factors that may influence narrative adoption and the implications for persuasion using narratives. For example, Eliaz and Spiegler (2020) formalize narratives as causal models that can be represented using directed acyclical graphs (DAGs) that capture the connections between different variables. The authors assume that agents can be persuaded to adopt "hopeful" narratives, i.e., those that induce optimistic beliefs, and investigate the consequences for public-opinion battles involving competing narratives. In contrast, Schwartzstein and Sunderam (2021) model narratives as likelihood functions that map data to beliefs, with agents adopting models on the basis of their likelihood fit. The most persuasive narratives are those with the highest likelihood fit. In earlier work, Froeb, Ganglmair, and Tschantz (2016) study a model

[3]Currently, there is not a does not exist a consensus on a single definition of the term *narrative* in the economics literature. S&S, for example, tend to use *narrative* and *model* interchangeably, and Eliaz and Spiegler (2020) also refer to *narratives* and *causal models* interchangeably. In an earlier working paper version of this study, we discussed the relationship between different conceptualizations of the concept in the literature in Appendix Section A (Barron and Fries, 2023). Here, we use the term to refer to *a causal explanation that makes sense of a collection of events*.

with a similar fit-based adoption rule in an application to court decision-making. A number of follow-up papers build on these ideas and study the implications of a likelihood-based decision rule in several scenarios, including those where persuaders: anticipate the arrival of future data (Aina, 2024), gain control over the data-generating process (Ichihashi and Meng, 2021), or are constrained by a benevolent planner who provides data strategically (Jain, 2023). Lang (2023) provides a different perspective on the use of narratives as a persuasive tool by considering the implications of narratives for mechanism design and Ispano (2023) shows how a receiver who only adopts a narrative if it satisfies a coherence criterion might benefit from weakening that criterion. While these recent approaches to analyzing narrative persuasion employ a variety of ways to formalize a narrative, they all study settings where a persuader endows their *claim* (the belief they want to induce) with a broader sense-making *explanation* (a justification for the claim).[4] This is a key feature that differentiates these frameworks from the classical treatment of communication in economics (discussed in more detail below).

Our experiment provides a sandbox for testing several key ideas from this theoretical literature. We do this by empirically investigating 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 facts. To generate predictions for our experiment, we draw on the framework developed by Schwartzstein and Sunderam (2021). Their key assumption is that the receiver will adopt a narrative if it explains the data sufficiently well. This assumption 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 central goal of our experiment is to test whether this key assumption provides an accurate description of the adoption decisions of narrative-receivers and whether narrative-senders account for this when constructing their narratives.

Aside from testing ideas developed in the theoretical narrative persuasion literature, our persuasion setup also relates naturally 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 and Sobel, 1982). While this work has given rise to a large body of experimental research studying cheap talk models (see, e.g., Blume, DeJong, Kim, and Sprinkle, 1998; Blume, DeJong, Neumann, and Savin, 2002; Wang, Spezio, and Camerer, 2010) and disclosure games (see, e.g., King and Wallin, 1991; Hagenbach and Perez-Richet, 2018; Jin, Luca, and Martin, 2021), our paper differs from this previous literature due to the focus on the *interpretation of facts*. Specifically, advisors in our experiment send messages that not only make claims about the payoff-relevant parameter but also contain assertions about the non payoff-relevant parameters. They may also be uninformed about the true value of these parameters. Importantly, advisors

[4]The term "narrative" is also used by Bénabou, Falk, and Tirole (2020) to describe the justification that an image-concerned agent may provide to explain their behavior (e.g., arguing that the positive impact that a charity has is low to excuse not donating), with Foerster and van der Weele (2021) and Hillenbrand and Verrina (2022) providing related theories and empirical evidence. In these settings, the agent uses the narrative to influence how their decision is interpreted, with the aim of convincing an observer that they are a "good" type.

may design their messages such that they choose an explanation (payoff-irrelevant parameters) to justify their claim (about the payoff-relevant parameter). 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 typically 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.

Finally, our work relates to a recent empirical literature in economics that explores how narratives, stories, and explanations shape behavior.[5] For example, Andre et al. (2022) study households' subjective beliefs about the responsiveness of key economic variables to macroeconomic shocks and Andre et al. (2023) 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. In the domain of pro-social behavior, Barron et al. (2023) show that when parents believe certain narratives about refugees, this can affect the pro-social behavior of their children, while Hillenbrand and Verrina (2022) also show that stories can be used to influence prosocial behavior. Graeber, Zimmermann, and 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, while Morag and Loewenstein (2023) find evidence that the act of telling a story about an owned object increases one's valuation of the object. To explore the role of qualitative explanations, Graeber, Roth, and Schesch (2024) investigate how listening to a verbal explanation influences an individual's willingness to imitate the choice of another individual (for related work from social psychology on explanations, see, e.g., Lombrozo, 2006, 2012). In the two experiments most closely related to ours, Charles and Kendall (2024) demonstrate how narratives as represented by DAGs influence beliefs and study their (nonstrategic) transmission, while Ambuehl and Thysen (2024) provide a thorough investigation into how individuals prioritize different features of narratives when choosing between competing causal interpretations that vary on a number of dimensions, such as narrative-fit, complexity, or the optimism of the embedded claim.[6]

Our study differs from this body of empirical work in several important ways. First, different

[5]The theoretical work on narratives discussed above has been followed swiftly by a rapid growth in the empirical interest in the topic. Some recent and contemporary contributions to this fast-developing empirical body of work on narratives include the following: Laudenbach, Weber, and Wohlfart (2021); Andre et al. (2023); Andre, Pizzinelli, Roth, and Wohlfart (2022); Barron, Harmgart, Huck, Schneider, and Sutter (2023); Gehring, Harm Adema, and Poutvaara (2022); Harrs, Berger, and Rockenbach (2023); Hillenbrand and Verrina (2022); Morag and Loewenstein (2023); Ambuehl and Thysen (2024); Ash, Gauthier, and Widmer (2024); Charles and Kendall (2024); Hagmann, Minson, and Tinsley (2024); Frechette, Vespa, and Yuksel (2024). This empirical work has approached the topic from several different methodological angles.

[6]Further related work includes contributions by Liu and Zhang (2023), who show that an initial exposure to a narrative can persistently change beliefs, Hüning, Mechtenberg, and Wang (2022), who study persuasion with free-text messages and demonstrate that the degree to which a message persuades is positively correlated with the number of arguments it contains, and Alysandratos, Boukouras, Georganas, and Maniadis (2020) who provide evidence suggesting that participants mistakenly perceive populist advice as expert advice. Finally, shifting from social learning to individual learning, Frechette et al. (2024) investigate how individuals construct models from data by themselves.

from this literature, we focus on the use of narratives in a strategic setting, where one individual wishes to use a narrative to persuade another by influencing their interpretation of facts. 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 et al., 2023, Ambuehl and Thysen, 2024, and Charles and Kendall, 2024, adopt a similar approach, but focus more on using the machinery of directed acyclic graphs (DAGs)). Third, while much of this work does not try to fully account for the information sets 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 captures the idea that narratives can be used to shape the way that individuals *interpret* objective data. Our framework draws heavily on the one proposed by S&S. We use it as a lens to zoom in on specific predicted behavioral patterns in the investor-advisor setup that we study empirically with our experiment. 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). We also discuss the predictions of a model in which investors are strategically sophisticated and provide a comparison of the two theoretical approaches. Importantly, we do not view our experiment and empirical analysis as providing a general horse race between the two theoretical approaches; rather, we wish to test whether there are scenarios where the S&S framework provides a useful lens for predicting behavior and a traditional game-theoretic approach does not. By providing clean evidence regarding the existence of a class of such scenarios, we believe that our paper makes a strong case for economists to add the narrative perspective on communication to their toolboxes.[7]

3.1 Basic Setup

We consider a setup with an investor ("he") and an advisor ("she"). Our experimental design will closely follows this setup. In this setting, the investor's goal is to form an accurate belief about a

[7]Economists have tended to use an Ockham's razor-type argument to advocate for the use of a single parsimonious framework that can explain diverse evidence across a range of scenarios (see, e.g., Abeler, Nosenzo, and Raymond, 2019, for a recent example). Therefore, we aim to provide evidence that behavior in certain persuasion scenarios cannot be parsimoniously rationalized by a game-theoretic approach, implying that another perspective that provides predictive accuracy would be valuable. When we write that we identify cases where the S&S framework makes *useful* predictions, we mainly mean cases where the S&S framework can explain behavior which alternative commonly used frameworks cannot.

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 a vector of successes and failures from years 1 to 10; $h \equiv (s_1, s_2, \ldots, s_{10})$.

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 dataset into a *pre* and a *post* period. In the experiment, this structural change is framed as a change of the company's CEO.[8] The structural change takes place at some point after Year 2 and before Year 9. Second, the parameter θ^T_{pre} denotes the company's success probability in the years 1 to c^T. Finally, the parameter θ^T_{post} denotes the company's success probability in the years $c^T + 1$ to 10. The true underlying model is thus given by $\boldsymbol{m^T} = (c^T, \theta^T_{pre}, \theta^T_{post}) \in \mathcal{M} \equiv \{2, \ldots, 8\} \times [0,1]^2$. The parameter values are drawn from independent uniform distributions. The investor and advisor both know that the true underlying model is part of this set. They also know the underlying distribution of the model parameters.

Actions. The advisor's advice comes in the form of a message $\boldsymbol{m^A} \in \mathcal{M}$. After receiving the message, the investor makes an assessment $\theta^I_{post} \in [0,1]$.

Payoffs. The investor's objective is to make an assessment that is as close as possible to the truth. The utility function is

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

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 persuasion target. We consider three different incentive types; up, down and aligned, which we also denote using $\uparrow, \downarrow$, and $\rightarrow$ respectively. The advisor's utility depends on the investor's assessment, θ^I_{post}, and her incentive type φ;

$$U^{\varphi}(\theta^I_{post}) = \begin{cases} 1 - (1 - \theta^I_{post})^2 & \text{if } \varphi = \uparrow, \\ 1 - (0 - \theta^I_{post})^2 & \text{if } \varphi = \downarrow, \\ 1 - (\theta^T_{post} - \theta^I_{post})^2 & \text{if } \varphi = \rightarrow . \end{cases} \tag{2}$$

This utility is maximized if θ^I_{post} equals the persuasion target: the up-advisor wants the investor to form an assessment as close as possible to the highest value, $\theta^I_{post} = 1$, the down-advisor wants

[8] This choice of frame was motivated by the goal of ensuring that the experiment is as simple and easy to understand as possible. We viewed the change of a CEO as a simple example of a structural change that might affect the company's performance in a meaningful way. Importantly, in the context of our experiment, this is not public information—in particular, investors will never be able to verify with certainty when exactly the CEO changed. We discuss these design features more in the experimental design section below.

the investor to form an assessment as close as possible to the lowest value, $\theta^I_{post} = 0$, and the aligned advisor wants the investor to form an accurate assessment.

The investor's belief about θ_{post} is *payoff relevant* for the investor and the advisor (though, while the investor wishes their θ_{post} belief to be accurate, the advisor might not necessarily wish this). Nonetheless, the advisor sends a message not only about θ_{post} but also about c and θ_{pre}. We call these parameters *auxiliary*. They play a key role in the narrative approach to communication, as explained below.

Information. The investor is uninformed about $\boldsymbol{m}^T$ while the advisor may hold superior information when choosing the advice they provide to the investor. For example, the advisor might be *informed* and know $\boldsymbol{m}^T$ when sending the message. This will be the case in our ASYMMETRIC treatment. Crucially, in the SYMMETRIC treatment of our experiment, we will also consider another key point on the information asymmetry spectrum where the advisor is *uninformed*, i.e., where she does not know $\boldsymbol{m}^T$. In this scenario, the advisor and the investor will have identical information sets, and communication will only be about the interpretation of the historical data.

We consider a benchmark condition where the investor also does not know his advisor's incentive type when making the assessment.[9]

3.2 Communication outcomes

The investor's problem in this setup is that he is uncertain about the true model, $\boldsymbol{m}^T$, governing the success and failure of the company. He, therefore, forms a belief about it. Based on this belief, he reports his assessment, θ^I_{post}, 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, he constructs a subjective model or default narrative—his own initial interpretation of the data. Second, he also receives advice. This advice arrives in the form of message $\boldsymbol{m}^A \in \mathcal{M}$, sent by the advisor. The investor then forms a final assessment using an *assessment rule*, which maps the information available to him into an assessment, θ^I_{post}. There exist different theoretical frameworks which can be used to analyze this setting that yield different assessment rules. In the main text below, we lay out S&S's assessment rule and describe how advisors construct narratives optimally according to this rule. We compare it with the assessment rule implied by a more conventional game-theoretical analysis. As noted above, we view each of these frameworks as being more suitable for analyzing a distinct set of communication scenarios characterized by specific features. We focus on those where S&S may be useful.

S&S's narrative approach. One key ingredient of S&S's framework is the assumption that, when faced with two different narratives, an agent will adopt the narrative that wins in a "Bayesian hypothesis test". This narrative will, in turn, determine the agent's beliefs and actions. In a persuasion scenario, the investor faces a decision: to either retain their pre-existing default narrative

[9] In the experiment, we will also consider a treatment condition where the investor does know his advisor's incentives.

$\boldsymbol{m}^{I,0}$—which is assumed to be degenerate and exogenous to the model—or to adopt the narrative $\boldsymbol{m}^{\boldsymbol{A}}$ proposed by the advisor.[10] In a Bayesian hypothesis test, the investor will adopt the narrative suggested by the advisor if and only if it is at least as likely that the advisor's narrative generated the observed history as it is that the investor's default narrative generated it. The assessment rule is described by the following:

$$\theta^{I}_{post}(\boldsymbol{m}^{I,0}, \boldsymbol{m}^{A}) = \begin{cases} \theta^{A}_{post} & \text{if } \Pr(h|\boldsymbol{m}^{\boldsymbol{A}}) \geq \Pr(h|\boldsymbol{m}^{I,0}) \\ \theta^{I,0}_{post} & \text{otherwise.} \end{cases} \tag{3}$$

Two features of this assessment rule stand out. First, the rule follows a binary structure; the investor either adopts the advisor's narrative or sticks with his default narrative. In an earlier working paper version of their paper, S&S introduce the notion of "value-adjusted fit", where the receiver averages between the narratives he is exposed to (including the sender's narrative). We choose to employ the simpler binary structure primarily due to its simplicity, which sharpens the focus on the specific components of persuasion that are of interest to us. Adopting the value-adjusted fit approach would not, however, alter the qualitative conclusions of our analysis. The second, more pivotal, feature is the central role that the assessment rule ascribes to narrative fit; the investor adopts the narrative with the better fit as measured by its likelihood given the data.We denote the log likelihood function by $\ell(\boldsymbol{m})$ and the narrative in $\mathcal{M}$ that maximizes the likelihood function by $\boldsymbol{m}^{\boldsymbol{DO}}$. 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 (3) will always adopt $\boldsymbol{m}^{\boldsymbol{DO}}$ if he receives it as a message. For most histories, the data-optimal narrative is unique.[11] 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 log likelihood values obtained from $(c, \theta^{DO}_{pre}(c), \theta^{DO}_{post}(c))$ across all possible values 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 both $\theta^{DO}_{pre}(c)$ and $\theta^{DO}_{post}(c)$ parameter values that are closer to the extremes (i.e., 0 and 1) dominate the other narratives in terms of empirical fit.[12] This captures the following intuition: A narrative that partitions the historical data into a *pre* and *post* period such that the company is either very successful or very unsuccessful in each partition will usually have a high log likelihood value. Narratives that instead partition the data to have a more equal

[10]Instead of assuming a degenerate default narrative, one can also think about the investor sampling from a set of narratives $\mathcal{M}^{I,0}$ after seeing the data. The investor might then compare the narratives in $\mathcal{M}^{I,0}$ and the advisor's narrative $\boldsymbol{m}^{\boldsymbol{A}}$ using a Bayesian hypothesis test and pick the narrative that wins. This problem is equivalent to a problem where the investor only compares the narrative with the highest fit from $\mathcal{M}^{I,0}$ with $\boldsymbol{m}^{\boldsymbol{A}}$. Therefore, assuming that the advisor holds a single default narrative is not as limiting as it might appear at first glance.

[11]It is not unique for 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. However, this is an exception and histories with a unique data-optimal model are far more common.

[12]The functions $\theta^{DO}_{pre}(c)$ and $\theta^{DO}_{post}(c)$ denote the proportion of successes and failures in *pre* and *post* when the structural break is in year c. When comparing two narratives $m' = (c', \theta^{DO}_{pre}(c'), \theta^{DO}_{post}(c''))$ and $m'' = (c'', \theta^{DO}_{pre}(c''), \theta^{DO}_{post}(c''))$, m' will have a higher fit than m'' if $\min\{\theta^{DO}_{pre}(c'), 1-\theta^{DO}_{pre}(c')\} < \min\{\theta^{DO}_{pre}(c''), 1-\theta^{DO}_{pre}(c'')\}$ and $\min\{\theta^{DO}_{post}(c'), 1-\theta^{DO}_{post}(c')\} < \min\{\theta^{DO}_{post}(c''), 1-\theta^{DO}_{post}(c'')\}$.

proportion of successful and unsuccessful years within the *pre* and *post* period will have lower log likelihood values. In this sense, narratives that more coherently explain success and failure in the data are more likely to be adopted.

To describe how the advisor optimally chooses her advice given the assessment rule in Equation (3), we close the model by specifying the advisor's beliefs about the investor's default narrative. We denote the pdf of the advisor's beliefs by $f(\boldsymbol{m}^{I,0})$. Unless stated otherwise, we assume that f has full support on $\mathcal{M}$.[13] Anticipating the possible default narratives that the investor may hold and taking into account his assessment rule, the advisor then chooses $\boldsymbol{m}^{\boldsymbol{A}}$ to maximize her expected utility:

$$\boldsymbol{m}^{\boldsymbol{A}} \in \underset{\boldsymbol{m}\in\mathcal{M}}{\arg\max} \quad \mathbb{E}_f[U^{\varphi}(\theta^{I}_{post}(\boldsymbol{m}^{I,0}, \boldsymbol{m}))|\boldsymbol{m}].$$

In the equation above, the expectation operator, $\mathbb{E}_f$, reflects the advisor's expectation, given her beliefs over the set of possible default models held by investors.

We highlight a number of properties of the advisor's optimal narrative. Proposition 1 shows that the advisor distorts the payoff-relevant parameter of the narrative, θ^{A}_{post}, towards her persuasion target and uses the supporting narrative components, c^A and θ^{A}_{pre}, to improve the fit of the narrative. Essentially, the c^A and θ^{A}_{pre} are used to *corroborate* the θ^{A}_{post} she sends.[14]

Proposition 1 (S&S framework)**.** *For an advisor with persuasion target ϕ, it holds that:*

(i) (Attempting persuasion) If the persuasion target is different from the data-optimal value, the advisor moves the θ^{A}_{post} in their narrative away from the data optimum, θ^{DO}_{post}, and towards the persuasion target:

$$\theta^{A}_{post} \geq \theta^{DO}_{post} \text{ if } \theta^{DO}_{post} \leq \phi \text{ and } \theta^{A}_{post} < \theta^{DO}_{post} \text{ if } \theta^{DO}_{post} > \phi.$$

(ii) (Improving coherence) Among all messages in $\mathcal{M}$, the advisor will only consider sending those with a c and θ_{pre} that maximize the log likelihood function conditional on θ_{post}. This implies that:

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

Two motives guide the advisor's optimal narrative choice. First, advisors wish to construct narratives that are coherent with the data. As the likelihood fit of $\boldsymbol{m}^{\boldsymbol{A}}$ increases, so does the probability that $\boldsymbol{m}^{\boldsymbol{A}}$ wins in the Bayesian hypothesis test. Therefore, the investor becomes more likely to adopt the advisor's narrative as the fit increases. Second, advisors wish to shift investor's beliefs towards their persuasion target. Conditional on the investor adopting the advisor's narrative, he will move his assessment from $\theta^{I,0}_{post}$ to θ^{A}_{post}. Therefore, the advisor biases θ^{A}_{post} in direction of her persuasion target. When constructing the narrative, the advisor therefore considers narrative *fit* and belief *movement*. There is typically a tension between the two motives. While the advisor

[13] This is a departure from S&S, who assume that the sender knows the receiver's default narrative with certainty. Here, we relax this assumption.

[14] The proof of Proposition 1 and a more detailed discussion on S&S's narrative approach is included in Appendix C.

can ensure narrative adoption by sending the data-optimal narrative, this will usually not be the optimal choice since θ^{DO}_{post} will typically not coincide with the advisor's persuasion target. Part *(i)* of the proposition above states that, on the margin, the advisor will find it optimal to move θ^{A}_{post} away from the data-optimum towards the persuasion target, trading off narrative fit and belief movement. Part *(ii)* then notes that, since the targeted movement operates entirely on θ^{A}_{post}, the advisor will select the auxiliary parameters θ^{A}_{pre} and c^{A} such that they maximize the likelihood fit *conditional* on θ^{A}_{post}.

Model selection criteria based on fit—such as the Bayesian hypothesis test assumed by S&S—are frequently used in practice. For example, empirical researchers may choose between theories based on how well each fits the data and, in everyday life, we may also focus on coherence when forming our mental model of how the world works. One crucial implication of this is that in contexts where one individual is trying to persuade another, this model selection rule allows for the systematic distortion of beliefs. Specifically, the beliefs of an investor with the S&S adoption rule do not follow the law of iterated expectations. This means that the advisor can *systematically* bias the investor's expectation away from his prior. This sort of decision rule is, therefore, not consistent with Bayesian rationality. Intuitively, in a strategic equilibrium-type framework, the investor should anticipate that the advisor's choice of narrative will be guided by her persuasion target and accordingly be skeptical. To delineate the differences between the predicted outcomes when investors follow a S&S fit-based adoption rule and those expected when investors exhibit strategic sophistication, we also examine the setup through the lens of conventional game theory.

Cheap talk benchmark. In this section, we provide a summary of the key differences from the S&S model when investors are are strategically sophisticated. In Appendix D, we provide a more detailed discussion of how we apply a standard game theoretic approach to our setting. The cheap talk benchmark that we consider has the following features. First, both advisors and investors are assumed to hold a common belief over models before communication takes place. They both arrive at this belief by incorporating the information contained in the historical data through Bayesian updating. Second, the investor takes into account the information that he has about the advisor's three possible incentive types. Third, we allow for a fraction of advisors, $\lambda \in [0, 1]$, to be *honest* and always truthfully report their beliefs about the model, $\boldsymbol{m}$. Here, we focus on the case where the advisor is *informed* and knows the truth, $\boldsymbol{m^T}$. Therefore, an honest advisor will always send $\boldsymbol{m}^T$. The non-honest (strategic) types choose a utility-maximizing strategy.

In the main text, we focus on Perfect Bayesian Equilibria (PBE) where the (strategic) aligned advisor finds it optimal to follow an honest strategy. We comment on the implications of this refinement at the end of this section. Equilibria in which the aligned advisor follows an honest strategy are characterized by two unique thresholds, θ_L and θ_H, and the investor can be induced to make any assessment in $[\theta_L, \theta_H]$. In such an equilibrium, the up-advisor follows a strategy of mixing between all possible narratives with $\theta^{A}_{post} \geq \theta_H$ and the down-advisor mixes between all possible narratives with $\theta^{A}_{post} \leq \theta_L$. These misaligned advisors will mix in such a way that, upon receiving a narrative with $\theta^{A}_{post} \geq \theta_H$ ($\theta^{A}_{post} \leq \theta_L$) the investor finds it optimal to make assessment θ_H (θ_L). In contrast, only advisors who are following an honest strategy send messages inside the

interval—i.e., with $\theta^A_{post} \in (\theta_L, \theta_H)$—which induces the investor to always make an assessment equal to θ^A_{post} if θ^A_{post} is between the thresholds. We can show that the following result holds in any such equilibrium:

Proposition 2 (Cheap talk framework)**.** *Consider any equilibrium in which the aligned advisor follows an honest strategy and compare any two narratives* $\boldsymbol{m}' = (c', \theta'_{pre}, \theta_{post})$ *and* $\boldsymbol{m}'' = (c'', \theta''_{pre}, \theta_{post})$ *that are sent with positive probability and where* $c' \neq c''$ *and/or* $\theta'_{pre} \neq \theta''_{pre}$*. Then, the investor's assessment of* θ^T_{post} *is the same after receiving either narrative.*

This proposition essentially says that an investor's reaction to receiving a narrative depends entirely on the θ^A_{post} and is invariant to the auxiliary parameters, c^A and θ^A_{pre}, of the narrative. To see that, consider that the investor can receive two broad types of narratives. First, there are narratives which are only sent by an honest/aligned advisor in equilibrium (i.e., those within the interval). Because these are honest and well-informed narratives, it is optimal for the investor to follow the θ^A_{post} of the narrative in forming his assessment, irrespective of the auxiliary parameters. Second, there are narratives which are sent by either the honest/aligned advisor or one type of misaligned advisor (i.e., those outside the interval). Equilibrium requires that the investor's assessment after receiving such a narrative is either θ_L or θ_H. This decision again depends only on θ^A_{post} and not on the auxiliary parameters. Therefore, the auxiliary parameters are *irrelevant* for persuasion in the cheap talk benchmark.

How general is this result? Since the aligned advisor does not have any incentive to deceive the investor, focusing on equilibria where she is honest seems natural. In addition, Appendix D shows that such an equilibrium is also *most informative* and that the result above extends to *any* most informative equilibrium if $\lambda > 0$.[15]

One may, however, conjecture that the equilibrium predictions would ascribe more communication relevance to the auxiliary parameters if one were to assume that the investor is not fully strategic. This would imply partially relaxing the strategic reasoning assumption of cheap talk, but not going as far as S&S in assuming fully non-strategic reasoning. Perhaps surprisingly, this conjecture does not, in fact, hold when considering several natural ways to model investors as being partially strategic. For example, the cheap talk literature has considered credulity, where a message-receiver interprets messages in a literal sense, as a natural deviation from fully strategic reasoning (Kartik, Ottaviani, and Squintani, 2007; Chen, 2011). In the spirit of S&S, we can imagine an investor who displays fit-based credulity, i.e., who puts a positive weight on the literal interpretation of the message, with this weight increasing in the fit of the message. Under this assumption, there exists an equilibrium that essentially shares the key qualitative features of the

[15]An equilibrium is referred to as "most informative" if it is the equilibrium where the investor learns the most, i.e., that minimizes the investor's squared assessment error. In our setting, all players prefer the most informative equilibrium outcome to any other equilibrium outcome ex-ante, and the most informative equilibrium is essentially unique. In Appendix D, we also discuss the case with no honest advisors ($\lambda = 0$). In this case, an equilibrium like the one sketched out in the main text remains a most informative equilibrium. However, because messages only take meaning in equilibrium, one can obtain multiple most informative equilibria that are essentially the same but simply involve a relabling of messages. Allowing for this would imply that the discussion of a result like the one in the main text would need to be more qualified to account for such relabeling possibilities. We leave this discussion to the Appendix.

one described above. In particular, it has the features highlighted in Proposition 2. When investors display fit-based credulity, the relevance of empirical fit does certainly increase: because the investor interprets messages with a higher fit more literally than those with a lower fit, misaligned advisors have an incentive to "tailor" the auxiliary parameters to the data more often. However, as long as the investor is not fully credulous, in equilibrium, he will discount these messages more because he expects misaligned advisors to send them more often. In equilibrium, therefore, his reaction to the auxiliary parameters is fully muted.[16]

3.3 Predictions for the Experiment

In this section, we describe several predictions that we will test using our experimental data. Our goal is to highlight how the predictions generated by the S&S narrative approach differ from those derived from a more conventional strategic analysis. This will help us to evaluate whether the S&S approach offers an additional useful lens for understanding persuasion in certain contexts. In particular, we wish to focus on scenarios where the interpretation of information may be an important margin of persuasion. This differs from the traditional focus on information transmission as the key margin of persuasion. Since many real-world situations that we care about provide scope for persuasion on both margins, we consider two types of scenarios. First, we examine a 'pure interpretation scenario', where there is no scope for persuasion via information transmission, and persuasion must operate through influencing the interpretation of information. Second, we consider a 'hybrid scenario,' which may involve both information transmission and influencing the interpretation of information. The former allows us to cleanly assess the predictive power of the S&S narrative approach, while the latter is more representative of a broader class of real-world scenarios of interest.

Investor behavior

One way to reformulate S&S's assessment rule is as follows: the investor will adopt the advisor's narrative if it fits the data sufficiently well. Below, we outline a set of predictions derived from this fit-based assessment rule.

Prediction 1 says that if investors use a fit-based criterion to evaluate narratives, then persuasion is possible in both pure interpretation and hybrid scenarios. This is important because these predictions differ from those of cheap talk. According to the cheap talk benchmark, which conceptualizes communication as a game where the advisor signals her private information, it clearly matters whether the advisor knows $\boldsymbol{m}^T$ or not. In contrast, under the S&S assessment rule, the fit of the narrative is what matters (irrespective of whether the advisor is more informed). Therefore, persuasion is possible even in the pure interpretation scenario. Under the cheap talk framework, persuasion is not possible in this scenario.

[16] See Appendix D.6 for a formal discussion of this case and of cases where the investor may be cursed or simply credulous. To provide S&S's assessment rule with a strategic rationale that survives equilibrium analysis, one may instead turn to biases on the advisor side. In Appendix D.7, we discuss how a model where advisors are differently skilled in tailoring their lies to data could potentially provide such a rationale.

Prediction 1 (Persuasion in pure interpretation and hybrid scenarios)**.** *The narrative sent by the advisor influences the investor's assessment:*

(i) in the pure interpretation scenario, where it is common knowledge that the advisor has no additional information about $\boldsymbol{m^T}$ *relative to the investor.*

(ii) in the hybrid scenario, where advisors know $\boldsymbol{m^T}$*, while investors do not.*

Prediction 2 focuses on the direct influence of narrative fit. There are two sub-predictions, with the second providing a more demanding test of the fit-based assessment rule relative to the first. Part *(i)* provides a simple statement that an S&S fit-based rule predicts a relationship between the advisor's narrative fit and the investor's assessment: as fit increases, the investor forms an assessment that is closer to the advisor's narrative. Therefore, the distance between θ^I_{post} and θ^A_{post} should decrease in the fit. However, it is important to note that this pattern can also be rationalized by a cheap talk equilibrium. Essentially, in a most informative equilibrium, the investor follows θ^A_{post} if and only if it is sufficiently close to his prior assessment of θ_{post}. Since closeness to the prior is likely correlated with fit, the fit-assessment relation could also occur in this equilibrium. To address this concern, Part *(ii)* predicts that a better fit should matter even when fixing θ^A_{post}. That is, improvements in fit caused only by changes in the auxiliary parameters should make a narrative more persuasive. In the experiment, we introduce a number of treatments tailored to test predictions *(i)* and *(ii)*.

Prediction 2 (Influence of narrative fit)**.** *Take the distance* $|\theta^I_{post} - \theta^A_{post}|$ *as a measure of closeness between the advisor's narrative and the investor's assessment. The following holds:*

(i) The distance $|\theta^I_{post} - \theta^A_{post}|$ *decreases as the likelihood fit of* $\boldsymbol{m^A}$ *increases.*

(ii) Fixing θ^A_{post}*, the distance* $|\theta^I_{post} - \theta^A_{post}|$ *decreases as the likelihood fit of* $\boldsymbol{m^A}$ *increases.*

Advisor behavior

Moving to advisors, we derive predictions for how they construct narratives if they expect investors to follow an S&S fit-based assessment rule. Consider a data generating process where the components of $\boldsymbol{m^T}$ are drawn from independent uniform distributions, the history h is generated by $\boldsymbol{m^T}$ and the advisor chooses $\boldsymbol{m^A}$ knowing h. When choosing $\boldsymbol{m^A}$, an up-advisor will move θ^A_{post} away from the data-optimum towards the persuasion target of 1. In order to "justify" sending θ^A_{post}, she will increase the fit by choosing data-optimal auxiliary parameters, conditional on θ^A_{post}. For example, the up-advisor can justify sending a high θ^A_{post} by adjusting her claim about the position of the structural change, c^A, to artificially increase the apparent fraction of successes in *post*. As a consequence of doing this, the advisor will mechanically decrease the apparent fraction of successes in *pre*. To maintain the fit of her narrative to the data, this will induce her to decrease her stated θ^A_{pre}. Therefore, the up-advisor who increases her stated θ^A_{post} above the data-optimal value θ^{DO}_{post} correspondingly decreases her stated θ^A_{pre} below θ^{DO}_{pre}. Conversely, a down-advisor will increase her stated θ^A_{pre} *because* she has incentives to decrease her stated θ^A_{post} below θ^{DO}_{post}. This yields the following testable prediction: When comparing the up- and down-advisor, the expected

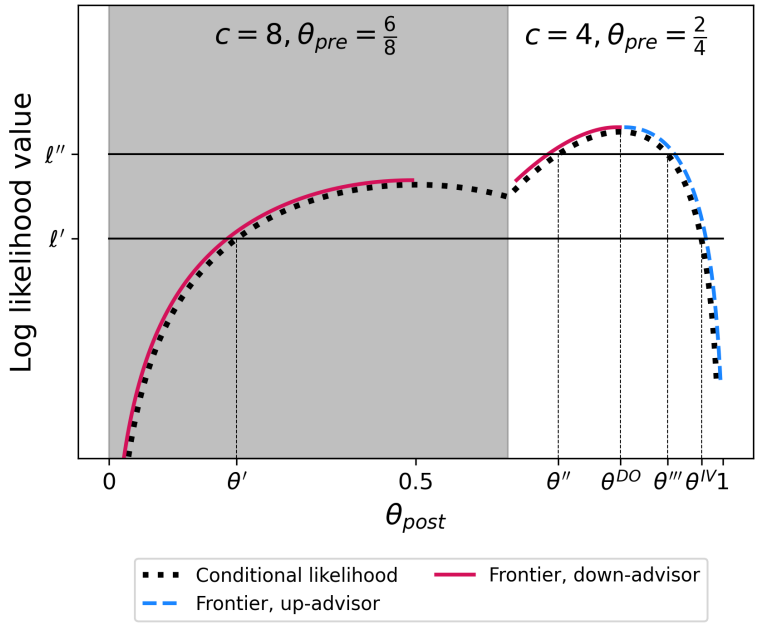
value of θ^A_{post} should be higher for the up- than for the down-advisor and the expected value of θ^A_{pre} should be lower for the up- than for the down-advisor.

Prediction 3 (Fit-movement tradeoff in narrative construction)**.** *Misaligned advisors shift the θ^A_{post} of their narrative towards their persuasion target and shift θ^A_{pre} in the opposite direction, yielding the following statistical regularities:*

$$\mathbb{E}[\theta^A_{post}|\varphi = \uparrow] > \mathbb{E}[\theta^A_{post}|\varphi = \downarrow] \text{ and } \mathbb{E}[\theta^A_{pre}|\varphi = \uparrow] < \mathbb{E}[\theta^A_{pre}|\varphi = \downarrow].$$

Finally, the S&S assessment rule generates interesting implications for how the advisor should adjust her narrative in response to changes in her beliefs about the fit of the investor's (competing) default narrative. Consider a situation where the advisor is assumed to know the investor's default model for sure. Now, because she needs to send a narrative that fits better to be persuasive, she should clearly take the fit of the investor's model into account when constructing the narrative that she will send.

Figure 2: Likelihood frontier and comparative statics of increasing the default narrative fit



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. The horizontal lines denote possible likelihood values of the default narrative fit. The advisor's optimal narrative for a given default narrative is given by the intersection between the horizontal line and her likelihood frontier.

To illustrate this, consider the set of narratives represented by what we call the "likelihood frontier". This is the set of all messages that are not dominated on both their fit and movement by any other message. 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 dotted black 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^A_{post}. It takes its maximum value when the message equals the data-optimal model,

$(c^{DO} = 4, \theta^{DO}_{pre} = 2/4, \theta^{DO}_{post} = 5/6)$.[17] 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 sending each of these messages can be rationalized under some intensity level 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. The figure also includes two horizontal lines displaying examples of the investor's default narrative fit. The advisor's optimal narrative is at the point where the likelihood frontier and the horizontal line intersect. For example, if the default narrative fit is ℓ', then the down-advisor optimally sends narrative $(8, 6/8, \theta')$, the narrative with the lowest θ_{post}-value that still has a fit of at least ℓ' and will thus be adopted by the investor. As the default narrative fit increases to ℓ'', the down-advisor increases θ^{A}_{post} from θ' to θ'' in order to increase fit. Likewise, as the default narrative fit increases from ℓ' to ℓ'', the up-advisor shifts from a narrative with θ^{IV} to a narrative with θ'''.

This example is important because it illustrates the impact that the investor's default narrative has on the advisor's narrative construction. As the (competing) default narrative gains explanatory power, the advisor sacrifices belief movement in order to improve the fit of her narrative. The reason is intuitive: if she anticipates that the investor will have access to another narrative that fits very well, she has to send a narrative that fits well to "beat" it. In contrast, if the investor only has access to a competing narrative that fits poorly, she can "beat" it easily, which provides her with the flexibility to be more ambitious on the movement dimension.

Prediction 4 (Responding to a competing narrative)**.** *Let the investor's default narrative be known to the advisor and fix $\theta^{I,0}_{post}$. Then, the following properties hold:*

(i) (Competing narrative constrains fit) The likelihood fit of $\boldsymbol{m}^{A}$ increases in the likelihood fit of $\boldsymbol{m}^{I,0}$.

(ii) (Competing narrative constrains movement) The distance between the advisor's persuasion target and the advisor's θ^{A}_{post} increases in the likelihood fit of $\boldsymbol{m}^{I,0}$.

4 Experimental Design

This section describes the experimental implementation of the investor-advisor setup. In designing the experiment, we aimed to construct a controlled environment that allows us to cleanly test the predictions of the previous section. For identification, we will often rely on within-treatment variation that is provided by the exogeneously assigned advisor incentive-types. For example,

[17] 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. This is not just the case in this example; it is generally true for all histories considered in our experimental setup. Conditional on a particular structural change value, c, the data-optimal θ_{pre} and θ_{post} are always equal to the proportion of successes in their respective periods.

when investigating whether narratives are persuasive, we examine whether investors' assessments depend on the type of advisor they meet.

4.1 Asymmetric and Symmetric Treatments

We start by introducing two of our three core treatments, the Asymmetric and Symmetric treatments. In both treatments, participants are assigned to the role of either the investor or the advisor and they remain in that role for the entire ten rounds of the experiment. In a typical round, they are matched into investor-advisor pairs. Each pair observes historical data from a hypothetical company. This data shows whether the company was "successful" or "unsuccessful" in each of the past ten years. These years are labeled Year 1 to Year 10. The key difference between the two treatments is that, in Asymmetric, advisors are more informed about the company than investors, while both players are provided with identical information about the company in Symmetric.

4.1.1 Choices and Incentives

Choices: To influence the investor's assessment of the data, the advisor can send him a message. This message consists of the advisor's proposed narrative—i.e., her choice of the parameters $(c, \theta_{pre}, \theta_{post})$. When composing the message, the advisor can choose any combination of the three parameters that are part of $\mathcal{M}$. Specifically, the advisor can choose any integer which lies between 2 and 8 for c, and any integers between 0 and 100 to express a percentage value for each of θ_{pre} and θ_{post}.[18]

The investor receives the message from the advisor and can inspect the data himself. He then submits his own assessment of the company's current probability of success, θ_{post}.

Incentives: Both players' payments depend on the investor's decision. The investor's payment is increasing in the accuracy of his own θ_{post}-assessment. The advisor's payment also always depends exclusively on her matched investor's θ_{post}-assessment. Importantly, however, advisors are randomly assigned to one of three incentive conditions. Every advisor is either: (a) an up-advisor, whose payoff is increasing in the investor's assessment of θ_{post}, (b) a down-advisor, whose payoff is decreasing in the investor's assessment of θ_{post}, or (c) an aligned advisor, whose payoff is increasing in the accuracy of the investor's assessment of θ_{post}.

We incentivize participants by providing them with the opportunity to win a bonus payment. The probability of winning the bonus is given by the corresponding payoff functions specified in equations (1) and (2): The investor's probability of winning is maximized if his assessment is equal to the true value, θ^T_{post}, and he suffers a quadratic loss (in probability) when moving away from it. This implies that the investor is incentivized to set his assessment equal to his true belief about θ_{post}. The advisor's probability of winning is maximized if the investor's assessment is equal

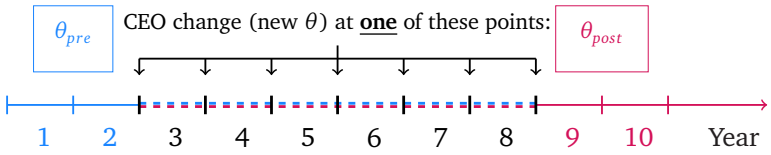
[18] In the instructions for the experiment, the year of the structural change was described as denoting the *first* year under the new CEO. Therefore, advisors could actually choose numbers between 3-9. For expositional clarity and coherence between the discussion of the theory and the experiment, throughout 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.

to the advisor's persuasion target. She also suffers a quadratic loss (of probability) as the investor's assessment moves away from her persuasion target. By mapping the investor's assessment into the probability of winning the bonus, the implemented payoff functions are essentially (strategic) versions of the binarized scoring rule (BSR; Hossain and Okui, 2013).[19]

4.1.2 Information Environment

The data generating process: The investor and the advisor are both informed about the structure of the data-generating process (DGP). Under this DGP, the probability that the company is successful in each year is determined by an underlying fundamental parameter, θ. This fundamental is drawn randomly at two points in time in the company's ten-year history. The initial draw takes place before Year 1. Thereafter, the fundamental is redrawn once at some point after Year 2 and before Year 9. The investor and advisor know this. They are also truthfully told that both the initial probability of success (θ_{pre}) and the current probability of success (θ_{post}) are each drawn from a uniform distribution, $\theta_x \sim U[0,1]$. Likewise, they are informed that the year of the structural change is drawn from a discrete uniform distribution—i.e., with an equal probability of drawing each of the years 2 to 8. All three parameter values are independent of one another. In the experiment, the structural change is framed as a change of the CEO of the company. Figure 3 illustrates the structure of the historical data.

Figure 3: Structure of the historical data



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

Investors do not know the parameters of the true DGP, $\boldsymbol{m}^T = (c^T, \theta^T_{pre}, \theta^T_{post})$, when forming their assessment. They only observe the data and the advisor's message. On the advisor side, we consider two different information conditions in which the advisor is either *informed* or *uninformed* about the true DGP. This constitutes the key treatment difference between Symmetric and Asymmetric. More specifically, we consider the following two scenarios:

[19]The 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 the advisor receiving 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. It, 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.

(i) **A Pure Interpretation Scenario.** In SYMMETRIC, advisors are provided with exactly the same information about the DGP as investors—they are also completely uninformed about the parameters of the true DGP. This is common knowledge. Therefore, when an investor receives a message from the advisor in SYMMETRIC, he knows that it was based only on the (commonly observed) company data; not on any additional information about the true DGP.

(ii) **A Hybrid Scenario.** In ASYMMETRIC, advisors are fully informed about the true parameters of the DGP before constructing their messages. This is common knowledge. Therefore, when an investor receives a message in ASYMMETRIC, he knows that it might have been informed by the advisor's superior information about the true DGP.

Strategic Information about Incentives: Investors are fully informed about the different types of advisors that they may face. Specifically, they are told about the incentives of 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, investors are not informed about the specific incentives of the advisor that they are matched with in a particular round.[20]

Feedback. Neither investors nor advisors receive feedback before the end of the experiment. This means that they do not learn anything about their payoffs until the end of the experiment. In addition, advisors do not receive feedback about their matched investors' assessments between rounds. We do this to minimize learning and the potential interdependence of choices across the ten rounds of the experiment.

4.2 COMPETITION Treatment

Our third core treatment is the COMPETITION treatment. COMPETITION is similar to SYMMETRIC, with the key difference being that we introduce an additional robot advisor to the setting. The robot advisor also sends a message to the investor. Therefore, in this treatment, investors receive two messages—one from the robot advisor and one from the human advisor. The robot's message is determined by a strategy chosen by us, the experimenters, which means we can exogenously vary features of this competing narrative. The aim of the COMPETITION treatment is to allow us to cleanly identify some of the mechanisms behind persuasion and narrative construction (i.e., the influence of the fit of a competing narrative on the advisor's narrative construction, and on

[20]The experimental instructions also take care to fix higher-order beliefs. However, we adopted a slightly different approach to fixing them in ASYMMETRIC and SYMMETRIC. In SYMMETRIC, advisors know that investors do not know the incentives of the particular advisor they are matched with in any given round. They also know that investors know the distribution of incentive-types they might encounter. Conversely, investors are told that advisors know that investors do not know the specific incentive type of the advisor they are matched with in a given round. In ASYMMETRIC, advisors are told that investors "may or may not know" their specific incentives in a given round, and investors know that advisors know that investors "may or may not know" their specific incentives in a given round. The rationale behind this design feature in ASYMMETRIC is that it allows us to introduce our DISCLOSURE treatment, where the advisor's incentives are disclosed to investors in every round. By doing this, we can identify how investors react to the disclosure of incentives, while keeping the advisor's instructions completely constant (and avoiding any deception). Appendix Section A discusses this treatment in more detail. Importantly, none of our analysis will involve a direct statistical comparison of results from SYMMETRIC and ASYMMETRIC. For this reason, we do not view the adjustment of more than one experimental design feature between these two treatments to be problematic.

investor adoption). We discuss how the insights we gain from this treatment relate to those from ASYMMETRIC and SYMMETRIC in the next section.

In each of the five rounds of COMPETITION, investors, advisors, and robots are matched randomly. Each round proceeds as follows. First, the robot constructs a message $\boldsymbol{m}^R = (c^R, \theta^R_{pre}, \theta^R_{post})$ after observing the company data and the true DGP (we describe the robot's strategy in the next paragraph). Second, the human advisor then observes the company data and the robot advisor's message. She then constructs her own message to send to the investor. Third, the investor observes the company data and both advisors' messages. These messages are labeled as coming from "Advisor A" or "Advisor B", and investors do not know which of these labels refers to the human or the robot. The investor must then choose one of the two messages. He is incentivized to pick the message with the more accurate *claim* (i.e., θ_{post}—value). As in the previous treatments, the human advisor takes one of three incentive types: She is incentivized to induce the investor to choose a message that contains either an accurate, low, or high assessment of θ_{post}.[21]

The robot advisors construct their messages, $(c^R, \theta^R_{pre}, \theta^R_{post})$, in the following way: They always send the true value of the company's current probability of success ($\theta^R_{post} = \theta^T_{post}$). Between robot advisors, however, we vary how they choose the *explanation*, (c^R, θ^R_{pre}). Specifically, in Round 1 of COMPETITION, robot advisors always send the true values for θ^T_{pre} and c^T. This implies that, in Round 1, one of the two messages received by the investor is the true DGP, $(c^T, \theta^T_{pre}, \theta^T_{post})$. This allows us to examine how often the human advisor can *beat the truth*. In Rounds 2-5, the robot advisor chooses the two auxiliary parameters (the *explanation*) to either fit very well or fit rather poorly. In two of the four rounds, the robot advisor chooses auxiliary parameters with a HIGH fit: It calculates the data-optimal (best-fitting) values of θ_{pre} and c, *conditional* on the company data and θ^T_{post}. In the other two rounds, the robot advisor chooses auxiliary parameters with a LOW fit: It draws θ_{pre} and c randomly from the uniform distributions, $U[0,1]$ and $U\{2,8\}$, respectively. In designing the experiment, we aimed to carefully control the information environment to allow us to compare the HIGH and LOW fit scenarios as cleanly as possible. We do this by ensuring that for every company history that participants encounter in Rounds 2 to 5, we have pairs of observations where one robot advisor chooses the auxiliary parameters with a HIGH fit and the other chooses the parameters with a LOW fit.[22]

Participants are not provided with a precise description of exactly how the robots construct their messages. Rather, they are told that the robot advisor is "always trying to help" the investor

[21] In particular, we used the same (strategic versions of) the BSR to determine the probability of an investor or advisor winning the bonus that we used in ASYMMETRIC and SYMMETRIC.

[22] To complete the description of the robot advisor's strategy, it is necessary to mention the following three additional details: First, to avoid human advisors being able to easily detect the robot when sending data-optimal auxiliary parameters, we added a small noise term $\eta \sim U[-.03, .03]$ to θ_{pre} in the robot's message. The rationale for this is that setting the data-optimal θ_{pre} is exactly equal to the success frequency in the *pre* period might allow an observant participant to recognize this matching and detect the robot. Introducing small perturbations to θ_{pre} reduces this risk. Second, if a robot (data optimal or random) had a θ_{pre}—value of 0 or 1, we replaced it with a value that was randomly drawn from either $U[.01, .1]$ or $U[.9, .99]$, respectively. Third, if one of the auxiliary parameters of the message generated by a robot's random strategy coincided with the corresponding auxiliary parameter in the message generated by a robot's data-optimal strategy for that same history and true DGP, we replaced the message generated by the random strategy with a new randomly generated message until none of the auxiliary parameters coincided. This was to ensure that the random message was different and had a lower fit.

to achieve his goal of being accurate. In addition, they are told that "not all robot advisors are equally skilled" in forming an accurate assessment of the company. The rationale for this design decision was that we wanted to provide participants with a description of the robot advisor's strategy that was accurate and easy to understand; and to avoid providing them with a detailed and complex description of the algorithm followed by the robot advisor.

4.3 General Comments about the Design

There are several features of our experimental design that warrant further explanation.

Asymmetric vs. Symmetric. In the Asymmetric treatment, there are two potential channels for influential communication. The first is persuasion by influencing the interpretation of data, as highlighted by the narrative approach to communication. The second is persuasion by drawing on expertise, as highlighted by game-theoretical approaches. In contrast, Symmetric only leaves room for persuasion through the data-interpretation channel. Therefore, Symmetric provides a clean test of the pure data-interpretation channel. However, Asymmetric also offers valuable additional insights. In particular, it aims to replicate an essential feature of advisor-investor relations in real life, namely that advisors are typically better informed. In Asymmetric, we achieve this information gap by informing advisors of the true DGP. In addition to capturing this advisor-investor information asymmetry, this serves several purposes. First, it allows us to exogenously fix advisors' beliefs about the true narrative, and study narrative construction conditional on these beliefs. Second, it creates a clear normative distinction between messages that are truth-telling and those that are lies. 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 provide advice that is accurate to the best of their knowledge. In such settings, it is very difficult to observe whether an individual is lying relative to their privately formed beliefs. 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. This complements the insights that can be gained from the Symmetric treatment, which provides a clean test of the data interpretation-based persuasion channel in a context where these normative expectations are not fixed.

Humans and robots. We employ both human and robot advisors in the experiment because they deliver complementary insights. In Asymmetric and Symmetric, investors know the incentive and information environment in which the (human) advisors construct their messages. This allows them to form beliefs about the motives behind the messages they receive. This is key if we want to use our design to examine persuasion not only in relation to the narrative approach, but also the game theoretic approach, which emphasizes the role of beliefs in strategic communication. For example, if we relied solely on robot advisors and remained opaque about how they construct

their messages, we would not be able to establish a benchmark for how a strategically sophisticated investor should interpret them.[23] In addition, having human advisors allows us to not only study how humans interpret messages but also how they construct them. However, importantly, one advantage of adding robot advisors in COMPETITION is that we are able to exogenously control the messages constructed by these (robot) advisors. Furthermore, we also gain some control over the alternative narratives that investors consider when assessing the human advisor's message: First, by comparing instances where investors see the same data and the robot's choice of θ^R_{post} is the same, but where the robot's choice of auxiliary parameters is either data-optimal or random, we can assess how the empirical fit of a narrative determines narrative adoption (i.e., it allows us to test Prediction 2 *(ii)*). Second, by comparing how human advisors construct messages in situations where everything except the auxiliary parameters of the robot's message is the same, we learn about the factors determining the human advisor's narrative construction (i.e., it allows us to test Prediction 4).

The investor's default narrative. We chose not to elicit investor's prior beliefs about the default model (i.e., the belief based on seeing only the historical data) in any of our main treatments. 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.

4.4 Procedures

The experiment was programmed in oTree (Chen, Schonger, and Wickens, 2016) and participants were recruited via the Prolific platform. Participants in the experiment were balanced by gender.[24] In designing the experiment, we devoted substantial attention to 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

[23]The alternative—not remaining opaque about the strategy of the robot advisor—comes with a different problem: An experiment which carefully explains the robot's strategy to human investors can only investigate whether human investors best-respond to a pre-specified mapping from states of the world into messages. This would capture only a small subset of the mechanics we are interested in when thinking about how persuasion via messages operates between two (human) individuals.

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

correctly before proceeding.[25] We preregistered the experiment and provide a populated preregistration for interested readers (see Banerjee, Duflo, Finkelstein, Katz, Olken, and Sautmann, 2020, for a general discussion of populated preregistrations).[26]

We recruited 360 participants (180 advisors and 180 investors) to participate in the Asymmetric treatment in March 2022. In June 2023, we recruited another 360 participants (180 advisors and 180 investors) to participate in Symmetric and Competition. The Symmetric and Competition treatments were conducted within-participant, i.e., after finishing the ten rounds of Symmetric, participants received the instructions for Competition and then completed five rounds of that treatment. (Note, the setup of Competition built on the basic setup of Symmetric, adding the robot advisor.)

Participants took part in the experiment in groups of 6. Within each group, 3 participants were randomly assigned to the role of the advisor and 3 are assigned to the role of the 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.

In every round of the experiment, each investor was randomly matched with an advisor within their group of 6 (i.e., the three investors were randomly matched with the three advisors). All matched investor-advisor pairs across all groups saw data generated by the same true underlying DGP in each round of Asymmetric and Symmetric. Specifically, we drew ten triplets of fundamentals, $(c^T, \theta^T_{pre}, \theta^T_{post})$, before the first session of Asymmetric. The sequence in which participants were exposed to each underlying true DGP was constant in all sessions of Asymmetric and Symmetric. For Competition, we drew new true DGPs. In Round 1 of Competition, we randomly drew one true DGP for every matching group of investors and advisors. In Rounds 2-5, we drew one true DGP for each round.

Conditional on the true DGP, the observed historical data of success and failure of the company was drawn independently for each investor-advisor pair and round in Asymmetric. In Symmetric and Competition, we further refined our design to gain greater control and enhance the comparability of the information sets of specific different participants (e.g., to compare participants with identical information sets, but different incentives). Specifically, in Symmetric and in Round 1 of Competition, we instead randomly drew the historical data on the matching group level. This meant that we had three advisors with different incentives who observed the same historical data. This allows us to control for Round×History fixed effects when analyzing data from Symmetric. In rounds 2-5 of Competition, we drew one historical data set for every two matching groups. This allowed us to take two human advisors with the same incentives—one in each matching group—and then to pair one of these advisors with a robot advisor which chose

[25]Instructions for all treatments can be found under the following link: https://tilmanfries.github.io/assets/pdf/NarrativePersuasionInstructions.pdf

[26]The populated preregistration contains a discussion of the full set of preregistered analyses. This populated preregistration, as well as the original preregistration documents, can be accessed via the following link: https://tilmanfries.github.io/assets/pdf/NarrativePersuasionPreregistration.pdf. The preregistrations for each of the two waves of data collection reported in this paper were originally uploaded to the AEA registry and can be located using the unique identifiers AEARCTR-0009103 and AEARCTR-0011565.

the auxiliary parameters of its narrative in a data-optimal way, while the other human advisor was paired with a robot advisor which chose them randomly. Therefore, when controlling for Round×History fixed effects in the analysis of rounds 2-5 of COMPETITION, we can essentially isolate the variation in assessments and narrative construction that was caused by variations in the robot's choice of auxiliary parameters, keeping other aspects of the decision environment fixed.

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 bonus payment depended on the relevant binarized scoring rule described above, which was evaluated in relation to the investor's assessment. After finishing all rounds of their session, participants answered a short demographic questionnaire. Participants took around 20-25 minutes for ASYMMETRIC and 30-35 minutes for SYMMETRIC plus COMPETITION.

5 Main Results

In this section, we present our key results. Our empirical exercises reported in this section use the within-treatment variation generated in our SYMMETRIC, ASYMMETRIC and COMPETITION treatments. In Section 6, we discuss the results from additional treatments that serve to extend and provide robustness checks for these core results.

5.1 Investor Behavior (Narrative Adoption)

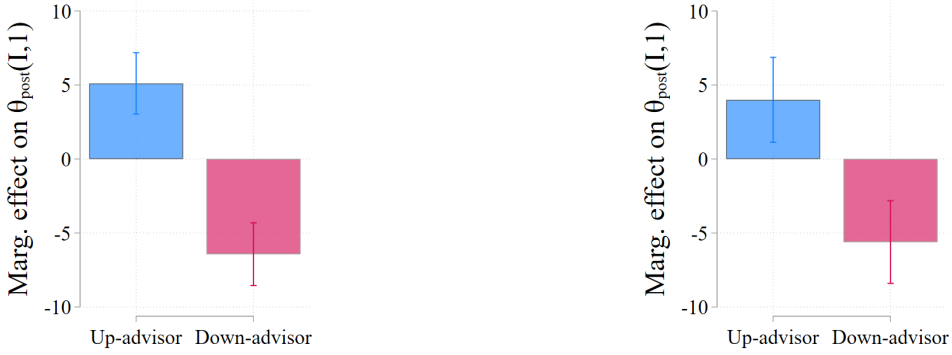
We first discuss the behavior of investors (Predictions 1 and 2), examining whether narratives shift investors' beliefs and evaluating the role of narrative fit.

Persuasion in the hybrid and pure interpretation scenarios: A key question of interest is whether persuasion using narratives is effective (Prediction 1): Are advisors able to successfully distort investors' beliefs? Figure 4 addresses this question, showing that investors' beliefs are systematically shifted towards the advisors' persuasion targets. The figure reports the coefficients from regressing the investor's posterior beliefs, θ^I_{post}, on indicator variables for the up- and down-advisor. The left panel displays the estimated coefficients using data from the ASYMMETRIC treatment, while the right panel uses data from our SYMMETRIC treatment. In both treatments, investors report higher beliefs when matched with an up-advisor and lower beliefs when matched with a down-advisor. Figure B.1 in the Appendices shows that this gap in beliefs due to persuasion is present across the full distribution, since beliefs of investors matched with up-advisors stochastically dominate those of investors matched with down-advisors. This evidence shows that advisors are able to use narratives to persuade investors to shift their beliefs. Narrative persuasion is effective.

Result 1 (Persuasion in pure interpretation and hybrid scenarios)**.** *The narrative sent by the advisor shifts the investor's assessment towards the advisor's persuasion target in both the ASYMMETRIC and*

Symmetric treatments. This indicates that narrative persuasion is effective in both the hybrid and pure interpretation scenarios.

Figure 4: Effect of the advisor type on investor assessments



(i) Asymmetric (ii) Symmetric

Notes: (i) The figure reports the coefficients from regressing the investor's assessment θ^{I}_{post}, on indicator variables for the up- and down-advisor, *(ii)* Error bars represent 95% confidence intervals that were derived from regressions that cluster standard errors at the matching group level, *(iii)* In the regression for the Asymmetric treatment (left panel), we include round fixed effects, *(iv)* In the regression for the Symmetric treatment (right panel), we are able to include round×history fixed effects due to the more sophisticated experimental design used in that treatment, *(v)* The regression output is reported in the Appendices in Table B.2.

Influence of narrative fit: Having shown that narrative persuasion is effective, we illustrate which types of narratives influence investors the most. According to the S&S framework, investors find narratives that are more coherent with the empirical data to be more plausible. This means that investors will be more willing to believe a narrative that fits the empirical data well (Prediction 2). In this section, we test this assertion.

We quantify narrative fit using a metric that we refer to as the Empirical Plausibility Index (EPI). To derive the EPI, we calculate the likelihood value of each narrative, given the relevant historical data. The EPI is then equal to this likelihood value divided by the likelihood value obtained by the data-optimal (best-fitting) 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 best-fitting narrative and the minimum value of 0 is obtained if the advisor sends the worst-fitting narrative.[28] We use the EPI to investigate the relation between fit and persuasion in several exercises.

First, we use the exogenous variation in the fit of the robot advisor's narrative that we induce in our Competition treatment. Here, we ask whether investors choose to follow the human advisor's narrative more often when the fit of the competing robot advisor's narrative is exogenously worsened.[29] Essentially, we examine how the probability that the investor follows the human ad-

[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 page 24 in Section A.3 and also on pages 33-35 in Section A.5.

[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., there will always be a narrative containing either $\theta_{pre} = \theta_{post} = 0$ or $\theta_{pre} = \theta_{post} = 1$ that will have a likelihood value of zero.

[29]Remember that, in Competition, investors face a binary choice between either adopting the human advisor's narrative or the robot's narrative, without knowing which is which. They are incentivized to choose the narrative with the more accurate θ_{post}. Therefore, we measure investor adoption directly by recording this binary decision.

visor's narrative changes when the fit of the competing robot narrative is exogenously increased from a Low to a High fit.

Table 1: Likelihood of adopting the human advisor's narrative in Competition

	(1) $\mathbb{I}$(adopt m^A)	(2) $\mathbb{I}$(adopt m^A)
$\mathbb{I}$(Robot Narrative Fit = High)	-0.0778***	
	(0.0235)	
EPI of Robot Narrative (cont. fit)		-0.138***
		(0.0447)
Round× History FE	Yes	Yes
Observations	720	720

Notes: (i) The dependent variable, $\mathbb{I}$(adopt m^A), is an indicator variable that takes a value of one when the investor chooses to adopt the human advisor's narrative, (ii) The independent variable, $\mathbb{I}$(Robot Narrative Fit = High), takes a value of one when the auxiliary narrative components are chosen to be optimal, and a value of zero when they are chosen randomly, (iii) Standard errors are clustered at the matching group level, implying that there are 60 clusters, and are reported in parentheses; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

In Column (1) of Table 1, we do this by regressing a binary variable that indicates that the investor chose the human advisor's narrative, $\mathbb{I}$(adopt m^A), on an indicator variable that takes a value of 1 when the robot advisor's narrative fit is High. The coefficient shows that the investor is 7.8pp less likely to choose the human advisor's narrative when the robot advisor proposes a narrative that fits well. Crucially, we exogenously vary the fit of the robot advisor's narrative using only the supporting narrative components, c and θ_{pre}. We hold both the historical company data and the θ_{post} sent by the robot advisor constant. This means that we can interpret the reduction in investors' willingness to adopt the human advisor's narrative as the causal influence of a reduction in the fit of the supporting narrative components. Therefore, this provides direct causal evidence that narrative fit is a key determinant of narrative adoption. Column (2) of Table 1 provides further support for this conclusion by showing that when we replace the binary variable for fit with a continuous measure (i.e., the EPI), we get the same result: the better the fit of the robot advisor's narrative, the less likely the investor is to follow the narrative of the human advisor.

As a second exercise, in Table B.3 in Appendix B.2, we show that in our Symmetric and Asymmetric treatments, the better the fit of the narrative sent by the advisor, the closer the investor's assessment of θ_{post} is to the θ^A_{post} sent by the advisor. Essentially, we show that as the EPI of the advisor's narrative increases, the gap between the investor and advisor assessments, $|\theta^I_{post} - \theta^A_{post}|$, gets smaller. Last, to examine belief updating by investors, we conducted an additional treatment, InvestorPrior, which is similar to our Asymmetric treatment except that we elicit investors' prior beliefs before they meet their advisor. Using this treatment, we show that investors update their beliefs more when the advisor sends a narrative that fits better. We discuss the details of this treatment in more detail in Section A.1.

Result 2 (Influence of narrative fit)**.** *As the fit of a narrative increases, the investor becomes more likely to adopt it. This is the case even when θ^A_{post} is held constant and the fit of the narrative is*

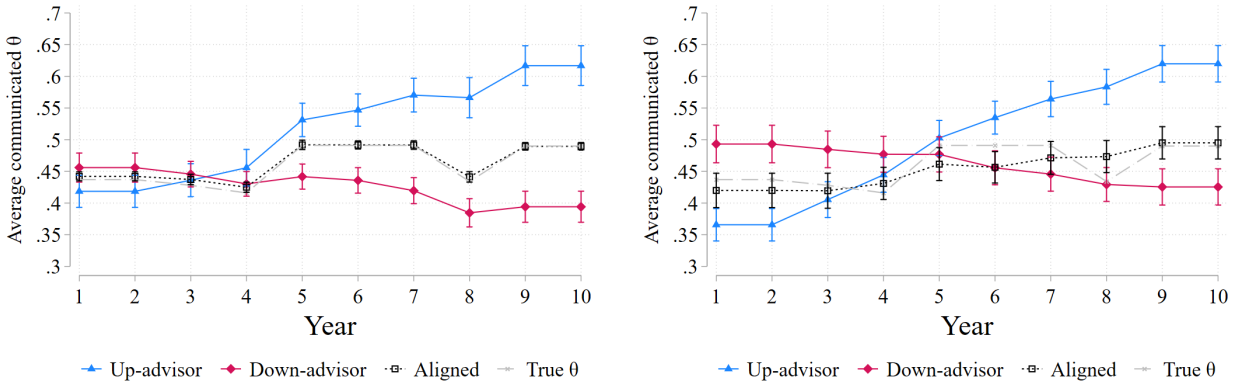
exogenously varied by only changing the supporting narrative components. Therefore, the fit of a narrative is a key determinant of narrative adoption.

5.2 Advisor Behavior (Narrative Construction)

We now turn to advisors with the aim of identifying systematic regularities in the features of the narratives that advisors construct (Predictions 3 and 4).

Fit-movement tradeoff in narrative construction: In our theoretical framework, advisors face a tradeoff between belief *movement* and empirical *fit* when constructing narratives. This arises from advisors anticipating that investors are more willing to believe narratives that fit the empirical data well. An implication of this is that advisors are predicted to construct narratives with a negative correlation between θ^A_{post} and θ^A_{pre}, since they shift θ^A_{post} towards their self-interest (*movement*) and use θ^A_{pre} to improve the narrative *fit* (Prediction 3). Here, we test this prediction.

Figure 5: Average narrative communicated by advisors (by advisor type)



(i) Asymmetric (ii) Symmetric

Notes: The left panel presents the average narratives of advisors in Asymmetric and the right panes presents average narratives of advisors in Symmetric. Error bars represent 95% confidence intervals that were derived from regressions which cluster standard errors at the advisor level.

Figure 5 provides a visual illustration of how advisors of different types construct narratives. The figure depicts the average narrative transmitted by advisors of each type.[30] We see that, in line with Prediction 3, misaligned advisors bias θ^A_{post} and θ^A_{pre} in opposite directions in both the Asymmetric and Symmetric treatments. Up-advisors (denoted by the blue line) construct narratives with a *higher* θ in year 10 than down-advisors (denoted by the red line). Conversely, up-advisors choose narratives with a *lower* θ in year 1 than down-advisors. This shows that, on average, misaligned advisors shift θ^A_{post} towards their persuasion target and use θ^A_{pre} to justify it in view of the historical company data.

To provide further support for this finding, in Table B.5 in the appendices, we report regression results indicating that up-advisors send higher θ^A_{post} and lower θ^A_{pre}-values than aligned advisors

[30]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 θ_{pre} in the period before the CEO change and θ_{post} in the period after the CEO change. To obtain Figure 5, we take the average θ for each year across all messages sent by advisors of each type.

in both treatments, while the opposite is true for down-advisors. Finally, Figure B.3 in Appendix B.3 shows visually that, in each of the ten rounds, the up-advisor sends a higher θ^A_{post} and lower θ^A_{pre} than the down-advisor.

Result 3 (Fit-movement tradeoff in narrative construction)**.** *Misaligned advisors shift the θ^A_{post} of their narrative towards their persuasion target and shift θ^A_{pre} in the opposite direction. This suggests that advisors anticipate the key role of narrative fit for investor adoption, using θ^A_{post} to try to shift investors' beliefs and θ^A_{pre} to improve the fit of the narrative.*

Responding to a competing narrative: In this section, we investigate whether advisors adjust their narrative construction in response to their beliefs about the narrative they are competing with. The S&S framework predicts that when advisors believe that they are competing with a narrative that fits the data well, they will be more constrained in their narrative choice in comparison to when they compete with a narrative that fits poorly (Prediction 4). The key reason is that in order to be persuasive, they need the narrative they send to appear more plausible than the alternative.

Table 2: How narrative fit and bias depends on the fit of the competing narrative

	(1) EPI^A	(2) EPI^A	(3) EPI^A	(4) EPI^A	(5) Bias	(6) Bias
Competing EPI	0.285***	0.286***	0.301***	0.301***	-4.787*	-5.260**
	(0.0259)	(0.0264)	(0.0346)	(0.0353)	(2.403)	(2.516)
Round× History FE	Yes	No	Yes	No	Yes	No
Round FE	No	Yes	No	Yes	No	Yes
History FE	No	Yes	No	Yes	No	Yes
Incl. round 1	No	Yes	No	Yes	No	Yes
Included advisor types	All	All	Misaligned	Misaligned	Misaligned	Misaligned
Observations	720	900	480	600	480	600

Notes: (i) The dependent variable is either the human narrative's *fit* (EPI^A) or the human narrative's *bias*, which is defined as the absolute distance between the advisor's persuasion target (1 for up and 0 for down) and θ^A_{post}, (ii) The main regressor is the fit (EPI) of the competing robot advisor's narrative, which is exogenously varied while holding θ^R_{post} constant, (iii) The sample contains data from all advisors in COMPETITION, (iv) For each advisor we have 5 observations—one for each round, (v) Due to the structure of the experimental design, in our regressions that exclude the Round 1 data (where the robot advisor reports the true narrative), we are able to include Round×History fixed effects. In other regressions, we can include round fixed effects and history fixed effects, (vi) Standard errors are clustered at the matching group level and are reported in parenthesis; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

Our COMPETITION treatment allows us to directly examine how advisors react to the fit of a competing narrative, because we exogenously vary the fit of the robot advisor's narrative while keeping the θ^R_{post} constant. Since the human advisors observe the narrative they are competing with before choosing their own, we can assess the causal influence that this competing narrative exerts. Table 2 shows how the fit of the competing narrative influences the *(i)* fit, and *(ii)* bias, of the narrative chosen by the human advisor. Columns (1) to (4) show that when the fit of the competing robot narrative increases, the human advisor also chooses a narrative that fits better. Essentially, the human advisor needs to choose a narrative that fits better in order to "beat" the robot advisor's narrative and be persuasive. Consequently, as the fit of the robot advisor's narrative improves, the human advisor becomes more constrained in the set of potential better-fitting narratives she can select from. A predicted implication of this is that human advisors

are forced to become less ambitious about their intended belief movement when the fit of the competing narrative increases. Columns (5) and (6) provide evidence in support of this prediction, showing that when the fit of the competing robot advisor's narrative improves, misaligned human advisors choose a θ^A_{post} that is less biased towards their persuasion target.

These findings clearly illustrate how advisors systematically adjust the narratives they construct in response to a shift in the fit of the narrative they are competing with. Taken together, they provide strong support for the systematic patterns in narrative construction behavior predicted by the S&S framework, highlighting the usefulness of this framework for analyzing the use of narratives as a persuasive tool.

Result 4 (Responding to a competing narrative)**.** *When an advisor constructs a narrative, they are constrained in their choice by the fit of the competing narrative. When the fit of the competing narrative increases, advisors construct narratives that (i) fit better, and (ii) are less ambitious about how far they try to move the investor's belief towards their persuasion target.*

6 Extensions and Robustness Exercises

In the previous section, we discussed the results from our three core treatment conditions. These analyses test our main predictions. To augment these findings and shed further light on the robustness and underlying mechanisms of narrative persuasion, we conducted several additional exercises.

First, to investigate the sensitivity of our findings with respect to specific features of the choice environment, we conducted a series of three "intervention" treatments. In each of these treatments, we altered one aspect of the choice environment to test whether we could reduce the persuasive influence of narratives in our setting. These exercises serve two purposes: *(i)* they provide evidence regarding the robustness of our results to specific experimental design choices, *(ii)* they yield lessons about whether and how one might protect narrative-recipients from harmful persuasion. We find that none of the three interventions has a statistically significant protective effect for the average investor. This suggests that narrative persuasion is fairly robust to contextual factors. Appendix Section A.1 contains a detailed discussion of these treatments and results.

Second, to provide further evidence on the role played by *explanations* (i.e., the auxiliary narrative components, c and θ_{pre}), we implemented a pair of treatments in which we exogenously varied whether these two parameters were revealed to the investor or not. We find that the quality of explanations matters—claims supported by a good explanation are more persuasive than claims supported by a bad explanation. We discuss these treatments and the associated analysis in more detail in Appendix Section A.2.

Third, we relax the assumption that decision-makers (advisors and investors) are perfectly consistent when choosing between narratives, acting without error. To do this, we estimate decision noise in a simple discrete choice model that allows us to shed light on the role it plays in the strategic interaction between advisors and investors. Perhaps unsurprisingly, we find that decision-making does contain noise. More interestingly, we also find that advisors appear to antic-

ipate the noise in investors' decisions, accounting for it when constructing their narratives. This exercise is discussed in Appendix Section A.3.

7 Concluding Discussion

The discussion above has provided empirical evidence showing how narratives can be used as a tool for persuasion, with one individual shaping the way that another interprets objective data. This analysis is relevant for the class of situations where there exists some public information that can accommodate more than one possible interpretation, potentially allowing some individuals to try to encourage others to adopt their preferred interpretation. Our results are largely in line with the persuasion mechanics outlined in the S&S theoretical framework. Specifically, when examining investor behavior, we find that *(i)* exposure to narratives shifts their beliefs, and *(ii)* the degree to which their beliefs are shifted increases in the narrative's empirical fit. These results hold in both a *pure interpretation scenario*, where it is common knowledge that the advisor does not have any additional private information viz-a-viz the investor, and in a *hybrid scenario*, in which advisors do hold additional information.

Turning to advisors, an important feature of both these scenarios is that the advisor can construct the narrative *ex-post*, tailoring it to the public data. This ex-post tailoring means that she can construct a narrative that fits the data well and can then, in turn, 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 their *claim* in the direction of their private self-interest and use their *explanation* to make their narrative more plausible by improving its fit. This behavior is consistent with a narrative construction approach that trades off *fit* and *movement*. We also show that the presence of a competing narrative predictably constrains the advisor's narrative construction: As the fit of the competing narrative improves, advisors send better-fitting and less biased narratives themselves.

While our experiment is designed to study persuasion through the lens of S&S, an additional advantage of our setup is that we can also test whether the behavior we observe is consistent with the Nash equilibrium predictions of the underlying cheap talk game. Our results provide several pieces of evidence that relate to these Nash equilibrium predictions. First, in cheap talk models, the key factor that sustains persuasive equilibria is an information asymmetry between the sender and the receiver; without this information asymmetry, persuasion is not sustained in equilibrium. Therefore, our finding that narratives remain persuasive in our SYMMETRIC treatment, where there is no information asymmetry, is not consistent with a traditional cheap talk persuasion channel. Second, cheap talk equilibria typically do not ascribe a persuasive role to auxiliary (non-payoff-relevant) parameters. In a model with a strategically sophisticated sender and receiver, when the receiver receives a message, he will not evaluate the fit of the message; rather, he will try to assess which type of sender sent it. In equilibrium, sophisticated advisors who hold misaligned incentives will adjust their messages to appear *as if* they are not misaligned

by mimicking aligned advisors. The consequence of this is that the receiver's reaction to the auxiliary parameters is fully muted in equilibrium. This result carries over to variants of the model that *(i)* relax the receiver's degree of strategic sophistication, and *(ii)* introduce honest senders. In contrast to these predictions, our empirical results clearly demonstrate the relevance of auxiliary parameters (*explanations*) for persuasion in the class of scenarios we consider. In COMPETITION, investors are more likely to adopt a robot narrative with high- rather than low-fitting auxiliary parameters. In addition, in the EXPLANATION treatments (reported in Appendix A), we use variation in the narratives constructed by human advisors in a fully specified strategic environment to show that investors are less persuaded by narratives with auxiliary parameters that fit poorly. To summarize, the persuasion we observe in the experiment goes beyond the channels traditionally highlighted in the cheap talk literature. This underscores the complementarity of the narrative framework viz-a-viz the cheap talk literature—i.e., it offers an additional lens through which to analyze scenarios where persuasion operates via manipulating the interpretation of facts.

Our analysis in this paper has focused predominantly on testing predictions that emerge from the S&S framework. To do this, we constructed a simple setting in which we could capture and vary some core features of narratives in a controlled way. However, narratives are rich and multifaceted and may contain additional features that are excluded in our experiment, such as emotional content and allusions to existing ideas stored in the individuals' memory. They are also extremely important drivers of human behavior. Therefore, there is scope and need for further research to paint a more complete picture of how narratives are formed and communicated between people. And also how they may be used as a persuasive tool. As is often the case when exploring new research areas, our analysis has raised several 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, the success of expert persuaders might not only depend on being more knowledgeable than the individuals they are persuading, but may also be a consequence of being able to skillfully construct narratives that tie together publicly available information. Some individuals might be particularly creative and sophisticated in doing this. Such individuals may disproportionately select into occupations and positions where there is scope to benefit from constructing convincing narratives. This arguably includes a wide range of occupations, such as advertising, politics, consulting, real estate, lobbying, and law. It would be interesting to investigate whether the skill of constructing a persuasive narrative is particularly well developed amongst individuals in these professions (either due to the selection of individuals with that trait into the profession or due to learning the skill within the profession).

Shaping how individuals see data: In the analysis above, we have worked within a theoretical framework in which a receiver either adopts the narrative he receives fully or does not adopt it at

all and maintains his prior understanding of the data. However, it seems reasonable to entertain alternative, weaker assumptions where narratives are not either fully adopted or not adopted, but still influence the investor's beliefs. For example, an investor may not fully adopt the narrative sent to him by an advisor, but may still be influenced by some of its features. In particular, the narrative may shift his attention to particular parts of the data, or may influence the way that he looks at the data.[31] To explore this idea, in Appendix B.7, we present an empirical exercise that examines whether advisors change how investors extract information from the historical performance data: We find that investors' beliefs are influenced more by successful and unsuccessful years in the company data that occur after the advisor's assertion about when the CEO changed, controlling for the fact that investors place more weight on later years.[32] Therefore, while an investor might dismiss what the advisor tells him about θ_{post}, he might still be influenced by the advisor's suggested partition of the data into *pre* and *post*, possibly because it provides him with a previously unconsidered way of seeing the data. A full analysis, however, goes beyond the scope of this paper; it would be valuable to investigate how narratives can be partially influential in future research.[33]

Avoiding narratives that are "too good to be true": The theoretical framework that we use points towards a monotone relationship between the fit of a narrative and its persuasiveness. However, in settings where the narrative sender knows the true DGP (and the receiver knows this), as in Asymmetric, one may expect non-monotonicities around the best-fitting narrative. The rationale for this is the following. Receivers know that advisors know the true DGP. Due to the random nature of the process generating successes and failures, this true DGP is is unlikely to be exactly the best-fitting narrative, given the data. Therefore, when a receiver sees a message from an advisor that contains a narrative with a perfect fit, they might become skeptical and take this as a sign that the advisor is likely to be lying. In Symmetric, this mechanism not present since receivers know that advisors do not know the true DGP and therefore a narrative that fits perfectly is not a signal of bias. While our experiment was not set up to be powered to investigate such skeptical thinking, we explore related evidence in Appendix B.8. There, we

[31]For an analogous example, consider the famous duckrabbit illusion, which contains an ambiguous image that can either be perceived as depicting a duck or a rabbit depending on how the viewer looks at it. If an individual is given the picture and told that the image shows *"a quacking duck"*, this may draw their attention towards seeing the features of a duck rather than a rabbit, thereby shaping how they see the image. However, they may not fully believe the message they received in the sense that they may or may not agree with the assertion that the duck in the picture is quacking.

[32]Specifically, we can examine the marginal effect that a success vs failure in each year of the historical data has on the investors' assessment, θ^I_{post}. For example, one would expect that a success in more recent years (e.g., Years 9 and 10) will have a larger influence than a success in years that are further in the past (e.g., years 1 and 2). Here, we are particularly interested, however, in the marginal effect of intermediate years and whether advisors can change how much information an investor draws from a success in, say, Year 7, depending on whether the advisor assigns this year to the *pre* or the *post* period in her narrative. We find evidence that advisors do directly influence how investors extract information from the data.

[33]A closely related avenue for further investigation would be to consider settings that combine narratives with the (selective) disclosure of evidence. Consider, for example, a variant of our design where the advisor receives the historical company data first and then can decide whether to reveal (parts of) the historical company data to the investor in addition to providing a narrative. In real-world settings, it is often the case that individuals are able to present data along with an interpretation of the data they are presenting. This allows other mechanisms, such as data selection to play a role in persuasion. Such extensions to our setting seem extremely important.

show that, in ASYMMETRIC, advisors rarely send narratives whose θ—parameters perfectly match the empirical success frequencies in the *pre* and *post*-periods implied by their c^A's, while they do match *in expectation*. This picture changes quite substantially in SYMMETRIC, where the perfect (and near-perfect) matching rate is dramatically higher. These findings are consistent with the idea that advisors are worried that their narratives might be perceived as being *too good to be true* in cases where they know the truth. Future research could investigate whether these non-monotonicities in receiver skepticism exist—i.e., whether receivers do disregard narratives that fit perfectly, implying that senders' worries about narratives that are too good to be true are justified. Embarking down this avenue might also yield more general insights into how strategic concerns (i.e., higher-order beliefs about the opponent's strategy) influence individual behavior in a scenario like the one we study.

The relevance of narrative persuasion in the age of the internet and social media: With the arrival of the internet, recent decades have brought near-instant access to a wealth of information to most people. However, it is non-trivial to sort through all of this information and process it to form accurate beliefs. This means that the way that individuals interpret this information is extremely important in determining how they understand the world around them. Alongside this rapid rise in access to information, the proliferation of social media has resulted in a highly interconnected modern society where different narratives proposing interpretations of the available information may be transmitted nearly instantaneously and can spread rapidly through large networks of individuals. In concert, these two changes in society (social media and ever-expanding information access) create conditions conducive to persuasion using narratives. Examining the topic from this perspective raises an array of interesting and important questions. For example, which factors determine the spread and survival of narratives on a social media platform? How is the persuasiveness of narratives affected when the number of available narratives increases? Do the motives of the narrative-recipient influence their susceptibility to believing particular narratives? (e.g., are individuals more willing to believe narrative explanations of the facts that support beliefs that they want to hold?) In addition, Schwartzstein and Sunderam (2024) provide a formal theoretical analysis of how individuals make sense of the world together as a community, which raises further interesting empirical questions. We leave these questions to future work.

References

Abeler, J., D. Nosenzo, and C. Raymond (2019). Preferences for Truth-Telling. *Econometrica 87*(4), 1115–1153.

Aina, C. (2024). Tailored Stories. *Mimeo*.

Alysandratos, T., A. Boukouras, S. Georganas, and Z. Maniadis (2020). The Expert and The Charlatan: An Experimental Study in Economic Advice. *SSRN Electronic Journal*.

Ambuehl, S. and H. C. Thysen (2024). Competing Causal Interpretations: An Experimental Study. *Mimeo*.

Andre, P., I. Haaland, C. Roth, and J. Wohlfart (2023). Narratives About the Macroeconomy. *CESifo Working Paper No. 10535*.

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

Ash, E., G. Gauthier, and P. Widmer (2024). Relatio: Text semantics capture political and economic narratives. *Political Analysis 32*(1), 115–132.

Ba, C. (2024). Robust Misspecified Models and Paradigm Shifts. *Mimeo*.

Banerjee, A., E. Duflo, A. Finkelstein, L. Katz, B. Olken, and A. Sautmann (2020). In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics. *NBER Working Paper 26993*.

Barron, K. and T. Fries (2023). Narrative persuasion. *CESifo Working Paper No. 10206*.

Barron, K., H. Harmgart, S. Huck, S. O. Schneider, and M. Sutter (2023). Discrimination, Narratives, and Family History: An Experiment with Jordanian Host and Syrian Refugee Children. *The Review of Economics and Statistics 105*(4), 1008–1016.

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

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

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

Bruner, J. (1991). The Narrative Construction of Reality. *Critical Inquiry 18*(1), 1–21.

Charles, C. and C. Kendall (2024). Causal narratives. *Mimeo*.

Charnysh, V. (2023). Remembering Past Attrocities—Good or Bad for Attitudes toward Minorities? In J. S. Kopstein, J. Subotić, and S. Welch (Eds.), *Politics, Violence, Memory*, The New Social Science of the Holocaust, pp. 245–266. Cornell University Press.

Chen, D. L., M. Schonger, and C. Wickens (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance 9*, 88–97.

Chen, Y. (2011). Perturbed communication games with honest senders and naive receivers. *Journal of Economic Theory 146*(2), 401–424.

Converse, P. E. (2006). The Nature of Belief Systems in Mass Publics. *Critical review 18*(1-3), 1–74.

Crawford, V. P. and J. Sobel (1982). Strategic Information Transmission. *Econometrica 50*(6), 1431–1451.

Eliaz, K. and R. Spiegler (2020). A Model of Competing Narratives. *American Economic Review 110*(12), 3786–3816.

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

Foerster, M. and J. J. van der Weele (2021). Casting doubt: Image concerns and the communication of social impact. *The Economic Journal*, 2887–2919.

Fong, M.-J., P.-H. Lin, and T. R. Palfrey (2023). Cursed Sequential Equilibrium. *Mimeo*.

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.

Frechette, G., E. Vespa, and S. Yuksel (2024). Extracting models from data sets: An experiment. *Mimeo*.

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

Gehring, K., J. A. Harm Adema, and P. Poutvaara (2022). Immigrant narratives. *CESifo Working Paper No. 10026*.

Graeber, T., C. Roth, and C. Schesch (2024). Explanations. *Mimeo*.

Graeber, T., F. Zimmermann, and C. Roth (2022). Stories, statistics, and memory. *CESifo Working Paper No. 10107*.

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

Hagmann, D., C. Minson, and C. Tinsley (2024). Personal narratives build trust across ideological divides. *Journal of Applied Psychology (forthcoming)*.

Haidt, J. (2007). The New Synthesis in Moral Psychology. *Science 316*(5827), 998–1002.

Haidt, J. (2013). *The Righteous Mind: Why Good People are Divided by Politics and Religion*. Harlow, England: Penguin Books.

Harbaugh, R. and E. Rasmusen (2018). Coarse Grades: Informing the Public by Withholding Information. *American Economic Journal: Microeconomics 10*(1), 210–235.

Harrs, S., L. M. Berger, and B. Rockenbach (2023). How Narratives Impact Financial Behavior. *ECONtribute Discussion Paper No. 91*.

Heidhues, P., B. Kőszegi, and P. Strack (2018). Unrealistic Expectations and Misguided Learning. *Econometrica 86*(4), 1159–1214.

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

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

Hirschman, A. O. (2013). *The Passions and the Interests: Political Arguments for Capitalism before Its Triumph*. Princeton: Princeton University Press.

Hossain, T. and R. Okui (2013). The Binarized Scoring Rule. *Review of Economic Studies 80*(3), 984–1001.

Hüning, H., L. Mechtenberg, and S. Wang (2022). Using Arguments to Persuade: Experimental Evidence. *SSRN Electronic Journal*.

Ichihashi, S. and D. Meng (2021). The Design and Interpretation of Information. *Mimeo*.

Ispano, A. (2023). The perils of a coherent narrative. *Mimeo*.

Jain, A. (2023). Informing agents amidst biased narratives. *Mimeo*.

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

Kamenica, E. and M. Gentzkow (2011). Bayesian Persuasion. *American Economic Review 101*(6),

2590–2615.

Kartik, N., M. Ottaviani, and F. Squintani (2007). Credulity, lies, and costly talk. *Journal of Economic Theory 134*(1), 93–116.

King, R. R. and D. E. Wallin (1991). Market-induced information disclosures: An experimental markets investigation. *Contemporary Accounting Research 8*(1), 170–197.

Koschorke, A. (2018). *Fact and Fiction: Elements of a General Theory of Narrative*. Walter de Gruyter.

Lang, M. (2023). Mechanism Design with Narratives. *CESifo Working Paper Series No. 8502*.

Laudenbach, C., A. Weber, and J. Wohlfart (2021). Beliefs about the stock market and investment choices: Evidence from a field experiment. *CEBI Working Paper 17/21*.

Little, A. T. (2023). Bayesian explanations for persuasion. *Journal of Theoretical Politics 35*(3), 147–181.

Liu, M. and S. Zhang (2023). The Persistent Effect of Narratives: Evidence from an Online Experiment. *Mimeo*.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences 10*(10), 464–470.

Lombrozo, T. (2012). Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, 260–276.

Mailath, G. J. and L. Samuelson (2020). Learning under Diverse World Views: Model-Based Inference. *American Economic Review 110*(5), 1464–1501.

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

McKelvey, R. D. and T. R. Palfrey (1995). Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior 10*(1), 6–38.

Milgrom, P. R. (1981). Good News and Bad News: Representation Theorems and Applications. *The Bell Journal of Economics 12*(2), 380–391.

Montiel Olea, J. L., P. Ortoleva, M. M. Pai, and A. Prat (2022). Competing Models. *The Quarterly Journal of Economics 137*(4), 2419–2457.

Morag, D. and G. Loewenstein (2023). Narratives and valuations. *CESifo Working Paper No. 10714*.

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

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

Schwartzstein, J. and A. Sunderam (2021). Using Models to Persuade. *American Economic Review 111*(1), 276–323.

Schwartzstein, J. and A. Sunderam (2024). Sharing Models to Interpret Data. *Mimeo*.

Shiller, R. J. (2019). *Narrative economics*. Princeton: Princeton University Press.

Sobel, J. (2013). Giving and Receiving Advice. In D. Acemoglu, M. Arellano, and E. Dekel (Eds.), *Advances in Economics and Econometrics*, pp. 305–341. Cambridge: Cambridge University Press.

Spiegler, R. (2016). Bayesian Networks and Boundedly Rational Expectations. *Quarterly Journal of Economics 131*(3), 1243–1290.

Spiegler, R. (2020a). Behavioral Implications of Causal Misperceptions. *Annual Review of Economics 12*(1), 81–106.

Spiegler, R. (2020b). Can agents with causal misperceptions be systematically fooled? *Journal of the European Economic Association 18*(2), 583–617.

Wang, J. T.-y., M. Spezio, and C. F. Camerer (2010). Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *American economic review 100*(3), 984–1007.

ONLINE APPENDICES

A Extensions and Robustness Exercises

This section contains a discussion of three additional exercises that we conducted that help to shed light on the robustness of the main results and to reveal further insight into the underlying mechanisms of narrative persuasion.

A.1 Sensitivity of Narrative Persuasion to Context

The discussion of our ASYMMETRIC and SYMMETRIC treatments in the main text demonstrates that narratives provide an effective tool for persuasion. This raises the question of whether the narrative-based persuasion we observe is sensitive to the characteristics of the choice environment. A closely related—but conceptually different—question is whether we can protect investors from this type of persuasion by intervening to alter specific elements of the environment. To explore these issues, we conducted three additional treatments. In each of these treatments, we intervened on a specific feature of the choice environment to evaluate whether this helped to protect investors.

Overview of the treatment variation introduced in our "intervention" treatments: The three treatments all follow a similar structure to our ASYMMETRIC treatment, but each introduces one specific change to the setting.

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. As discussed by S&S, the simple fit-based narrative adoption rule can easily be adapted to allow for the investor to be *skeptical*. For example, he might penalize the fit of 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 (3) 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. One natural reason why an individual's skepticism parameter, s, might increase is because they learn that the person they are receiving a narrative from has incentives that are entirely different from their own, indicating a conflict of interest. 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 their decision screen.

INVESTORPRIOR. In the narrative persuasion framework, the advisor can only convince the investor to adopt a new narrative when it fits better than the investor's own prior interpretation of the data (i.e., his default narrative). When the fit of the investor's default narrative improves, the set of potential better-fitting narratives available to the advisor shrinks, and persuasion becomes

more difficult.[34] To encourage investors to improve the fit of their default narrative, we implement the InvestorPrior treatment in which we ask investors to form their own assessment of the process that they think generated the observed history *before* being exposed to the advisor's narrative. Specifically, instead of receiving the historical data and the advisor's message simultaneously and only then forming a belief about the data-generating process, as in Asymmetric, in InvestorPrior 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.[35]

Both our Disclosure and InvestorPrior treatments only adjust the decision environment of investors; not advisors. Therefore, in both treatments, advisors receive identical instructions to the advisors in Asymmetric. We implemented this design choice in order to ensure that we can attribute any potential treatment effect to changes in investor behavior due to changes in their decision environment.

PrivateData. According to the theory, the advisor tries to send a narrative which fits the historical data well. The investor can be persuaded by such a narrative because he disregards the fact that the advisor constructed the narrative ex-post, after observing the data. If access to the data is restricted such that only the investor may access it, this makes it more difficult for the advisor 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, implying that persuasion should be more challenging. To investigate whether having access to private data serves a protective role against persuasion, we introduce the PrivateData treatment in which the advisor does not observe the historical performance data (the investor knows this).[36] 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. However, she is unable to precisely tailor the message to the data that the investor observes (she can only tailor it to her expectation of the data). This may make it more difficult for the advisor to send a message that is both deceptive and persuasive.

Procedures: We recruited 180 participants (90 advisors and 90 investors) per treatment via

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

[35] 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.

[36] 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.

the Prolific platform in March 2022. Participants received a participation fee of £3.50 and could receive an additional bonus payment of £3.75.

Main Results: To evaluate whether narrative persuasion is sensitive to the contextual changes considered in each of the three "intervention" treatments, we ask whether investors form beliefs that are closer to the truth in these scenarios when compared to Asymmetric. Table A.1 presents the findings from these exercises. 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 specific intervention being considered. The regressions only include rounds in which investors are matched with advisors with misaligned advisors, since these are the rounds where advisors may try to persuade investors to move their beliefs away from the truth. The coefficient associated with "Intervention=1" in each of the (*a) columns shows the average effect of the intervention denoted in the column header. Surprisingly, we see that none of the three interventions has a statistically significant protective effect for the average investor.

Table A.1: 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)
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 Asymmetric treatment as well as the relevant treatment mentioned in the column header, (iii) The value of the constant is the same in all regressions as it is the mean of the dependent variable for the Asymmetric treatment and equals 15.3 (iv) 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), (v) Standard errors are clustered at the Interaction Group level, reported in parentheses, (vi) there are 90 clusters (v) The results in columns (*a) relate to Hypotheses 2, 3 and 4 from the pre-registration, (vii) ** $p < 0.05$, *** $p < 0.01$.

Figure B.4 in Appendix Section B.4 provides visual evidence regarding the distributions of investor behavior, plotting the distribution of the distance between investors' beliefs and the truth in each of these treatments. The figure shows that in the InvestorPrior and PrivateData, investor behavior is very similar to that in Asymmetric. However, investor behavior is substantially different 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.

This raises the following question: Why does increased skepticism not protect investors in Disclosure? One potential explanation is the following. Out of all narratives sent by misaligned advisors, approximately 30% are actually truthful. However, investors only know when advisors are *incentivized* to be truthful or not truthful and not whether they *are* truthful or not truthful.

They cannot easily distinguish advisors who are being honest from those who are being dishonest. In DISCLOSURE, investors become skeptical of narratives received from misaligned advisors, but this includes both honest and dishonest misaligned advisors. Therefore, this increased skepticism in DISCLOSURE could lead investors to do better when matched with dishonest advisors, but worse when matched with an honest advisor. Column (1b) provides some support for this explanation by showing that investors do indeed do better 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 do worse when matched with an honest advisor (although, this variable is not statistically significant).

A.2 Exploring the Influence of Explanations

In our experiment, the advisor and investor are only incentivized to care about θ_{post}. The other components of the narrative, c and θ_{pre}, serve only to substantiate the claim made by the advisor about θ_{post} when comparing the narrative to the empirical data. While our experiment is fairly abstract, it aims to capture real-world scenarios in which one individual makes some claim and then provides a explanation for their claim. The aim of the explanation is to make the main claim more convincing by providing the receiver with a more nuanced and detailed set of connected claims ("narrative") that they can evaluate relative to their own information set. If the explanation is coherent with the receivers' existing information set, then he might find the claim more convincing than if the claim was made without an explanation attached. Conversely, if the explanation contradicts the receivers information set, then he might find the claim less convincing. We explore these ideas in our EXPLANATION and NOEXPLANATION treatments.

Overview of the treatment variation introduced in our "explanation" treatments: These two treatments build on our INVESTORPRIOR treatment. Since we are interested in studying the influence of providing the investor with an explanation while holding advisor behavior constant, we "borrow" the narratives sent by advisors in the INVESTORPRIOR treatment (along with the corresponding historical company data). The investors in the EXPLANATION and NOEXPLANATION treatments therefore complete almost exactly the same task as those in INVESTORPRIOR, with one key exception: In the EXPLANATION treatment, investors observe all three components of the advisors' narrative, $(c, \theta_{pre}, \theta_{post})$, while in NOEXPLANATION, investors only observe the parameter of interest, θ_{post}.[37] To isolate the effect of revealing or not revealing these two explanation parameters to investors, we hold all other features of the choice environment constant. In particular, we hold constant the historical data and the full narrative sent by the advisor.

Procedures: We recruited 180 participants per treatment via the Prolific platform in June 2023. Participants received a participation fee of £3.50 and could receive an additional bonus payment of £3.75.

[37] In the interest of expositional brevity, we have omitted describing several important design details that allow us to isolate the causal effect of explanations, while also avoiding the use of deception. For example, each pair of investors in the EXPLANATION and NOEXPLANATION treatments is matched to an investor from the INVESTORPRIOR treatment. This pair of investors follows exactly the same trajectory through the game, observing the same historical data and receiving the same advice. For a complete description of the experimental design details for these two treatments, please refer to the preregistration document. Please note that in the pre-registration document, we refer to the EXPLANATION treatment as 3PARAMETERS and refer to NOEXPLANATION as 1PARAMETER.

Table A.2: The influence of good and bad explanations on persuasion

	(1) Posterior Distance	(2) Posterior Distance	(3) Posterior Distance
Prior Distance	0.366***	0.365***	
	(0.0271)	(0.0271)	
Explanation	0.0383	3.078*	2.649*
	(0.578)	(1.574)	(1.487)
Explanation × fit (APS)		-3.855**	-3.459**
		(1.811)	(1.729)
Round× linked investor FE	Yes	Yes	Yes
$\lvert\theta^{I,0}_{post} - \theta^{A}_{post}\rvert$ FE	No	No	Yes
Observations	3600	3600	3595

Notes: (i) The dependent variable, "posterior distance", is the distance between the investor's assessment and the advisor's message about θ_{post}, $D^{I,1}(\theta^{A}_{post}) := |\theta^{I,1}_{post} - \theta^{A}_{post}|$, (ii) The regressor, "prior distance", denotes the same distance metric *before* the investor meets the advisor, θ_{post}, $D^{I,0}(\theta^{A}_{post}) := |\theta^{I,0}_{post} - \theta^{A}_{post}|$ (iii) The fit metric (APS) provides a measure of fit of the explanation (i.e., it only measures the fit of the auxiliary justification parameters). It does this by constructing a score which, for a given θ_{post}, ranks all possible narratives from 1 (best likelihood fit) to 707 (worst likelihood fit), normalized between 0 (lowest-ranking narrative) and 1 (highest-ranking narrative), (iv) The sample contains data from all investors in Explanation and NoExplanation, (v) For each investor we have 10 observations—one for each round, (vi) Standard errors are clustered at the investor level and are reported in parentheses; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

Main Results: Table A.2 reports our main results from the "explanation" treatments. The regressions address two questions: *(i)* Does providing an explanation result in the investor forming a belief that is closer to the advisor's claim? *(ii)* Does the quality of the explanation matter? In column (1), we regress the posterior distance on the prior distance and a treatment indicator.[38] Each of the two distance metrics reflects the absolute distance between the investor's assessment, θ^{I}_{post}, and the θ^{A}_{post} sent by the advisor. The prior distance uses the investor's prior belief before meeting the advisor, $\theta^{I,0}_{post}$, while the posterior uses his assessment after meeting the advisor, θ^{I}_{post}. Importantly, we can include fixed effects, such that we are essentially comparing pairs of investors that have identical information sets in each round, where one is in the Explanation treatment and observes all three narrative components, and the other is in the NoExplanation treatment and only observes θ^{A}_{post}.

As one would expect, we see that there is a positive correlation between the prior and posterior distance metrics—investors who start off further from the advisor's assessment also end up further away. Interestingly, the coefficient on the treatment indicator is not statistically different from zero, indicating that, on average, allowing for explanations does not make advisors more persuasive in this setting. However, importantly, explanations can either be good or bad in the sense that they can either fit the objective data well or poorly. Adding a good explanation might make a claim more convincing, but adding a bad explanation may make the claim less convincing. To investigate this, in column (2), we interact our treatment indicator with a measure of the quality or fit of the explanation, which we refer to as the auxiliary parameter score (APS).[39] We find that the

[38]All results remain qualitatively unchanged if we do not control for the prior distance.

[39]To construct the APS, for any given θ_{post}, we consider all possible combinations of θ_{pre} and c. Therefore, after choosing a θ^{A}_{post}, the advisor has $101 \times 7 = 707$ possible choices for the explanation, θ_{pre} and c (since we impose that θ_{pre} can take 101 discrete values). We rank these from best (rank 1) to worst (rank 707) to give each explanation an auxiliary parameter rank (APR) value. We then normalize this to construct a score (the APS) between 0 and 1, such that the best explanation takes an APS value of 1 and the worst an APS value of 0:

$$APS = 1 - \left(\frac{APR-1}{707-1}\right) = \frac{707-APR}{707-1}$$

quality of the explanation matters—claims supported by a good explanation are more persuasive than claims supported by a bad explanation. The negative coefficient on the interaction term shows that the better the fit of the explanation, the closer the investor's posterior belief is to the advisor's assessment of θ_{post}. However, the positive coefficient on the treatment indicator variable, EXPLANATION, implies that when the explanation fits poorly, investors are even less persuaded than they are in the absence of an explanation in NOEXPLANATION. To check for the robustness of this result, in column (3), instead of including a continuous measure of the prior distance in the regression, we add prior distance fixed effects as more flexible controls. The results are very similar. Taken together, the evidence suggests that the influence of explanations for persuasiveness hinges on the quality of the explanation.

A.3 Accounting for Decision Noise

The evidence presented so far shows that many qualitative predictions of the S&S framework are borne out in the data. One of the assumptions made within the framework is that decision-makers (advisors and investors) are precise in their choices, acting without noise. As a final exercise, this section sets out to relax this assumption and allow for noisy decision-making in the analysis. In doing so, we empirically quantify the amount of decision noise, which enables us to estimate the degree to which noise explains our data. This exercise also sheds light on the degree to which our data are explained by the mechanics of the S&S model once we allow for noisy decisions.

We estimate a discrete choice model of narrative adoption and construction using data from COMPETITION. This data is well suited for such an exercise: In COMPETITION, investors make a binary choice between the human advisor's narrative and a competing narrative that we assigned exogenously (the robot advisor's narrative). This allows us to quantify how the empirical fit of the advisor's narrative relative to the competing narrative influences narrative adoption. Since advisors observe the competing narrative before constructing their own, the treatment also allows us to analyze how specific features of the competing narrative influence narrative construction.

Our empirical model is based on the framework in Section 3 but adds the assumption that individuals may make mistakes; i.e., it quantifies the amount of decision noise. The estimation proceeds in two steps. First, we estimate the investor's binary choice between the two narratives that he receives, modeling this choice as a function of the relative fit. This provides us with an estimate of the investor's decision noise by quantifying the frequency with which he adopts a worse-fitting narrative. We then turn to the advisor and estimate a discrete choice model of narrative construction, modeling the choice of narrative as a function of the relative fit of her narrative, the noise in the investor's narrative adoption rule, and her own decision noise.[40]

The presence of noise in the investor's narrative adoption rule implies that the advisor should not only condition her choice of narrative on the competing narrative's fit, but also on its θ_{post}—value. For example, as the competing narrative's θ_{post} increases, the fit of the optimal narrative of an up-advisor decreases while its bias increases. Intuitively, this is because an increase in the θ_{post} of the competing narrative means that the *worst-case scenario* for the up-advisor, where the competing narrative is adopted, becomes less bad. This allows her to become more ambitious.

[40]The idea of introducing decision noise is a common modeling approach in economics. Notably, it is a key component of the Quantal Response Equilibrium (QRE) solution concept in behavioral game theory (e.g., McKelvey and Palfrey, 1995). It is worth noting, however, that the exercise we conduct does not formally estimate a QRE since the framework we quantify is not an equilibrium framework.

We illustrate the implications of this noise-based narrative construction channel in an example at the end of this section.

Two-stage estimation: The investor chooses between the narrative of the human advisor, $\boldsymbol{m}^A$, and the competing $\boldsymbol{m}^R$ of the robot. We assume that he chooses $\boldsymbol{m}^A$ if $\text{EPI}(\boldsymbol{m}^A|h)+\varepsilon^A \geq \text{EPI}(\boldsymbol{m}^R|h)+\varepsilon^R$, where ε^A and ε^R are iid type-I extreme-value distributed noise parameters with location 0 and scale $1/\lambda^I$. This narrative adoption rule is similar to the one used in Froeb et al. (2016)'s model of adversarial justice. It essentially becomes equal to the S&S adoption rule as $\lambda \to \infty$. With this parametric specification, the probability of the investor adopting belief θ^A_{post} is equal to:

$$\Pr(\text{adopt } \boldsymbol{m}^A|\boldsymbol{m}^A, \boldsymbol{m}^R, \lambda^I) = \frac{\exp(\lambda^I \cdot \text{EPI}(\boldsymbol{m}^A))}{\exp(\lambda^I \cdot \text{EPI}(\boldsymbol{m}^A)) + \exp(\lambda^I \cdot EPI(\boldsymbol{m}^R))}.$$

The advisor anticipates the investor's adoption rule and chooses to send a narrative that maximizes:

$$\mathbb{E}[U^\varphi(\boldsymbol{m}^A)] = \Pr(\text{adopt } \boldsymbol{m}^A|\boldsymbol{m}^A, \boldsymbol{m}^R, \lambda^I)U^\varphi(\theta^A_{post}) + (1 - \Pr(\text{adopt } \boldsymbol{m}^A|\boldsymbol{m}^A, \boldsymbol{m}^R, \lambda^I))U^\varphi(\theta^R_{post}) + \eta,$$

where η is iid type-I extreme-value distributed with location 0 and scale $1/\lambda^A$. Allowing for noise in the advisor's choice, this means that the probability that an advisor sends $\boldsymbol{m}^A$ is:

$$\Pr(\text{send } \boldsymbol{m}^A|\boldsymbol{m}^R, \lambda^A, \lambda^I) = \frac{\exp(\lambda^A \cdot \mathbb{E}[U^\varphi(\boldsymbol{m}^A)])}{\sum_{m \in \mathcal{M}} \exp(\lambda^A \cdot \mathbb{E}[U^\varphi(\boldsymbol{m})])}.$$

Before turning to the results, it is worthwhile highlighting two caveats regarding the way in which we estimate the discrete choice model described above. First, the advisor's message space in our experiment is large—advisors have $101 \times 101 \times 7$ possible messages to choose from. We simplify the estimation problem by only considering the advisor's problem of choosing θ^A_{post} and year of change parameter c, which reduces the number of possible messages to 101×7. When calculating the empirical fit of a message, we assume that the advisor chooses the data-optimal θ^A_{pre} for any given c^A.[41] Second, we will estimate the discrete choice model of narrative construction using misaligned senders only because their persuasion target remains the same across all possible decision situations.

We use maximum likelihood to first estimate λ^I and then estimate λ^A, given the estimate of λ^I.[42] The two parameters quantify the decision noise of the investor and advisor, respectively. As $\lambda^I \to \infty$, the investor always adopts the better-fitting narrative. Similarly, as $\lambda^A \to \infty$, the advisor always sends the message that maximizes her expected payoff. Values of zero would instead imply that they choose randomly.

Column (1) in Table A.3 presents results from the two-stage estimation. We obtain estimates for λ that are positive and significant, implying that individual decisions are not random. Their absolute size, however, suggests that decisions are also partially determined by noise. The estimate of λ^I, for example, implies that increasing the EPI of the advisor's message by 0.1 increases the probability of the investor adopting it by 3.2 percentage points. Importantly, this influence of the

[41] We present parameter estimates using the whole message space in Appendix B.6. They are very similar to the estimates of the model with the smaller message space presented in the main text below. However, the computational demands of the full message space model make it difficult to comprehensively test the robustness of its parameter estimates in Monte Carlo experiments.

[42] We used Monte Carlo experiments to confirm that this procedure reliably identifies the true underlying parameter values. See Appendix B.6 for details.

EPI on adoption is continuous.[43] Column (2) presents results from using only the advisor data to estimate both noise parameters. Essentially, this involves only estimating Step 2—the estimation of the advisor's strategy. The estimate of λ^I here can be interpreted as the advisor's expectation of how noisy the investor's assessment rule is. Our estimates reveal that advisors' expectations of λ^I are positive and significant. Importantly, the estimated λ^I in Column (2) is not significantly different from the estimate in Column (1). Therefore, we cannot reject the hypothesis that advisors accurately anticipate the amount of noise in investors' assessments.[44] Finally, Column (3) uses only advisor data to derive an estimate of λ^A under the assumption that investors' adoption decisions contain very little noise. Essentially, we assume that advisors anticipate investors who are sufficiently precise and always choose the model with the better empirical fit. Here, a key insight is that this model achieves a worse likelihood fit when compared to the models estimated in Columns (1) and (2): likelihood ratio tests reject the hypothesis that the noise neglect model fits the data as well as either of the alternative models ($p < 0.001$). To summarize, the parameter estimates from this exercise indicate that the S&S framework is able to explain patterns that we observe in the data. Therefore, they are consistent with our reduced form results. Additionally, these results suggest that decision-making is somewhat noisy and, importantly, that advisors anticipate the noise in investors' decisions.

Table A.3: Estimated noise parameters

	(1) Accurate anticipation	(2) Subjective response	(3) Noise neglect
$\hat{\lambda}^A$	3.181***	3.285***	2.24***
	(0.364)	(0.388)	(0.339)
$\hat{\lambda}^I$	1.39***	2.002**	100
	(0.206)	(0.808)	–
Log likelihood	-3892.14	-3891.802	-3912.06
Observations: Investors	900	–	–
Observations: Advisors	600	600	600

Notes: Column (1) presents estimation results from a two-stage estimation procedure that first estimates λ^I using investor adoption decisions and then plugs the estimated λ^I into the advisor's discrete choice problem to estimate λ^A. Column (2) uses only advisor data on narrative construction to derive estimates of both noise parameters. Column (3) uses only advisor data to derive an estimate of the advisor noise parameter under the assumption that the investor's adoption decisions do not contain much noise. This is achieved by imposing a low value for the investor's scale parameter of $\frac{1}{\lambda^I} = \frac{1}{100}$. The estimates use data from COMPETITION and exclude aligned advisors in the estimation of λ^A. The log-likelihood row displays the log-likelihood value of the advisor's discrete choice problem. Standard errors in parenthesis. * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

Implications of noise: To illustrate the relevance of a noisy adoption rule, we consider an example. Figure A.1 consists of three panels that each display:

(a) a realization of historical company data (which is constant across all panels),

(b) a competing narrative (in green),

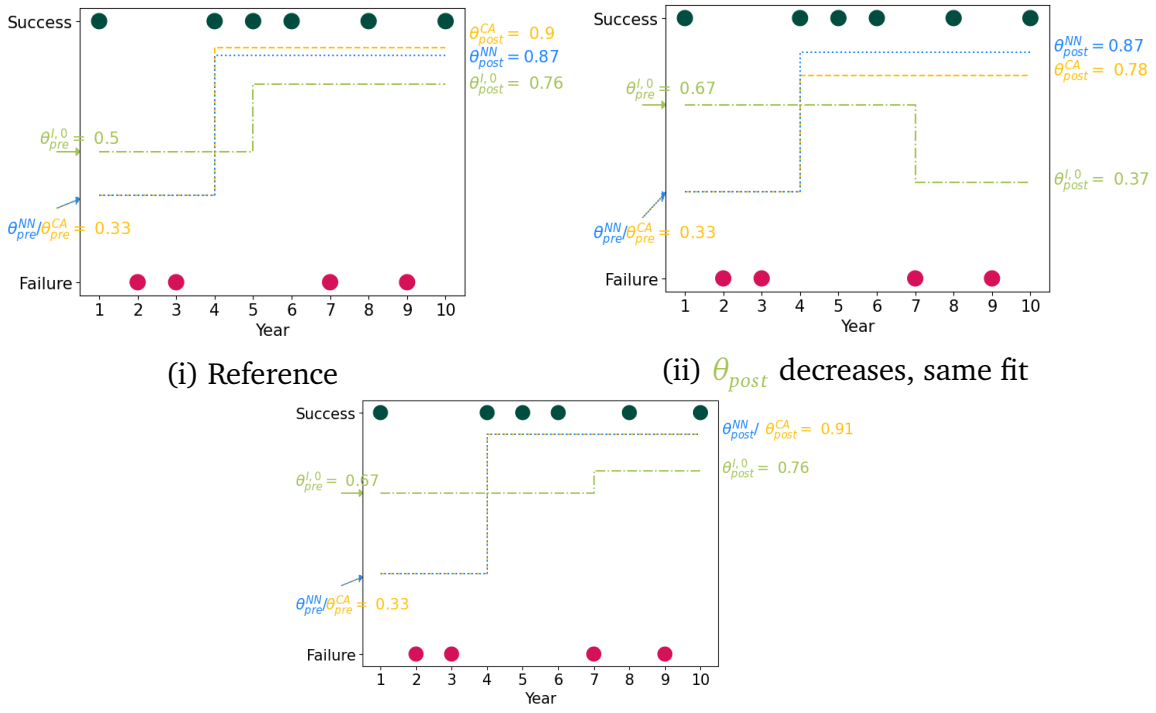
[43] A model without decision noise would instead predict a discontinuous jump in the adoption probability from 0 to 1 at the point where the advisor's EPI surpasses that of the competing narrative. In the logit probability function, this will be the case if λ^I becomes large. In practice, setting $\lambda^I = 100$ (as we do in Column (3) of Table A.3) is enough to generate a discontinuous jump.

[44] A likelihood ratio test also does not reject the null hypothesis that the subjective response and accurate anticipation models fit the data equally well ($p = 0.411$).

(c) the optimal narrative of two types of up-advisors—one who correctly anticipates noise in the investor's assessment rule (m^{CA}, in yellow), and one who neglects noise (m^{NN}, in blue).

When moving from Panel *(i)* to Panel *(ii)* in the figure, the empirical fit of the default model is held constant. Consequently, m^{NN} is predicted to remain the same. The rationale for this is the following. Since the advisor anticipates no noise in the investor's decision, she expects that the investor will adopt her narrative with certainty provided she proposes a narrative with a higher fit than the competing narrative. She, therefore, disregards all other features of the competing narrative aside from the fit. Since the fit is constant between panels *(i)* and (ii), she sends the same m^{NN}.

Figure A.1: Optimal narratives of an up-advisor in response to a competing narrative (by noise in the decision rule and competing narrative)



(i) Reference

(ii) θ_{post} decreases, same fit

(iii) Same θ_{post}, fit decreases

Notes: Each panel shows an example data set (which is constant across all panels), and a default competing narrative which is depicted in green. The yellow line displays the optimal narrative of an up-advisor given the data and competing narrative if the investor's decision noise parameter is equal to $\lambda^I = 1.39$. The blue line displays the optimal narrative of an up-advisor who neglects decision noise in the investor's narrative adoption and assumes that the investor will adopt the advisor's narrative if and only if it has a higher EPI than the default.

In contrast, with anticipated noise, the fit of m^{CA} increases, and its bias decreases when moving from Panel *(i)* to (ii). This happens because, when expecting decision noise, the advisor cannot be sure that the investor will adopt her narrative. As we move from Panel *(i)* to (ii), the consequences of the investor adopting the competing narrative become worse for the advisor. Therefore, she increases her own narrative's fit (by lowering her movement ambition) to ensure a higher adoption probability.

When moving from Panel *(i)* to (iii), the competing narrative's θ_{post} remains constant while the fit decreases. As a consequence, the bias of the advisor's optimal model under both variants increases while the fit decreases, illustrating the fit-movement tradeoff outlined by S&S. The results from this example highlight that, with noisy narrative adoption, the two θ_{post}—values (of the competing narrative and the advisor) are complements.

B Additional Results

B.1 Participant Demographics

Table B.1 below presents summary statistics of the demographics of the participants in the experiment. In general, the demographics are fairly balanced between treatments, with the notable exception being the participants' average age, which was higher in our second wave of data collection in 2023 than in our first wave in 2022. However, it is important to keep in mind that the tests in our main analysis rely on within-treatment variation. Therefore, we do not see this as a threat to our main results, which only test for differences in participant behavior within SYMMETRIC, COMPETITION, and the two EXPLANATION treatments, and refrain from directly comparing their behavior to that of participants the Wave 1 experiments.

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

	ASYMMETRIC mean/sd	SYMMETRIC/COMPETITION mean/sd	DISCLOSURE mean/sd	INVESTORPRIOR mean/sd	PRIVATEDATA mean/sd	EXPLANATION mean/sd	NOEXPLANATION mean/sd
Investors:							
Age	36.044	41.783	34.389	36.278	35.800	40.822	40.606
	(12.674)	(14.652)	(11.624)	(12.469)	(12.190)	(13.726)	(12.998)
Gender: Female	0.506	0.500	0.467	0.444	0.556	0.489	0.506
	(0.501)	(0.501)	(0.502)	(0.500)	(0.500)	(0.501)	(0.501)
Gender: Male	0.489	0.489	0.522	0.544	0.411	0.500	0.494
	(0.501)	(0.501)	(0.502)	(0.501)	(0.495)	(0.501)	(0.501)
Gender: Other	0.006	0.011	0.011	0.011	0.033	0.011	0.000
	(0.075)	(0.105)	(0.105)	(0.105)	(0.181)	(0.105)	(0.000)
Edu: Primary school	0.000	0.000	0.000	0.000	0.011	0.006	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.105)	(0.075)	(0.000)
Edu: Secondary school	0.078	0.117	0.089	0.067	0.111	0.122	0.083
	(0.269)	(0.322)	(0.286)	(0.251)	(0.316)	(0.328)	(0.277)
Edu: Higher secondary education	0.183	0.144	0.244	0.244	0.267	0.250	0.206
	(0.388)	(0.353)	(0.432)	(0.432)	(0.445)	(0.434)	(0.405)
Edu: College or university	0.478	0.528	0.467	0.467	0.411	0.389	0.489
	(0.501)	(0.501)	(0.502)	(0.502)	(0.495)	(0.489)	(0.501)
Edu: Post-graduate	0.244	0.211	0.189	0.211	0.189	0.233	0.222
	(0.431)	(0.409)	(0.394)	(0.410)	(0.394)	(0.424)	(0.417)
Edu: Prefer not to say	0.017	0.000	0.011	0.011	0.011	0.000	0.000
	(0.128)	(0.000)	(0.105)	(0.105)	(0.105)	(0.000)	(0.000)
Observations	180	180	90	90	90	180	180
Data Collection Wave	1	2	1	1	1	2	2
Advisors:							
Age	35.878	41.667	35.500	34.989	34.967		
	(12.030)	(13.398)	(11.619)	(12.257)	(12.471)		
Gender: Female	0.506	0.494	0.500	0.511	0.556		
	(0.501)	(0.501)	(0.503)	(0.503)	(0.500)		
Gender: Male	0.483	0.494	0.478	0.489	0.433		
	(0.501)	(0.501)	(0.502)	(0.503)	(0.498)		
Gender: Other	0.011	0.011	0.022	0.000	0.011		
	(0.105)	(0.105)	(0.148)	(0.000)	(0.105)		
Edu: Primary school	0.017	0.000	0.000	0.000	0.011		
	(0.128)	(0.000)	(0.000)	(0.000)	(0.105)		
Edu: Secondary school	0.056	0.094	0.089	0.078	0.111		
	(0.230)	(0.293)	(0.286)	(0.269)	(0.316)		
Edu: Higher secondary education	0.211	0.200	0.244	0.200	0.289		
	(0.409)	(0.401)	(0.432)	(0.402)	(0.456)		
Edu: College or university	0.467	0.439	0.389	0.489	0.389		
	(0.500)	(0.498)	(0.490)	(0.503)	(0.490)		
Edu: Post-graduate	0.250	0.261	0.267	0.233	0.189		
	(0.434)	(0.440)	(0.445)	(0.425)	(0.394)		
Edu: Prefer not to say	0.000	0.006	0.011	0.000	0.011		
	(0.000)	(0.075)	(0.105)	(0.000)	(0.105)		
Observations	180	180	90	90	90		
Data Collection Wave	1	2	1	1	1		

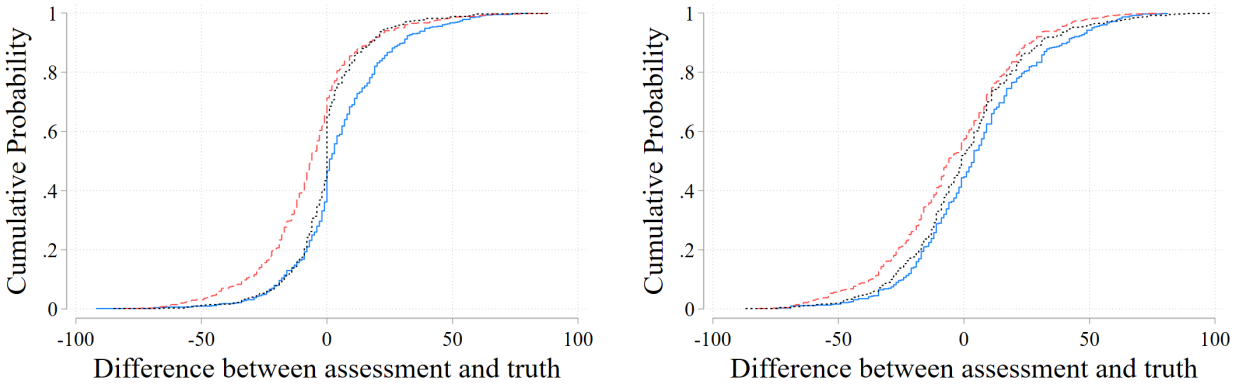
Notes: (i) Aside from the "Age" variable, each of the "Gender" and "Education" variables reports the fraction of the sample that falls into the relevant category, (ii) The reason why there are no observations for advisors in the EXPLANATION and NOEXPLANATION treatments is due to the way we designed these two treatments. Here, we reused messages sent by advisors in the INVESTORPRIOR treatment and only recruited new investors, (iv) The "Data Collection Wave" refers to the fact that we collected the data in two waves—the treatments with a "1" were collected in 2022, while those with a "2" were collected in 2023. (v) Standard deviations in parenthesis.

The table contains no information about advisors in the two EXPLANATION treatments. The reason for this is that we re-used the advisor messages from INVESTORPRIOR for the two EXPLANATION treatments. For more details, please refer to the description of our experimental design.

B.2 Investor Behavior

Figure B.1 reports the empirical cumulative density functions (CDFs) of the difference between investors' beliefs, θ^I_{post}, and the true value, θ^T_{post}. The black, blue, and red lines plot the CDFs for cases where the investor meets an advisor who is of the aligned-, up-, and down-type, respectively. The left panel uses data from the ASYMMETRIC treatment, while the right panel uses data from the SYMMETRIC treatment. The figure shows that, in both treatments, investors' beliefs are shifted to the right across the full distribution when comparing those matched with an up- and down-advisor. This indicates that the advisors are shifting investors' beliefs towards the advisor's persuasion target.

Figure B.1: CDF of distance between investors' beliefs and the truth (by advisor type)



(i) ASYMMETRIC (ii) SYMMETRIC

Notes: The figure reports the empirical CDF of the difference between the investor's assessment, θ^I_{post}, and the truth θ^T_{post} using data from ASYMMETRIC (left panel) and SYMMETRIC (right panel). The black, blue, and red lines plot the CDFs for cases where the investor meets an advisor who is of the aligned-, up-, and down-type, respectively. The regression results in Table B.2 show that the difference between the investors' assessments and the truth is significantly different between advisor types in both treatments. It is important to notice that the regressions control for round fixed effects. Since these fixed effects control for the true value, θ^T_{post}, which is constant within a given round, using θ^I_{post} as the outcome gives the same coefficients estimates as using the difference between assessment and truth.

To provide support for this visual evidence, Table B.2 reports regression results which examine the influence that being matched with different types of advisors has on investors' beliefs. To do this, we regress investors' beliefs, θ^I_{post}, on indicator variables for the up- and down-advisor, implying that the aligned advisor is the benchmark category. Column (1) reports the results for the ASYMMETRIC treatment, controlling for round fixed effects. We see that investors matched with an up-advisor report beliefs that are 5.1 pp higher than those matched with an aligned advisor, while those matched with a down-advisor report beliefs that are 6.4 pp lower. Column (2) reports the results for the SYMMETRIC treatment. Here, our experimental design allows us to control for Round×History fixed effects, thereby holding the historical company data constant in our statistical comparisons. We again see that investors matched with up-advisors report beliefs that are higher (3.9 pp) than those matched with aligned advisors, while those matched with down-advisors report beliefs that are lower (5.6 pp). All of these differences are statistically significant at the 1% level.

Table B.2: Investor assessment by advisor type and treatment

	(1) θ^I_{post}	(2) θ^I_{post}
Up-advisor	5.105***	3.993***
	(1.057)	(1.463)
Down-advisor	-6.430***	-5.612***
	(1.080)	(1.425)
Treatment	ASYMMETRIC	SYMMETRIC
Round FE	Yes	No
Round×History FE	No	Yes
Observations	1800	1800

Notes: (i) The dependent variable is the investor's assessment (ii) The sample used in Column (1) contains data from all investors in ASYMMETRIC, while Column (2) contains data from SYMMETRIC, (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; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

In the main text, we use data from the COMPETITION treatment to show that the fit of the robot advisor's narrative influences both narrative construction and adoption. Here, in Table B.3, we now present evidence from the ASYMMETRIC and SYMMETRIC treatments showing the relationship between the fit of the advisor's narrative and investor behavior in these two treatments. We do this by regressing the distance between the advisor's message and the investor's report, $|\theta^I_{post} - \theta^A_{post}|$, on the EPI of the advisor's narrative. Columns (1) and (2) report the results for ASYMMETRIC and SYMMETRIC, respectively. In both treatments, we find that when the advisor's narrative fits the objective data better, the investor's belief, θ^I_{post}, is closer to the θ^A_{post} of the advisor's message. Specifically, a move from the worst-fitting to the best-fitting narrative is associated with a 15 pp [11 pp] reduction in the distance between the advisor's message and the investor's report in ASYMMETRIC [SYMMETRIC]. This evidence is associative rather than causal, but it supports our results in the main text, suggesting that investors find narratives that fit the data well to be more compelling.

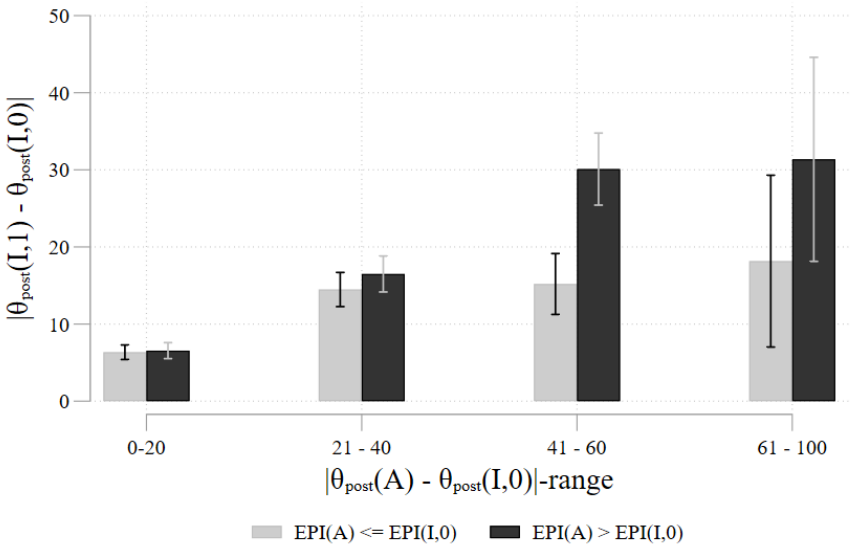
In addition to the exercises examining the influence of narrative "fit" that we conducted using the data from COMPETITION, ASYMMETRIC and SYMMETRIC, we are also able to make use of the fact that in INVESTORPRIOR we elicited investor's prior beliefs before they meet their advisor. This allows us to examine whether investor's belief updating in response to the narrative they receive is influenced by the fit of this narrative. By examining belief updating, we are able to control for heterogeneity in the prior beliefs that investors form themselves, and focus on how the fit of the narrative influences the *change* in their beliefs.

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

	(1) $\lvert\theta^I_{post} - \theta^A_{post}\rvert$	(2) $\lvert\theta^I_{post} - \theta^A_{post}\rvert$
Advisor message fit (EPI)	-14.59***	-11.14***
	(1.892)	(2.044)
Misaligned advisor = 1	0.691	0.0485
	(0.668)	(1.100)
Treatment	Asymmetric	Symmetric
Round FE	Yes	No
Round× History FE	No	Yes
Observations	1800	1800

Notes: (i) The dependent variable is the absolute distance between the investor assessment and the advisor narrative (ii) The sample used in Column (1) contains data from all investors in Asymmetric; that of Column (2) contains data from Symmetric, (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; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

Figure B.2: 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 B.2 provides a visual portrayal of investor belief updating. The y-axis shows the average absolute distance that investors update, $|\theta^I_{post} - \theta^{I,0}_{post}|$, and the x-axis disaggregates the data into categories according to the distance between the advisor's message and the investor's prior belief, $|\theta^A_{post} - \theta^{I,0}_{post}|$. The pairs of bars divide the data according to whether the empirical fit of the advisor's proposed narrative, EPI($\boldsymbol{m}^A$), is better or worse than the fit of the investor's default

model, EPI($\boldsymbol{m}^{I,0}$). 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 exercise less discretion when updating their beliefs in response to messages that are "close" to their prior assessment could be that investors perceive following the advisor's advice to be less risky than when the advice is "far" from their assessment.

Table B.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; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

To provide statistical evidence in support of the patterns in the data displayed in the figure, Table B.4 presents the results from regressions that investigate the influence of narrative fit 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 the investor's 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. Taken together, these results show that the fit of the advisor's narrative plays an important role in influencing investor belief updating.

B.3 Advisor Behavior

Table B.5 shows how advisors bias the narratives that they send as a function of their incentive-type. Columns (1a) and (1b) report results from the Asymmetric treatment. Column (1a) re-

gresses the advisor's θ^A_{post} on indicator variables for the up- and down-advisor, leaving the aligned advisor as the omitted category. The regression controls for round fixed effects, implying that we control for the value of the true θ_{post}. We see that, on average, up-advisors report θ^A_{post}'s that are 12.0 pp higher than those reported by aligned advisors, while down-advisors report θ^A_{post}'s that are 9.8 pp lower than aligned advisors. Column (1b) reports the results for the same regression, with the exception that the outcome variable is now the advisors, θ^A_{pre}. This regression reveals two insights. First, the coefficient signs are reversed relative to column (1a)—up-advisors report lower values of θ^A_{pre} than aligned advisors, while down-advisors report higher values of θ^A_{pre}. Second, the magnitude of the bias in the θ^A_{pre} due to advisor incentives is smaller than for θ^A_{post}.

Table B.5: Regressions on the impact of incentive type on narrative construction

	(1a) θ^A_{post}	(1b) θ^A_{pre}	(2a) θ^A_{post}	(2b) θ^A_{pre}
Up-advisor	12.02***	-2.951***	12.47***	-5.417***
	(1.009)	(0.806)	(1.711)	(1.491)
Down-advisor	-9.832***	2.488***	-6.963***	7.322***
	(0.945)	(0.764)	(1.648)	(1.780)
Treatment	Asymmetric	Asymmetric	Symmetric	Symmetric
Round FE	Yes	Yes	No	No
Round × History FE	No	No	Yes	Yes
N	3600	3600	1800	1800

Notes: (i) The dependent variable is either the θ^A_{post} (odd columns) or the θ^A_{pre} (even columns) sent by the advisor, (ii) Columns (1a) and (1b) contain data from all advisors who received the Asymmetric instructions (advisors in Disclosure and InvestorPrior also received the Asymmetric instructions and are included here, which is why there are 3600 observations); columns (2a) and (2b) contain data from advisors who participated in Symmetric, (iii) We can control for Round×History FE in Symmetric but not in Asymmetric due to adjustments to the experimental design, (iv) Aligned advisors serve as a reference group, (v) For each advisor we have 10 observations—one for each round, (vi) Standard errors are clustered at the advisor level, implying that there are 360 clusters in Asymmetric and 180 clusters in Symmetric, and are reported in parentheses; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

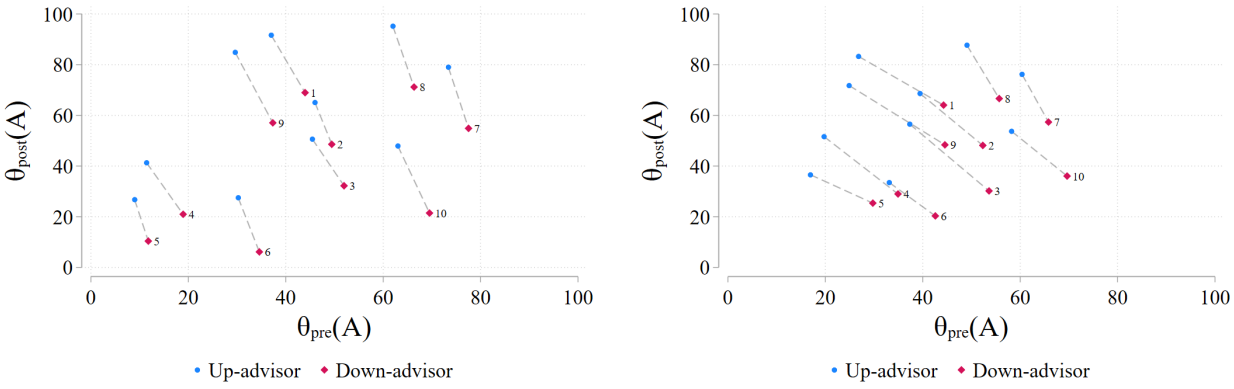
Columns (2a) and (2b) report similar regressions for the Symmetric treatment. The main difference is that we are able to control for Round×History FEs in Symmetric as we've discussed above. The general pattern of results in the Symmetric treatment is similar to that in Asymmetric, with a reversal of the parameter signs between columns (2a) and (2b). One interesting difference is that the magnitudes of the bias in column (2b) are closer to those in (2a), relative to the difference between (1a) and (1b). This suggests that in Symmetric, advisors are distorting θ^A_{pre} nearly as much as they distort θ^A_{post}. However, it is important to stress that we did not design the Asymmetric and Symmetric treatments to be directly compared to one another; rather, the treatments are designed to examine whether similar patterns are observed within each treatment using within-treatment variation (e.g., advisor incentive variation).

In general, the regression results are in line with the narrative construction behavior shown in Figure 5—on average, advisors tend to shift θ^A_{post} towards their persuasion target, and shift θ^A_{pre} in the opposite direction to improve the fit of the narrative.

Figure B.3 shows how up-advisors (blue markers) and down-advisors (red markers) systematically construct different average $(\theta^A_{post}, \theta^A_{pre})$ vectors in each of the ten rounds of the experiment. The y-axis reflects the θ^A_{post} value, while the x-axis denotes the θ^A_{pre} value. The left panel shows

advisor behavior in ASYMMETRIC, while the right reports behavior in SYMMETRIC. Each of the ten rounds is numbered next to the red marker. We see that, in each of the ten rounds, the red marker is to the south-east of the blue marker. This is the case in both treatments. This indicates that in every round, the average θ^A_{post} reported by an up-advisor is higher than that of a down-advisor, and the average θ^A_{pre} of up-advisors is lower than that of down-advisors. This shows visually in a fairly simple way that the systematic patterns in advisor narrative construction that we describe in the main text are not driven by one or two rounds, but are present in each and every round.

Figure B.3: Advisor θ_{post} and θ_{pre} reports in each round (by advisor type)



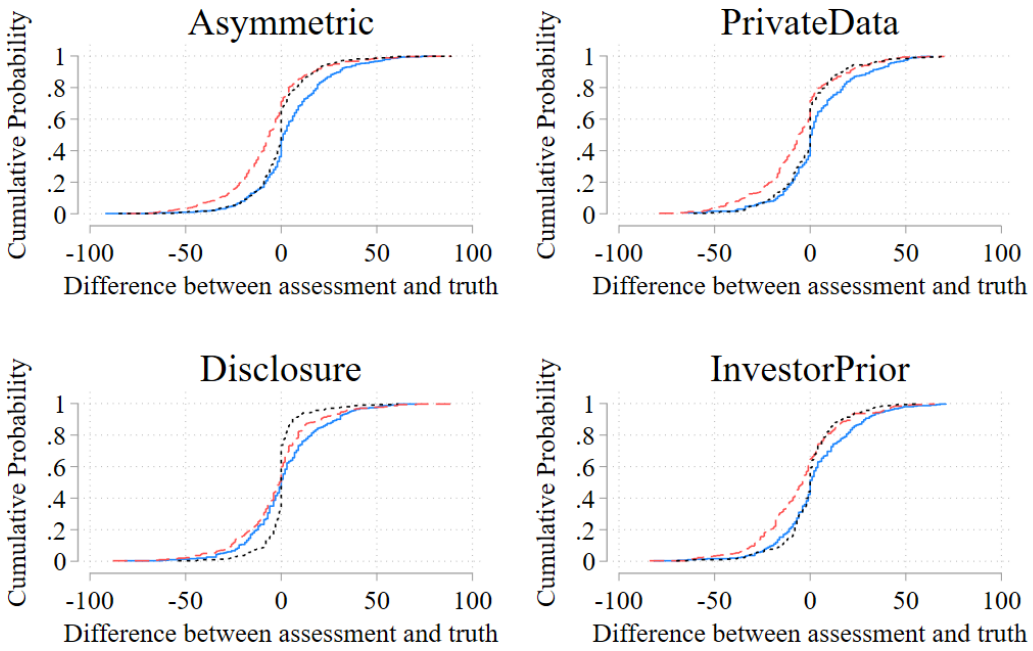
(i) ASYMMETRIC (ii) SYMMETRIC

Notes: (i) The left panel uses data of all advisors who received the ASYMMETRIC instructions while the right panel uses data of all advisors who received the SYMMETRIC instructions, *(ii)* The numbered labels in the figure denote the 10 rounds of the experiment, *(iii)* The blue markers show the average θ_{post} and θ_{pre} sent by up-advisors in each round, while the red markers report the same for down-advisor, *(iv)* 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. It is important to recall that in each round the true θ^T_{post} and θ^T_{pre} are held constant, implying that without a systematic distortion in narrative construction due to advisor incentives the red and blue markers should coincide.

B.4 Sensitivity to Context

Figure B.4 plots the cumulative density functions (CDFs) of the difference between investors' beliefs, θ^I_{post}, and the true value, θ^T_{post}. As in Figure B.1, the black, blue, and red lines plot the CDFs for cases where the investor meets an advisor who is of the aligned-, up-, and down-type, respectively. The top-left panel reports the data from the ASYMMETRIC treatment, with the other three panels containing data from the three "intervention" treatments: PRIVATEDATA (top-right), DISCLOSURE (bottom-left), and INVESTORPRIOR (bottom-right). The figure shows that in the three treatments aside from DISCLOSURE, investors report higher beliefs when matched with an up-advisor in comparison to when matched with a down-advisor. In DISCLOSURE, investors appear to become more skeptical of the advice they receive and we observe less of a gap between the beliefs of investors matched with up- versus down-advisors.

Figure B.4: Difference between θ^I_{post} and θ^T_{post} (by treatment and advisor type).

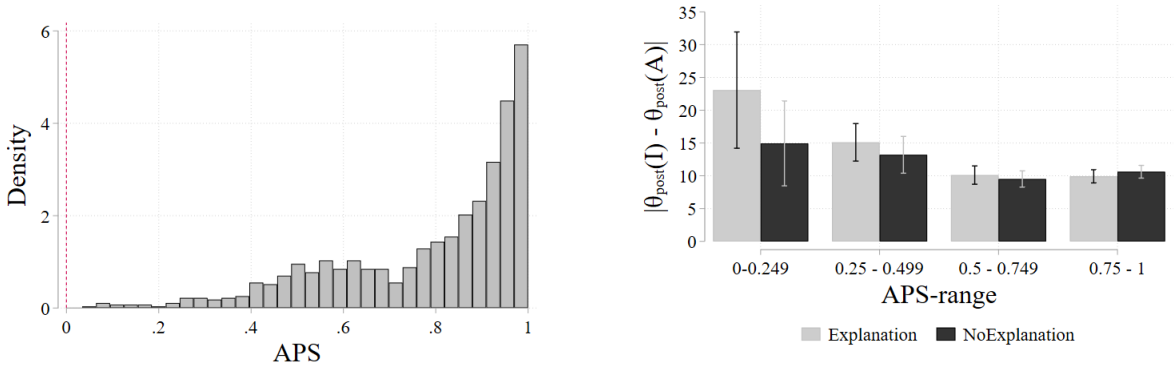


Notes: (i) The figure plots the CDF of the difference between the investor's belief and the truth, $\theta^I_{post} - \theta^T_{post}$, for all investor-rounds where the investor is matched with a particular advisor type, *(ii)* Each of the panels show this for a particular treatment condition, *(iii)* 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.

B.5 Explanations

In this section we provide additional evidence to illustrate the relationship between good and bad explanations and the investor's trust in the advisor's message. First, in Panel *(i)* of Figure B.5, we plot the distribution of the APS in the explanation treatments (keep in mind that messages were perfectly balanced between treatments, which implies that the APS distribution is exactly the same in both treatments). The figure shows that most narratives contain auxiliary parameters that obtain a relatively high APS score; 75% of all narratives sent by advisors include auxiliary parameters that are among the approximately 35% best-fitting pairs of auxiliary parameters that they could have sent. Second, Panel *(ii)* plots the mean distance between the investor's belief and the advisor's message about θ_{post}, $|\theta^I_{post} - \theta^A_{post}|$, as a function of the treatment (Explanation vs NoExplanation) and APS range. The results plotted here display a negative correlation between the distance and the APS in Explanation but not in NoExplanation. This indicates that the quality of the explanation matters and is in line with the regression results reported in Table A.2. They further suggest that, in the data, investors are particularly skeptical of narratives if they contain poorly-chosen auxiliary parameters (i.e., those that are among the least-fitting quarter of possible auxiliary parameters). Narratives that fit poorly are not convincing.

Figure B.5: Distribution of the APS and relation between the APS and the investor's distance to the message (by treatment)



(i) Distribution of the APS

(ii) Distance to the message (by APS and treatment)

Notes: Panel *(i)* presents the distribution of the Auxiliary Parameter Score (APS). Panel *(ii)* displays the average distance to the advisor's message by APS-range and treatment. The error bars display 95% confidence intervals that were derived from regressions which cluster standard errors at the investor level.

B.6 Accounting for Decision Noise

In this section, we provide two additional pieces of evidence to supplement the analysis of Section A.3. First, Table B.6 presents parameter estimates of a discrete choice model that considers the full message space that advisors have (i.e., a message space of size $101 \times 101 \times 7$). Compared to the estimated parameters presented in Table A.3, the point estimates of the full message space model are slightly smaller but they remain significantly different from zero at the same level as the parameter estimates reported in the main text.

Table B.6: Noise parameter estimates using the model with a full message space

	Accurate anticipation	Subjective response	Noise neglect
$\hat{\lambda}^A$	3.31***	3.501***	2.852***
	(0.383)	(0.46)	(0.441)
$\hat{\lambda}^I$	1.39***	2.02**	100
	(0.206)	(0.912)	
Log Likelihood	-6662.074	-6661.794	-6682.661
Observations: Investors	900	–	–
Observations: Advisors	600	600	600

Notes: Column (1) presents estimation results from a two-stage estimation procedure that first estimates λ^I using investor adoption decisions and then plugs the estimated λ^I into the advisor's discrete choice problem to estimate λ^A. Column (2) uses only advisor data on narrative construction to derive estimates of both noise parameters. Column (3) uses only advisor data to derive an estimate of the advisor noise parameter under the assumption that the investor's adoption decisions do not contain much noise. This is achieved by imposing a low value for the investor's scale parameter of $\frac{1}{\lambda^I} = \frac{1}{100}$. The estimates use data from Competition and exclude aligned advisors in the estimation of λ^A. The log-likelihood row displays the log-likelihood value of the advisor's discrete choice problem. Standard errors in parenthesis. * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

Second, Table B.7 presents the results of Monte Carlo experiments to verify the reliability of our estimation procedure. The Monte Carlo results indicate that the estimation procedure yields unbiased estimates for all parameters of interest. This is the case both: *(i)* when estimating the

accurate anticipation model that uses two stages, first estimating λ^I using investor data and then using this estimate to identify λ^A, and also *(ii)* when estimating the *subjective response model* that simultaneously estimates λ^I and λ^A using advisor data only.

Table B.7: Monte Carlo experiments

	$\lambda^A = 3, \lambda^I = 2$		$\lambda^A = 6, \lambda^I = 4$	
	Accurate anticipation	Subjective response	Accurate anticipation	Subjective response
$\hat{\lambda}^A$	3.024***	2.993***	5.951***	6.001***
	(0.344)	(0.344)	(0.423)	(0.543)
$\hat{\lambda}^I$	1.982***	2.032**	4.0***	4.031***
	(0.34)	(0.929)	(0.496)	(0.703)
Log Likelihood	-3889.169	-3889.495	-3790.913	-3785.953
Observations: Investors	900	–	900	–
Observations: Advisors	600	600	600	600

Notes: Columns (1) and (2) present the mean and standard deviations of parameter estimates that have been made based on Monte Carlo simulation that (i) randomly generate a true DGP, draw a data set from that DGP and randomly draw a competing model; (ii) draw an advisor's model conditional on the data, competing model, and the noise parameters $\lambda^A = 3$ and the advisor's expectation over $\lambda^I (= 2)$; (iii) draw an investor's assessment based on the data, the two narratives, and $\lambda^I = 2$. Column (1) presents the mean and standard deviations of parameter estimates that follow the two-step procedure, first estimating λ^I using the simulated assessment data and then estimating λ^A conditional on this estimate using advisor data. Column (2) presents the mean and standard deviations of estimates that only use advisor data for identification. Columns (3) and (4) repeat these exercises using the true parameter values $\lambda^A = 6$ and $\lambda^I = 4$ for simulation. All are based on 100 simulations and estimations for each column. * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

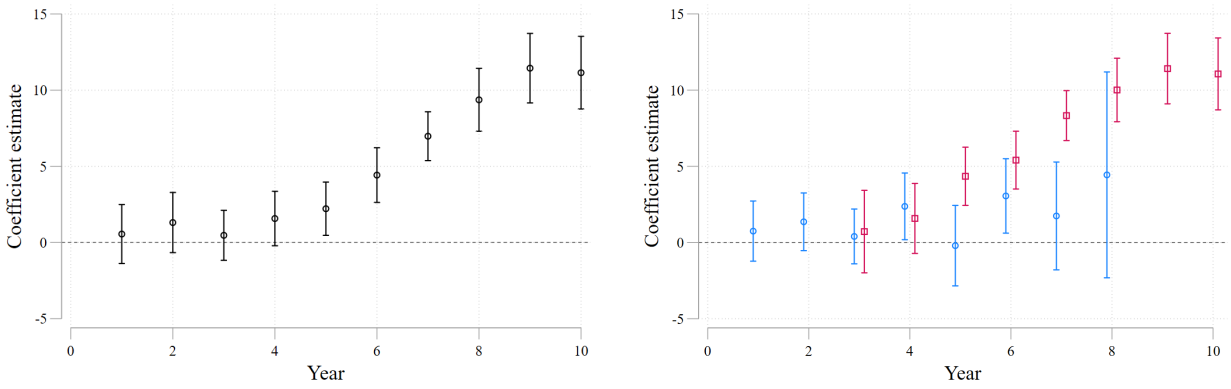
B.7 Shaping the Interpretation of Data: Direct evidence

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:

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

In the equation above, s_t indicates a success in year t and ρ are round fixed effects. The left panel of Figure B.6 plots the β-coefficient estimates. The qualitative pattern of the coefficient estimates implies that investors interpret the data in a reasonable way. 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.

Figure B.6: 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.

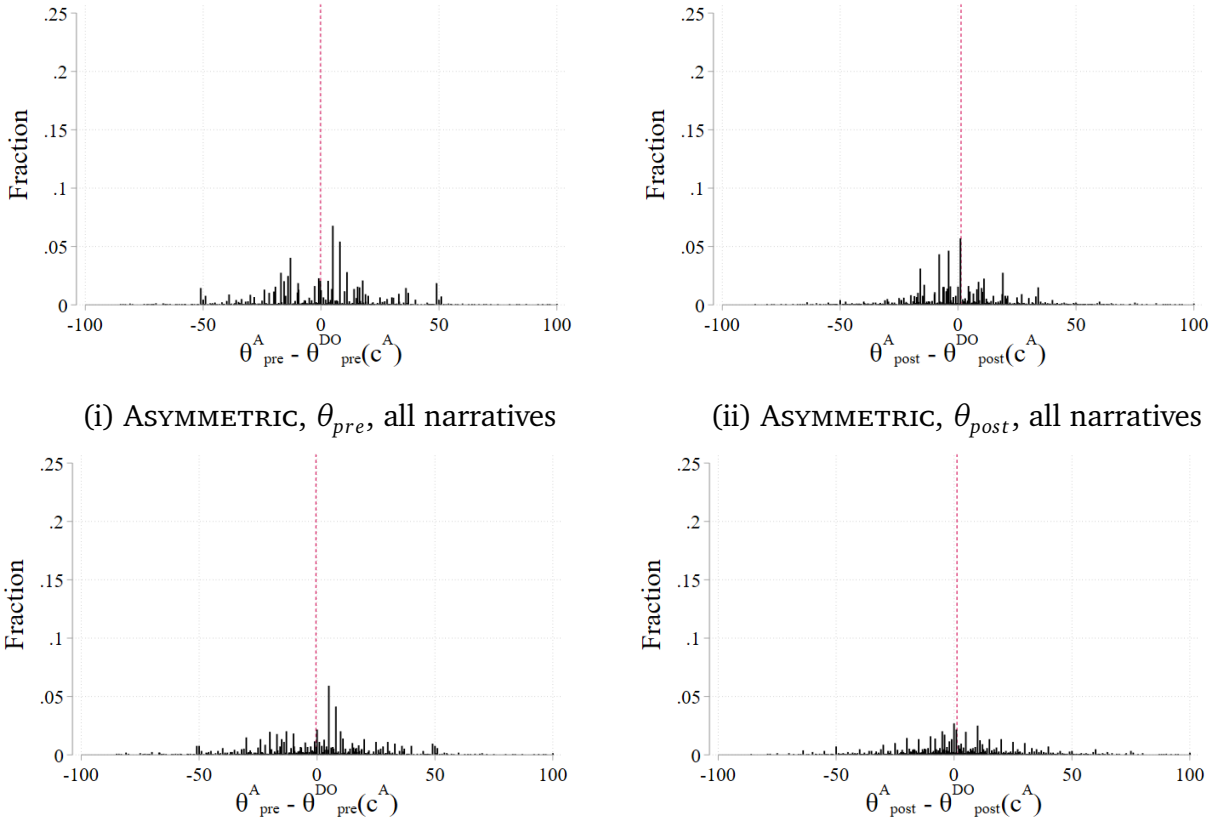
By sending a narrative, the advisor can potentially change how the investor interprets the data. In particular, by providing a suggestion regarding the year in which the CEO changed, the advisor essentially tells the investor which years to focus on to assess the company's future probability of success. The right panel of Figure B.6 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 provides 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 when those years are in the pre period (blue) according to the advisor's narrative ($p < 0.001$).[45] This result is consistent with the idea that the advisor influences which years in the data the investor deems relevant when making his assessment.

B.8 Do Advisors Disguise Lying by Adding Noise to their Narratives?

In the ASYMMETRIC treatment, advisors know the true DGP (and investors know this). If an advisor lies and constructs a narrative that differs from the truth, according to the S&S framework, they will want to ensure that the narrative they construct fits the data well. However, one can argue that they should not choose their narratives such that they fit *too well*. The argument is the following. Conditional on choosing a particular structural change parameter, c^A, the best-fitting θ_{pre} and θ_{post} parameters will be equal to the empirical fraction of successes in the pre and post-periods, respectively. However, typically, the data will not perfectly coincide with the truth in the sense that, given a true underlying DGP, in most instances, the true θ^T_{pre} and θ^T_{post} will not exactly equal the fraction of successes in the pre and post periods, respectively. Therefore, if an advisor constructs a narrative with θ—parameters that do equal the fraction of successes exactly, a sophisticated investor may find it suspicious. If advisors anticipate this, then they may choose to construct narratives that don't perfectly match the empirical success fractions in the data.

[45] In an earlier working paper version of this study, we included some additional discussion, regression output and formal tests providing further evidence in support of this. Please refer to Section 5.3.2 and Appendix Section B.4 of Barron and Fries (2023).

Figure B.7: Histograms of the difference between θ^A and the conditionally data-optimal parameter (Asymmetric).



(i) Asymmetric, θ_{pre}, all narratives

(ii) Asymmetric, θ_{post}, all narratives

(iii) Asymmetric, θ_{pre}, only narratives that are lies

(iv) Asymmetric, θ_{post}, only narratives that are lies

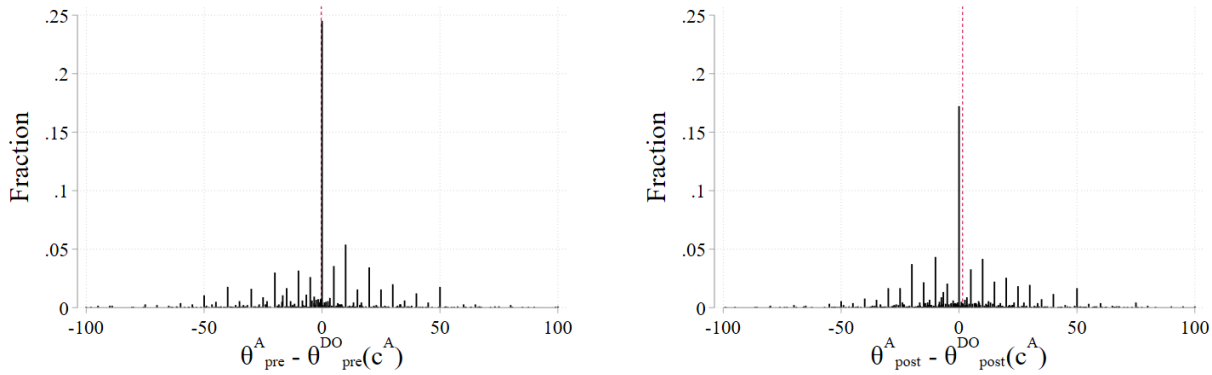
Notes: The figures use data from the Asymmetric treatment. The top two panels use data of all advisors, while the bottom two panels only use data of advisors who lie on at least one dimension of the narrative. The red vertical lines plot the mean difference.

Figure B.7 presents evidence on the distribution of the difference between the θ—value sent by the advisor and the corresponding data-optimal value, conditional on c^A in Asymmetric. If all advisors were to always match the success frequency exactly, we would expect the distributions to have a single mass point at 0. We see that the distributions are centered around 0, but, interestingly, advisors rarely send narratives with θ-parameters that closely match the success fractions in the data. Importantly, the bottom two panels of Figure B.7 shows that is true even when we consider only the advisors who send narratives that contain a lie (i.e., they deviate from the true DGP). The reason why it is useful to restrict attention to advisors who are lying is that truth-telling advisors will typically not have θ^A_{pre} and θ^A_{post} parameters that match the empirical success fractions in the data, precisely because the historical data is noisy relative to the truth. It is exactly this noise that a sophisticated advisor may be trying to simulate. To summarize, in Asymmetric, we find that advisors often send parameters that do not exactly match the empirical success probabilities. Their matching frequency further does not increase when we restrict the attention to advisors who lie.

In contrast, Figure B.8 shows that the pattern of behavior is strikingly different in Symmetric. Here, advisors choose narratives that match the empirical success frequencies far more often. This shows that advisors are able to do so. This evidence suggests that advisors in Asymmetric intentionally choose narratives that do not match the empirical success frequencies—potentially

to disguise their lies by adding some noise to ensure that their narratives do not fit *too well*.

Figure B.8: Histograms of the difference between θ^A and the conditionally data-optimal parameter (SYMMETRIC).



(i) SYMMETRIC, θ_{pre}, all narratives

(ii) SYMMETRIC, θ_{post}, all narratives

Notes: The figures use data from the SYMMETRIC treatment. The top two panels use data of all advisors, while the bottom two panels only use data of advisors who lie on at least one dimension of the narrative. The red vertical lines plot the mean difference.

C Additional Discussion of S&S's Narrative Approach

C.1 Notation

Throughout the discussion, we will use several 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

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

C.2 Proof of Proposition 1

(i). For a given pdf over default narratives of the investor $f(\boldsymbol{m}^{I,0})$, one can derive a distribution over the possible empirical fit of default narratives. In particular, the function

$$\tilde{f}(l) = \int_{\boldsymbol{m}\in\mathcal{M}} \mathbb{I}(\ell(\boldsymbol{m}) = l) f(\boldsymbol{m})\, \mathrm{d}\boldsymbol{m}$$

denotes the pdf of the log likelihood fit of default narratives and $\tilde{F}(l) = \int_{-\infty}^{l} \tilde{f}(s)\, \mathrm{d}s$ is the cdf. This distribution has full support on $(-\infty, l(\boldsymbol{m}^{DO})]$.

Using this notation, the advisor's expected utility from sending a message $\boldsymbol{m}$ is

$$\mathbb{E}_f[U^{\varphi}(\theta^I_{post}(\boldsymbol{m}^{I,0}, \boldsymbol{m}))|\boldsymbol{m}] = \tilde{F}(\ell(\boldsymbol{m}))U^{\varphi}(\theta^A_{post}) + (1-\tilde{F}(\ell(\boldsymbol{m})))\mathbb{E}_f[U^{\varphi}(\theta^{I,0}_{post})|\ell(\boldsymbol{m}) < \ell(\boldsymbol{m}^{I,0})].$$

Taking the first-order condition with respect to θ^A_{post} gives

$$\begin{aligned}\frac{\partial \mathbb{E}_f[U^{\varphi}(\theta^I_{post}(\boldsymbol{m}^{I,0}, \boldsymbol{m}^A))|\boldsymbol{m}^A]}{\partial \theta^A_{post}} &= \tilde{f}(\ell(\boldsymbol{m}^A))\frac{\partial \ell(\boldsymbol{m}^A)}{\partial \theta^A_{post}}\left(U^{\varphi}(\theta^A_{post}) - \mathbb{E}_f[U^{\varphi}(\theta^{I,0}_{post})|\ell(\boldsymbol{m}^{DO}) < \ell(\boldsymbol{m}^{I,0})]\right) \\ &\quad + \tilde{F}(\ell(\boldsymbol{m}^A))\frac{\partial U^{\varphi}(\theta^A_{post})}{\partial \theta^A_{post}} + (1-\tilde{F}(\ell(\boldsymbol{m}^A)))\frac{\partial \mathbb{E}_f[U^{\varphi}(\theta^{I,0}_{post})|\ell(\boldsymbol{m}^A) < \ell(\boldsymbol{m}^{I,0})]}{\partial \theta^A_{post}}.\end{aligned}$$

Now, if we evaluate the derivative at the data-optimal model, $\tilde{F}(\ell(\boldsymbol{m}^{DO})) = 1$ and $\frac{\partial \ell(\boldsymbol{m}^{DO})}{\partial \theta_{post}} = 0$. Therefore, the derivative simplifies to

$$\frac{\partial \mathbb{E}_f[U^{\varphi}(\theta^I_{post}(\boldsymbol{m}^{I,0}, \boldsymbol{m}^{DO}))|\boldsymbol{m}^{DO}]}{\partial \theta^A_{post}} = \frac{\partial U^{\varphi}(\theta^{DO}_{post})}{\partial \theta^A_{post}},$$

which is positive if $\frac{\partial U^{\varphi}(\theta^{DO}_{post})}{\partial \theta^A_{post}} \geq 0$ and otherwise negative. Since U^{φ} is strictly concave with an optimum at the persuasion target ϕ, Part *(i)* follows.

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

$$\mathcal{C} \equiv \left\{ \boldsymbol{m} \in \mathcal{M} | \theta_{post} \in [0,1], \theta_{pre} = \hat{\theta}_{pre}(\theta_{post}), c = \hat{c}(\theta_{post}) \right\}.$$

The proof will proceed by showing and combining a number of claims. These claims will culminate in the conclusion that the advisor's optimal message, that we denote by $\boldsymbol{m}^*$, is always part of the set $\mathcal{C}$.

Claim 1: For every $\theta_{post} \in [0,1]$, there are always parameter values $c \in \{2, \ldots, 8\}$ 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 message $(\hat{c}(\theta_{post}), \hat{\theta}_{pre}(\theta_{post}), \theta_{post})$ induces likelihood value $\bar{\ell}$. Now consider $\bar{\ell}$ taking on a value on the interior of the interval. We know that

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

Consider changing $\hat{\theta}_{pre}$ to a value 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 $\ell((\hat{c}(\theta_{post}), t, \theta_{post})) = \bar{\ell}$. We conclude that we can always fix θ_{post} and find auxiliary parameter values that induce any likelihood fit on $(-\infty, \ell^C(\theta_{post})]$.

The next claim makes the following comparison: Compare any messages that induce the same likelihood fit $\bar{\ell}$. Then, the advisor will prefer to send the message with a θ_{post}−value that is closest to ϕ. For the proof, we introduce a correspondence, which, for a given log-likelihood value $\bar{\ell}$, returns all messages whose fit is equal to that value:

$$\dot{\mathcal{M}}(\bar{\ell}) = \left\{ \boldsymbol{m} \in \mathcal{M} | \ell(\boldsymbol{m}) = \bar{\ell} \right\}.$$

Claim 2: Among all $\boldsymbol{m} \in \dot{\mathcal{M}}(\bar{\ell})$, the advisor chooses the $\boldsymbol{m}$ that minimizes the distance between θ^A_{post} and ϕ: $\boldsymbol{m}^*(\bar{\ell}) \in \underset{\boldsymbol{m} \in \mathcal{M}(\ell)}{arg\,min}(\phi - \theta_{post})^2$. Sending a message $\boldsymbol{m}' = (c', \theta'_{pre}, \theta'_{post}) \in \mathcal{M}(\bar{\ell})$ yields utility

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

Note that any alternative model in $\dot{\mathcal{M}}(\bar{\ell})$ only changes the value of $U^{\varphi}(\cdot)$ in the first term of the utility function, while the values of all other components remain fixed, as they only depend on $\bar{\ell}$. Therefore, choosing the message that maximizes utility for a given level of fit $\bar{\ell}$ is equal to

maximizing the utility the advisor receives if the investor adopts the message, $U^{\varphi}(\theta^{A}_{post})$, with respect to θ^{A}_{post}. This in turn is equal to minimizing $(\phi - \theta^{A}_{post})^2$.

Claim 3: Suppose that $\theta^{*}_{post} \neq \phi$. Then, $\boldsymbol{m}^{*} \in \mathcal{C}$. Suppose by contradiction that $\theta^{*}_{post} \neq \phi$ and $\boldsymbol{m}^{*} \neq \mathcal{C}$. 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 $(\phi - \theta'_{post})^2 < (\phi - \theta^{*}_{post})^2$. By Claim 1, we know that a message $\boldsymbol{m}' = (c', \theta'_{pre}, \theta'_{post})$ exists such that $\ell(\boldsymbol{m}') = \ell(\boldsymbol{m}^{*})$ as long as $\theta^{*}_{post} \notin \mathcal{C}$. By Claim 2, the advisor prefers message m' to message m^{*}, which contradicts the initial statement.

Claim 4: Consider two messages $\boldsymbol{m}' = (c', \theta'_{pre}, \phi)$ and $\boldsymbol{m}'' = (c'', \theta''_{pre}, \phi)$ and suppose that $\ell(\boldsymbol{m}') > \ell(\boldsymbol{m}'')$. The advisor prefers sending $\boldsymbol{m}'$ over sending $\boldsymbol{m}''$. Denote by $\Delta\tilde{F}$ the difference $\tilde{F}(\ell(\boldsymbol{m}')) - \tilde{F}(\ell(\boldsymbol{m}''))$. For notational brevity we will also use $\tilde{F}'' \equiv \tilde{F}(\ell(\boldsymbol{m}''))$, $\ell' \equiv \ell(\boldsymbol{m}')$, $\ell'' \equiv \ell(\boldsymbol{m}'')$, and $\ell^{I,0} \equiv \ell(\boldsymbol{m}^{I,0})$. We can then denote the expected utility of the sender from sending $\boldsymbol{m}'$ as

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

The inequality above follows from the fact that the advisor's prior over the investor's possible default narratives has full support on $\mathcal{M}$, so that the set of messages among which the investor's default narrative is if the investor follows message $\boldsymbol{m}'$ but not message $\boldsymbol{m}''$ will always include some message with a value $\theta_{post} \neq \phi$ with positive likelihood, which implies that $U^{\varphi}(\phi) > \mathbb{E}[U^{\varphi}(\theta^{I,0}_{post})|\ell^{I,0} \in (\ell'', \ell')]$. Therefore, $\mathbb{E}[U^{\varphi}(\theta^{I}_{post}(\boldsymbol{m}'', \boldsymbol{m}^{I,0})))|\boldsymbol{m}'] > \mathbb{E}[U^{\varphi}(\theta^{I}_{post}(\boldsymbol{m}'', \boldsymbol{m}^{I,0})))|\boldsymbol{m}'']$, which proves the claim.

Combining claims 3 and 4, the statement of the proposition directly follows.

Claim 5: $\boldsymbol{m}^{*} \in \mathcal{C}$. This follows directly by combining the statements of claims 3 and 4.

C.3 Discussion of the Predictions

When discussing the predictions for the experiment, we will generally consider a setup with a large pool of (heterogeneous) investors and advisors that are first randomly matched, then the advisor sends a message, and finally the investor makes an assessment. Denote the distribution of default narratives by $F(\boldsymbol{m}^{I,0})$ and the distribution of default model log likelihood fits (that can be derived from F, see the proof of Proposition 1) by $\tilde{F}(\ell)$.

Prediction 1 (Persuasion in pure interpretation and hybrid scenarios). The assessment rule in Equation (3) suggests that the investor will adopt if the fit of the advisor's message is sufficiently high, regardless of the advisor's knowledge relative to the investor. Therefore, messages can be influential with and without knowledge.

Prediction 2 (Influence of message fit). Consider two populations of investors who draw their default narrative from f and two populations of advisors who send the same θ^A_{post} but vary their messages in the auxiliary parameters, so that population α sends a message with likelihood fit ℓ_α and population β sends a message with fit $\ell_\beta > \ell_\alpha$.

There can be three cases:

With probability $F(\ell_\alpha)$, the investor's default narrative has a fit that is smaller than that of the narratives of $\alpha-$ and $\beta-$advisors. He will adopt and make assessment θ^A_{post} in either case.

With probability $F(\ell_\beta) - F(\ell_\alpha)$, the investor's default narrative fit is smaller than the $\beta-$ but larger than the $\alpha-$advisor's narrative fit. The investor will only adopt the $\beta-$advisor's narrative.

With probability $1 - F(\ell_\beta)$, the investor will not adopt the $\alpha-$ and the $\beta-$advisor's narrative.

This suggests that the expected distance to the narrative after meeting the $\alpha-$advisor is

$$(F(\ell_\beta) - F(\ell_\alpha))\mathbb{E}[|\theta^{I,0}_{post} - \theta^A_{post}||\ell^{I,0} \in (\ell_\alpha, \ell_\beta]] + (1 - F(\ell_\beta))\mathbb{E}[|\theta^{I,0}_{post} - \theta^A_{post}||\ell^{I,0} > \ell_\beta],$$

which is larger than the expected distance after meeting the $\beta-$advisor,

$$(1 - F(\ell_\beta))\mathbb{E}[|\theta^{I,0}_{post} - \theta^A_{post}||\ell^{I,0} > \ell_\beta].$$

Therefore, we expect higher-fitting messages to move the investor's assessment closer to the advisor's narrative.

Prediction 3 (Negative correlation). The prediction is based on the two observations from Proposition 1: *(i)*, the advisor always sends a message with a θ^A_{post} between θ^{DO}_{post} and the persuasion target ϕ and that, *(ii)*, the advisor sends the data-optimal auxiliary parameters conditional on θ^A_{post}. We call the set of messages that the proposition does not rule out the "likelihood frontier." Consider an up-advisor who constructs a message and perceives a relatively strict tradeoff between belief movement and empirical fit, which would indicate that she sends a message with a θ^A_{post} that is close to θ^{DO}_{post}. For such a message, it will typically be optimal to send along the data-optimal auxilliary parameters c^{DO} and θ^{DO}_{pre}; since they are optimal for θ^{DO}_{post}, they are also optimal for a θ^A_{post} that is close enough to θ^{DO}_{post}. Therefore, such an adivsor will always slightly exaggerate θ^A_{post} above the data-optimum while keeping θ^A_{pre} at the data-optimum. Since the expected values of θ^{DO}_{post} and θ^{DO}_{pre} are 0.5 if the data-generating parameters are randomly drawn from independent uniform distributions, we would expect that $\mathbb{E}(\theta^A_{post}|\varphi = \uparrow) > \mathbb{E}(\theta^A_{pre}|\varphi = \uparrow)$. If the advisor exaggerates θ^A_{post} by more, moving along the likelihood frontier might induce her to adjust the auxiliary parameters to support the empirical fit. Which alternative auxiliary parameters will the advisor entertain? The following results analytically characterizes regularities.

Proposition 3. *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 post period is lower under c^A than c^{DO}, i.e.,

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

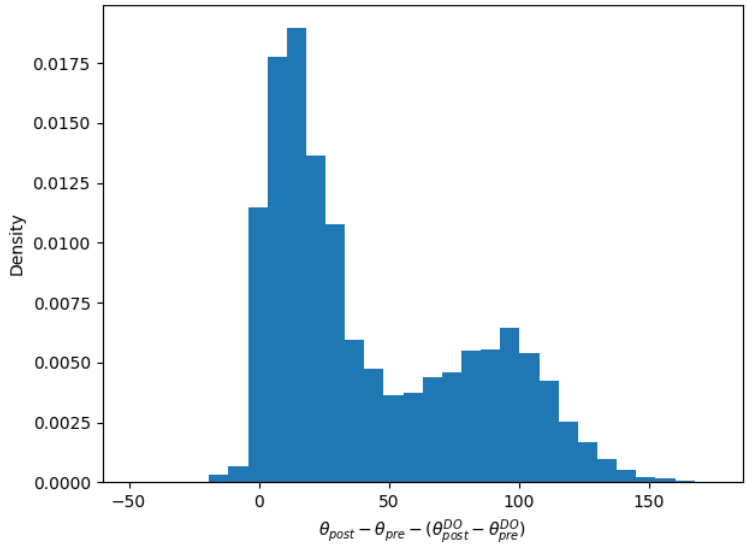
Proof. See Appendix E. □

These regularities will tend to induce the up-advisor to decrease and the down-advisor to increase θ_{pre} when moving away from sending θ^{DO}_{pre}, which is in line with the prediction: An up-advisor for example will only decrease c^A below the data-optimum if such a decrease increases the fraction of successes in *post* and only increase c^A if doing so decreases the number of failures in *post*. These adjustments will in turn tend to decrease the fraction of successes in *pre*. We can get a general sense of how systematic these tendencies are by calculating the likelihood frontiers for all 1024 possible histories that indiviudals could encounter. If we simply average the messages that are in the up-advisor's likelihood frontier for every history and then average the average messages over all histories,[46] we find that the average θ^A_{post} for the up-advisor is equal to 0.76 while the average θ^A_{pre} is equal to 0.45. Since this is a perfectly symmetrical problem, the expected values for the down-advisor are $\theta^A_{post} = 0.24$ and $\theta^A_{pre} = 0.55$. This is a further indication that we should expect the misaligned advisors to move θ^A_{post} and θ^A_{pre} into opposite directions.

We further look at how the difference between $\theta^A_{post} - \theta^A_{pre}$ evolves while moving along the up-advisor's likelihood frontier. Note that the expected value $\mathbb{E}(\theta^{DO}_{post} - \theta^{DO}_{pre})$ is equal to zero if the parameters of the D.G.P. are drawn randomly from independent uniform distributions. Therefore, if the difference in differences $\theta^A_{post} - \theta^A_{pre} - (\theta^{DO}_{post} - \theta^{DO}_{pre})$ is positive for points on the frontier, this also implies that $\mathbb{E}(\theta^A_{post}|\varphi = \uparrow) > \mathbb{E}(\theta^A_{pre}|\varphi = \uparrow)$. The histogram below plots all possible differences-in-differences for any message in the up-advisor's likelihood frontier for all possible histories. We can see that this difference is positive for almost any point on the frontier, and therefore for almost any possible message that the up-advisor is expected to send. In fact, out of the 1024 possible histories, only in a minority of them (139) there is a message under which the difference in differences is negative. This provides further arguments for Prediction 3.

[46] Taking the average twice ensures that the average message of each history is given equal weight when taking the overall average. If we instead average over the collection of all messages that could have been send by combining the likelihood frontiers of any history, we would implicitly attach more weights to histories with a larger likelihood frontier.

Figure C.1: Distribution of difference in differences between $\theta^A_{post} - \theta^A_{pre}$ and $\theta^{DO}_{post} - \theta^{DO}_{pre}$



Note: This histogram plots the diff-in-diff between $\theta^A_{post} - \theta^A_{pre}$ and $\theta^{DO}_{post} - \theta^{DO}_{pre}$ for all possible messages that are on the likelihood frontier of the up-advisor for all histories that she might encounter. The θ values of the plotted messages are discretized to $\{0, .01, \ldots, 1\}$.

Prediction 4 (Fit-movement tradeoff). Fix any history and consider the likelihood frontier of the up-advisor for this history. Messages sent by the advisor can induce any empirical fit from $\ell_{min} \geq -\infty$ to $\ell(\boldsymbol{m}^{DO})$. Suppose that the fit of the investor's default narrative, $\ell(\boldsymbol{m}^{I,0})$, is smaller than ℓ_{min}. Then, the advisor can send a message inducing the persuasion target ϕ. Instead, if $\ell(\boldsymbol{m}^{I,0}) \in (\ell_{min}, \ell(\boldsymbol{m}^{DO}))$, the advisor will always send a message with fit $\ell(\boldsymbol{m}^{I,0})$: First, the advisor is always weakly better of by getting the investor to adopt their message, and therefore she will only consider sending messages which have a weakly higher fit than the default narrative. Second, suppose by contradiction that the advisor sends a message with a strictly higher fit. Because the likelihood function is continuous in θ_{post}, the advisor could always move θ^A_{post} closer to the persuasion target and still guarantee that it is adopted by the investor as long as its fit is weakly larger than $\ell(\boldsymbol{m}^{I,0})$. Therefore, sending a message with a strictly higher fit is not optimal. Part *(i)* now directly follows from this argument, as increasing the fit of the default narrative leads to a one-to-one increase in the fit of the advisor's message. Part *(ii)* follows from a revealed preference argument. Consider two default narratives $\boldsymbol{m}^{I,0'}$ and $\boldsymbol{m}^{I,0''}$, with $\ell(\boldsymbol{m}^{I,0'}) < \ell(\boldsymbol{m}^{I,0''})$. The set of possible messages that are adopted by the investor is larger under $\boldsymbol{m}^{I,0'}$ than under $\boldsymbol{m}^{I,0''}$. Since the advisor can always induce the same assessment when facing $\boldsymbol{m}^{I,0'}$ than when facing $\boldsymbol{m}^{I,0''}$, she can never be worse off when the default narrative is $\boldsymbol{m}^{I,0'}$. For this reason, the distance between message and persuasion target must be smaller when facing $\boldsymbol{m}^{I,0'}$ than when facing $\boldsymbol{m}^{I,0''}$.

D Discussion of the Cheap Talk Benchmark

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

D.1 Setup

Consider a game between an advisor and an investor. There is an unknown true data generating process, or model, $\boldsymbol{m}^T = (c^T, \theta^T_{pre}, \theta^T_{post}) \in \mathcal{M} \equiv \{2,\dots,8\} \times [0,1]^2$. Nature draws this model from a distribution $G(\boldsymbol{m}^T)$ with pdf $g(\boldsymbol{m}^T)$. Denote the expectation of θ^T_{post} given G by $\bar{\theta}$. We comment on the exact shape of G below. The advisor observes $\boldsymbol{m}^T$. We comment on the case where the advisor does not observe $\boldsymbol{m}^T$ (where she is uninformed) further below. The investor does not observe $\boldsymbol{m}^T$, but it is common knowledge that $\boldsymbol{m}^T$ is distributed according to $G(\boldsymbol{m}^T)$ and. After observing $\boldsymbol{m}^T$, the advisor sends a message $\boldsymbol{m} \in \mathcal{M}$ to the investor. The investor then makes an assessment $\theta^I_{post} \in [0,1]$. The investor's utility depends on the assessment and θ^T_{post}. It is maximized if the investor makes an accurate assessment:

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

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 persuasion target. The advisor can be one of three incentive-types; up, down, and aligned, which we also denote using $\uparrow, \downarrow$, and $\rightarrow$ respectively. The advisor's utility depends on the investor's assessment, θ^I_{post}, and her incentive type φ;

$$U^{\varphi}(\theta^I_{post}) = \begin{cases} 1-(1-\theta^I_{post})^2 & \text{if } \varphi = \uparrow, \\ 1-(0-\theta^I_{post})^2 & \text{if } \varphi = \downarrow, \\ 1-(\theta^T_{post}-\theta^I_{post})^2 & \text{if } \varphi = \rightarrow. \end{cases} \tag{4}$$

This utility is maximized if θ^I_{post} equals the persuasion target; the up-advisor wants the investor to make the highest possible assessment of $\theta^I_{post} = 1$, the down-advisor wants the investor to make the lowest possible assessment of $\theta^I_{post} = 0$, and the aligned advisor wants the investor to make an accurate assessment. In the following, we denote the persuasion target by ϕ.[47] At the start of the game, nature draws the advisor's incentive type, and each type is equally likely. The advisor knows her incentive type, but the investor does not. Information about the incentive type distribution is common knowledge.

D.2 Behavioral Types

For our main analysis, we are going to assume that any advisor is a *honest type* ($h = 1$) with probability $\lambda \in [0,1]$. An honest advisor always follows a truth-telling strategy, i.e, she always sends $\boldsymbol{m} = \boldsymbol{m}^T$ regardless of her incentive-type. An advisor who is not honest is *strategic* ($h = 0$). A strategic advisor sends a message to maximize her expected utility. Information about honest types and the value of λ is common knowledge. One important implication of introducing honest types is that any message in the support of G is sent with positive probability, since there is a nonzero chance that an honest advisor observes it.

[47] I.e., $\phi = 1$ if $\varphi = \uparrow$, $\phi = 0$ if $\varphi = \downarrow$ and $\phi = \theta^T_{post}$ if $\varphi = \rightarrow$.

Observation 1. *If $\lambda > 0$, any message that is in the support of G is sent with positive probability in equilibrium.*

Types and terminology. The advisor's *type* is defined by her incentives (up,down,aligned), her information about the true model ($\boldsymbol{m}^T$), and her behavioral type (honest or strategic). We will abuse terminology and often omit the behavioral type when discussing different agents. For example, when we mention the up-advisor, we will typically mean the strategic up-advisor. We will use $\boldsymbol{\tau} = (\varphi, c^T, \theta^T_{pre}, \theta^T_{post}, h)$ to denote the advisor's type (we will sometimes write this as $\boldsymbol{\tau} = (\varphi, \boldsymbol{m}^T, h)$) and $\mathcal{T}$ to denote the type space. The investor's initial belief about the advisor's type is $\mu(\boldsymbol{\tau})$.[48] The updated belief after receiving message $\boldsymbol{m}^A$ is $\mu(\boldsymbol{\tau}|\boldsymbol{m}^A)$.

Remark: relating theory to design; how does G look like? We can think of the historical data, jointly with the information that the three parameter values $c^T, \theta^T_{pre}, \theta^T_{post}$ are uniformly distributed on $\{2,8\} \times [0,1]^2$ ex-ante, determining the belief $g_{\theta_{post}}(\theta^T_{post})$. Formally, upon seeing the data, the investor can form a Bayesian belief over θ^T_{post} which is equal to

$$g_{\theta^T_{post}}(\theta^T_{post}) = \sum_{c=2}^{8} \frac{\int_0^1 \mathcal{L}(c^T, \theta^T_{pre}, \theta^T_{post})\, \mathrm{d}\theta^T_{pre}}{\sum_{c=2}^{8} \int_0^1 \int_0^1 \mathcal{L}(c^T, \theta^T_{pre}, \theta^T_{post})\, \mathrm{d}\theta^T_{pre} \mathrm{d}\theta^T_{post}}. \tag{5}$$

This expression gives the pdf of the marginal distribution of θ^T_{post} given g. In the equation above, $\mathcal{L}(c^T, \theta^T_{pre}, \theta^T_{post}) = \theta_{pre}^{Tk_{pre}(c^T)}(1-\theta^T_{pre})^{f_{pre}(c^T)} \theta_{post}^{Tk_{post}(c)}(1-\theta^T_{post})^{f_{post}(c^T)}$ is the likelihood function, and $k_p(c^T)$, $f_p(c^T)$ denote the number of successes and failures in the *pre* and *post* period for a given structural change parameter value c^T. We can simplify this 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 (5), the marginal density of θ^T_{post} becomes

$$g_{\theta^T_{post}}(\theta^T_{post}) = \sum_{c=2}^{8} w_c h(\theta^T_{post} | k_{post}(c^T)+1, f_{post}(c^T)+1),$$
$$\text{where } w_c \equiv \frac{B(k_{pre}(c^T)+1, f_{pre}(c^T)+1) B(k_{post}(c^T)+1, f_{post}(c^T)+1)}{\sum_{c'=2}^{8} B(k_{pre}(c')+1, f_{pre}(c')+1) B(k_{post}(c')+1, f_{post}(c')+1)}.$$

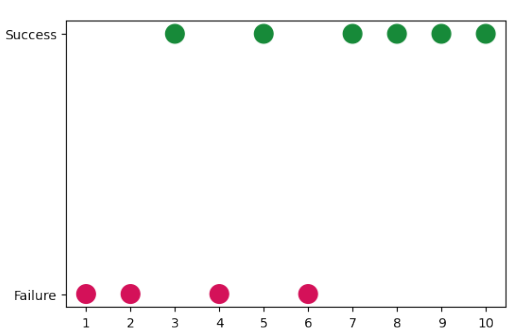
Therefore, the investor's belief distribution over θ^T_{post} is a mixture of beta distributions with expectation $\bar{\theta} \in (0,1)$. Figure D.1 shows the investor's belief density for an example historical data set.

In the text below, whenever we refer to G, we refer to a distribution that is derived in the way described above and whose marginal pdf with respect to θ^T_{post} is given above. We will also generally assume that G has full support on $\mathcal{M}$, which is true in all but a few special cases.[49] This is purely for simplification. The results below can be extended to the case where G has a restricted support, but the notation becomes more cumbersome.

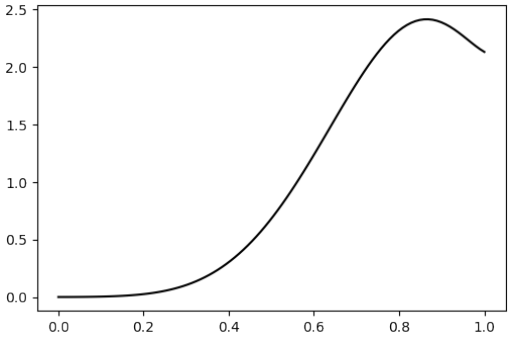
[48] The initial belief is equal to $\mu(\boldsymbol{\tau}) = \frac{1}{3} g(\boldsymbol{m}^T)(\lambda h + (1-\lambda)(1-h))$.

[49] The only cases where models leave the support are those where the data suggests that $\theta^T_{post} \neq 1$, $\theta^T_{post} \neq 0$, $\theta^T_{pre} \neq 1$, or $\theta^T_{pre} \neq 0$. For example, if there is at least one failure in period 10, then a model with $\theta_{post} = 1$ would not be in the support of G.

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



Historical data



Belief density over θ_{post}

D.3 Equilibrium

In the described game, the advisor's strategy maps from the advisor's type into a probability distribution over messages. Denote by $\sigma(\boldsymbol{m}^A|\boldsymbol{\tau})$ the probability that an advisor with type $\boldsymbol{\tau}$ sends $\boldsymbol{m}^A$. The investor's assessment rule then maps the received message into an assessment. Denote the investor's assessment rule by $\theta^I_{post}(\boldsymbol{m}^A)$ (by concavity of the investor's utility function, restricting the investor to pure strategies is without loss). We investigate PBE in which (strategic) players maximize utility and where the investor uses Bayes' rule to update μ whenever possible.

We are interested in persuasive equilibria of this game. Following Little (2023), 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 $\theta^I_{post}(\boldsymbol{m}) \neq \bar{\theta}$. A persuasive equilibrium is an equilibrium where a persuasive message is sent with strictly positive probability.*

Suppose an equilibrium exists. Given the equilibrium assessment rule, we can define two sets of messages that induce the investor to make the highest or lowest possible assessment. Formally, define

$$\mathcal{M}^{max} \equiv \{\boldsymbol{m} \in \mathcal{M} | \boldsymbol{m} \in \arg\max_{\boldsymbol{m}' \in \mathcal{M}} \theta^I_{post}(\boldsymbol{m}')\} \text{ and } \mathcal{M}^{min} \equiv \{\boldsymbol{m} \in \mathcal{M} | \boldsymbol{m} \in \arg\min_{\boldsymbol{m}' \in \mathcal{M}} \theta^I_{post}(\boldsymbol{m}')\}.$$

We denote the maximum assessment the investor can be induced to make by a^{max} and the minimum assessment by a^{min}. The following result states that up-advisors always send a message in $\mathcal{M}^{max}$, down-advisors always send a message in $\mathcal{M}^{min}$, and aligned advisors always send a message in $\mathcal{M}^{max}$ if θ^T_{post} is sufficiently high (and a message in $\mathcal{M}^{min}$ if θ^T_{post} is sufficiently low).

Lemma 1. *In any equilibrium,*

(i) The up-advisor sends a message $\boldsymbol{m} \in \mathcal{M}^{max}$.

(ii) The down-advisor sends a message $\boldsymbol{m} \in \mathcal{M}^{min}$.

(iii) If $\theta^T_{post} \geq a^{max}$, the aligned advisor sends a message $\boldsymbol{m} \in \mathcal{M}^{max}$.

(iv) If $\theta^T_{post} \leq a^{min}$, the aligned advisor sends a message $\boldsymbol{m} \in \mathcal{M}^{min}$.

Proof. This follows directly from utility maximization. □

A prominent type of a persuasive equilibrium is what we call a Two Threshold Equilibrium. In such an equilibrium, the investor can be induced to make any assessment on an interval $[\theta_L, \theta_H]$.

Definition 2. *A Two Threshold Equilibrium (TTE) is an equilibrium characterized by two thresholds* $\theta_L < \bar{\theta} < \theta_H$ *and where the investor can be induced to make any assessment on* $[\theta_L, \theta_H]$.

The following result is similar to Lemma 1 but applied to the definition of the TTE.

Corollary 1. *In any TTE:*

(i) The up-advisor always induces the investor to make assessment θ_H.

(ii) The down-advisor always induces the investor to make assessment θ_L.

(iii) The aligned advisor always induces the investor to make assessment θ_H *if* $\theta^T_{post} \geq \theta_H$, θ_L *if* $\theta^T_{post} \leq \theta_L$, *and* θ^T_{post} *if* $\theta^T_{post} \in (\theta_L, \theta_H)$.

Proof. Follows directly from utility maximization and the definition of the TTE. □

We are now going to define a specific TTE, which we call a Truthful Two Threshold Equilibrium (TTTE). In this equilibrium, the aligned advisor always follows a truth-telling strategy, and the up- and down-advisors follow a strategy that is independent of $\boldsymbol{m}^T$.

Definition 3. *A Truthful Two Threshold Equilibrium (TTTE) is a TTE characterized by the following properties:*

(i) The aligned advisor follows a truth-telling strategy:

$$\sigma(\boldsymbol{m}^A = \boldsymbol{m}^T | \rightarrow, \boldsymbol{m}^T) = 1 \text{ and } \sigma(\boldsymbol{m} \neq \boldsymbol{m}^T | \rightarrow, \boldsymbol{m}^T) = 0.$$

(ii) The up-advisor's strategy is given by:

$$\sigma(\boldsymbol{m}^A | \uparrow) = g(\boldsymbol{m}^A)\frac{2\lambda + 1}{1 - \lambda}\frac{\max\{\theta^A_{post} - \theta_H, 0\}}{\theta_H - \bar{\theta}}.$$

(iii) The down-advisor's strategy is given by:

$$\sigma(\boldsymbol{m}^A | \downarrow) = g(\boldsymbol{m}^A)\frac{2\lambda + 1}{1 - \lambda}\frac{\max\{\theta_L - \theta^A_{post}, 0\}}{\bar{\theta} - \theta_L}.$$

(iv) The investor's assessment rule is given by:

$$\theta^I_{post}(\boldsymbol{m}^A) = \begin{cases} \theta_H & \text{if } \theta^A_{post} \geq \theta_H, \\ \theta_L & \text{if } \theta^A_{post} \leq \theta_L, \\ \theta^A_{post} & \text{if } \theta^A_{post} \in (\theta_L, \theta_H). \end{cases}$$

Let us develop an argument for why the strategies of the TTTE constitute an equilibrium. Recall from Lemma 1 that the up-advisor will induce the highest possible and the down-advisor the lowest possible assessment in *any* equilibrium. Any message gets filtered through the investor's assessment rule. "Inducing" an assessment means that the advisor sends a message for which the assessment rule prescribes the investor to make that assessment. If the equilibrium is persuasive (so that the highest assessment is different from the lowest assessment), then the up- and down-advisor never send the same message with positive probability. Otherwise, they would induce the same assessment with positive probability, contradicting Lemma 1.

Now suppose that the TTTE is an equilibrium. It prescribes the aligned advisor to follow an honest strategy. Messages now fall in one of three broad categories. First, there are messages which are sent by either the aligned or the honest advisor. Since the investor knows that the aligned advisor follows an honest strategy, he will update his expectation to $\mathbb{E}_\mu(\widetilde{\theta}_{post}|\boldsymbol{m}^A) = \theta^A_{post}$ upon receiving such a message. His optimal assessment consequently is θ^A_{post}.

Second, there are messages which are sent by either the honest/aligned advisor or the up-advisor. The up-advisor does not condition her message on $\boldsymbol{m}^T$ while the other possible advisors do. Therefore, the investor's expectation about θ^T_{post} is a weighted average of her initial expectation and θ^A_{post};

$$\begin{aligned} &\mathbb{E}_\mu(\widetilde{\theta}_{post}|\boldsymbol{m}^A) = p(\boldsymbol{m}^A)\theta^A_{post} + (1-p(\boldsymbol{m}^A))\bar{\theta}, \qquad (6)\\ &\text{where } p(\boldsymbol{m}^A) = \frac{(2\lambda+1)g(\boldsymbol{m}^A)}{(2\lambda+1)g(\boldsymbol{m}^A)+(1-\lambda)\sigma(\boldsymbol{m}^A|\uparrow)}. \end{aligned}$$

The weight put on θ^A_{post} (p) increases in the relative likelihood of meeting an honest/aligned advisor ($\frac{2\lambda+1}{1-\lambda}$), in the fit of $\boldsymbol{m}^A$ ($g(\boldsymbol{m}^A)$), and decreases in the probability with which the up advisor sends m^A ($\sigma(\boldsymbol{m}^A|\uparrow)$).

In equilibrium, the up-advisor induces θ_H. Therefore, $\mathbb{E}_\mu(\widetilde{\theta}_{post}|\boldsymbol{m}^A) \stackrel{!}{=} \theta_H$. Solving this equation for σ returns the advisor's optimal strategy

$$\sigma(\boldsymbol{m}^A|\uparrow) = g(\boldsymbol{m}^A)\frac{2\lambda+1}{1-\lambda}\frac{\max\{\theta^A_{post}-\theta_H, 0\}}{\theta_H-\bar{\theta}}.$$

Under this strategy the up-advisor randomizes among all messages with a $\theta^A_{post} \geq \theta_H$. The function σ is a pdf which must integrate to 1. Therefore, the equilibrium θ_H is implicitly defined in the equation

$$\int_{\boldsymbol{m}\in\mathcal{M}} \sigma(\boldsymbol{m}|\uparrow)\,\mathrm{d}\boldsymbol{m} = 1.$$

The third category of messages consists of those which are either sent by the honest/aligned advisor or by the down-advisor. Steps similar to those that we took when discussing the second case would derive the down-advisor's optimal strategy.

Given the strategies of others, all players find it optimal to follow their own optimal strategy. The proposition below shows that θ_H and θ_L exist and are unique.

Proposition 4. *There exists a unique TTTE.*

Proof. See Appendix E. □

The TTTE is not the only persuasive equilibrium. However, we can show that it is a most informative equilibrium in a sense that we define below and that any most informative equilibrium is essentially unique, i.e., any most informative equilibrium generates the same payoff distribution for all players.

Definition 4 (Most informative equilibrium)**.** *Define the expected squared assessment error as*

$$SAE \equiv \mathbb{E}[(\widetilde{\theta}^I_{post} - \widetilde{\theta}^T_{post})^2].$$

An equilibrium is most informative if there is no other equilibrium with a lower expected squared assessment error.

Proposition 5. *The TTTE is a most informative equilibrium. Any most informative equilibrium is a TTE that is characterized by the same thresholds θ_L and θ_H that characterize the TTTE.*

Proof. See Appendix E. □

We are now ready to characterize the investor's assessment rule in any most informative equilibrium.

Proposition 6. *In any most informative equilibrium of the game where either (i) $\lambda > 0$ or (ii) $\lambda = 0$ and the aligned advisor follows an honest strategy, the investor's assessment rule is:*

$$\theta^I_{post}(\boldsymbol{m}^A) = \begin{cases} \theta_H & \text{if } \theta^A_{post} \geq \theta_H, \\ \theta_L & \text{if } \theta^A_{post} \leq \theta_L, \\ \theta^A_{post} & \text{if } \theta^A_{post} \in (\theta_L, \theta_H). \end{cases}$$

Proof. See Appendix E. □

We provide one additional result to connect the discussion with the one in the main text.

Proposition 7. *Any equilibrium in which the aligned advisor follows an honest strategy is a most informative equilibrium.*

Proof. See Appendix E. □

Combining the last two propositions directly implies Proposition 2 in the main text that the investor's assessment is independent of the auxiliary parameters.

D.4 The Role of Honest Behavioral Types

Messages of honest types should be interpreted with their literal meaning. Therefore, if $\lambda > 0$, the investor thinks that there is a nonzero chance that he should take a message literally. This rules out a typical source of equilibrium multiplicity; namely, that one can simply relabel message strategies of strategic advisors to construct new most informative equilibria (Sobel, 2013). To illustrate, suppose that $\lambda = 0$ and take the TTTE. Change the advisor's strategies so that she swaps what she says about θ_{pre} and θ_{post}. For example, whenever an advisor sends $\boldsymbol{m}' = (c = c', \theta_{pre} = \theta'_{pre}, \theta_{post} = \theta'_{post})$ in the TTTE above, she now sends $\boldsymbol{m}' = (c = c', \theta_{pre} = \theta'_{post}, \theta_{post} = \theta'_{pre})$. The new set of strategies makes up a most informative equilibrium as long as the investor's assessment rule is changed accordingly to

$$\theta^I_{post}(\boldsymbol{m}^A) = \begin{cases} \theta_H & \text{if } \theta^A_{pre} \geq \theta_H, \\ \theta_L & \text{if } \theta^A_{pre} \leq \theta_L, \\ \theta^A_{pre} & \text{if } \theta^A_{pre} \in (\theta_L, \theta_H). \end{cases}$$

This shows that, when no advisor follows an honest strategy, Proposition 6 does not hold and we would need to enrich the statement in the main text to allow for multiple meanings of messages. In the equilibrium sketched out above, we could, for example, make the statement that what the

advisor sends about c and θ_{post} does not influence the investor's assessment. These parameters become essentially the new auxiliary parameters. In the equilibrium without honest types, choosing which parameters are auxiliary and which are not is a coordination problem. However, it seems plausible that the labels of *post* and *pre* give a natural meaning to the different dimensions of the message, which would favor an assessment rule as the one of Proposition 6.

D.5 What if the Advisor Does not Know the True State of the World?

If the advisor does not know the true state of the world, no persuasive equilibrium exists. The advisor's type is now two-dimensional (incentive-type and honest/strategic). Therefore, when the investor updates $\mu(\tau|\boldsymbol{m})$, he does no longer update over $\boldsymbol{m}^T$. The advisor cannot tell him anything he does not already know.

D.6 Biases in Information Processing

Economists have modified standard game theoretical concepts to accommodate biases information processing. One example is cursed equilibrium (Eyster and Rabin, 2005). In a cursed equilibrium, players neglect the connection between who other players are and what they do. An example more specific to the communication context is credulity, where message-receivers interpret them literally (Chen, 2011).

Information processing biases lead to non-Bayesian belief updating rules. We consider a $\psi-$biased investor who has an updated posterior distribution over the advisor's type $\mu^{\psi}(\tau|\boldsymbol{m}^A)$. The value of $\psi \in [0,1]$ determines the extent of the bias in a sense that we make precise below. In our context, we are going to investigate biases which lead the investor to form the following conditional expectation about θ^T_{post} after receiving $\boldsymbol{m}^A$:

$$\mathbb{E}_{\mu^{\psi}}(\widetilde{\theta}^T_{post}|\boldsymbol{m}^A) = \psi\, \nu(\cdot) + (1-\psi)\mathbb{E}_{\mu}(\theta^T_{post}|\boldsymbol{m}^A). \tag{7}$$

A $\psi-$biased investor's conditional expectation is a weighted average of the Bayesian expectation and of a function $\nu(\cdot)$ whose shape depends on the investigated bias. A $(\psi = 0)-$biased investor forms the Bayesian expectation and a $(\psi = 1)-$biased investor forms an expectation that is completely governed by the bias.

If $\nu(\boldsymbol{m}^A) = \theta^A_{post}$, the investor is *credulous*. He puts too much weight on a literal interpretation of the message. This should intuitively benefit a misaligned advisor. We will test this intuition by sketching out a TTTE with a $\psi-$credulous investor. First, Lemma 1 still applies; in an equilibrium, the up-advisor induces the highest possible assessment while the down-advisor induces the lowest possible assessment. Therefore, if $\boldsymbol{m}^A \in \mathcal{M}^{max}$, equilibrium requires that

$$\theta_H = \psi\theta^A_{post} + (1-\psi)\mathbb{E}_{\mu}(\theta^T_{post}|\boldsymbol{m}^A). \tag{8}$$

In a candidate TTTE, aligned advisors are honest. Therefore, $\mathbb{E}_{\mu}(\theta^T_{post}|\boldsymbol{m}^A) = p(\boldsymbol{m}^A)\theta^A_{post} + (1 - p(\boldsymbol{m}^A))\bar{t}$, where $p(\boldsymbol{m}^A)$ was defined in Equation (6). When we solve Equation (8) with respect to $\sigma(\boldsymbol{m}^A|\uparrow)$, we obtain an expression for the up-advisor's optimal strategy

$$\sigma(\boldsymbol{m}^A|\uparrow) = g(\boldsymbol{m}^A)\frac{(2\lambda+1)}{1-\lambda}\frac{\max\{\theta^A_{post} - \theta_H, 0\}}{\theta_H - (1-\psi)\bar{\theta} - \psi\theta^A_{post}}.$$

The equilibrium condition for the up-advisor is that she draws her message from a pdf, which implies that

$$\int_{\boldsymbol{m}^A \in \mathcal{M}} \sigma(\boldsymbol{m}^A | \uparrow) \, \mathrm{d}\boldsymbol{m}^A \stackrel{!}{=} 1$$

This equation pins down a unique $\theta_H(\psi)$; i.e., the highest assessment that the $\psi-$credulous investor can be induced to make. Our initial guess was correct; θ_H increases in ψ and approaches 1 as ψ approaches 1. If the receiver is credulous, the up-advisor can get him to make a higher assessment. Symmetric results are true for the down-advisor.

The discussion shows that credulity changes the levels of the highest and lowest actions that the advisor can induce. It does not change any of the qualitative features of the TTTE. This includes the assessment rule: In a TTTE with a $\psi-$credulous investor, his assessment depends only on θ^A_{post}, not on θ^A_{pre} or c^A. Introducing credulity does not make the auxiliary parameters relevant for persuasion.

If $\nu(\bar{\theta}) = \bar{\theta}$, the investor is *cursed*. In the game above, advisors are defined by their type. The type consists of their incentives and their knowledge about the true DGP. A cursed investor fails to anticipate that the advisor's message is based on that type. We investigate cursed equilibria of our game.[50]

We will now show that cursedness leads to $\nu(\bar{\theta}) = \bar{\theta}$. When a cursed investor neglects that the message contains information about the type, he will instead believe that the advisor randomly drew a message from an *average message profile*. This average message profile is equal to

$$\bar{\sigma}(\boldsymbol{m}) = \int_{\tau \in \mathcal{T}} \mu(\tau) \sigma(\boldsymbol{m}|\tau) \, \mathrm{d}\tau.$$

The investor can be partially cursed. We call an investor $\psi-$cursed if he believes that the advisor randomly draws a message from the average message profile $\bar{\sigma}(\boldsymbol{m})$ with probability $\psi \in [0,1]$ and otherwise believes that the advisor sends a message based on her type. A $\psi-$cursed investor's belief about what the advisor sends when she is τ is equal to

$$\sigma^\psi(\boldsymbol{m}|\tau) = \psi \bar{\sigma}(\boldsymbol{m}) + (1-\psi)\sigma(\boldsymbol{m}|\tau).$$

We can now think about what the investor learns about the advisor upon hearing $\boldsymbol{m}$. He will update his beliefs using $\psi-$cursed Bayes' rule whenever possible:

$$\mu^\psi(\tau|\boldsymbol{m}) = \frac{\sigma^\psi(\boldsymbol{m}|\tau)\mu(\tau)}{\int_{\tau \in \mathcal{T}} \sigma^\psi(\boldsymbol{m}|\tau)\mu(\tau) \, \mathrm{d}\tau}.$$

The $\psi-$cursed Bayes' rule can be simplified to

$$\mu^\psi(\tau|\boldsymbol{m}) = \psi \mu(\tau) + (1-\psi)\mu(\tau|\boldsymbol{m}).$$

The rule nests the case of an investor who fully accounts for the connection between types and

[50] Since ours is a sequential game, we adopt the Cursed Sequential Equilibrium (CSE) solution concept developed by Fong, Lin, and Palfrey (2023).Eyster and Rabin (2005) develop a cursed equilibrium concept for simultaneous games.

actions (who is ($\psi = 0$)—cursed) and that of an investor who does not account for the connection at all (who is ($\psi = 1$)—cursed).

In a ψ—cursed equilibrium, advisor and investor maximize their utility and the investor uses ψ—*cursed* Bayes' rule to update beliefs whenever possible.

Consider an ψ—cursed investor who hears $\boldsymbol{m}$ and suppose that $\boldsymbol{m}$ is on-path. The investor will update his belief about the advisor's incentives and knowledge ($\boldsymbol{\tau}$). His assessment will then be equal to his expectation of θ^T_{post} given this updated belief:

$$\begin{aligned}\mathbb{E}_{\mu^\psi}(\widetilde{\theta}^T_{post}|\boldsymbol{m}^A) &= \int_{\tau\in\mathcal{T}} \theta^T_{post}\mu^\psi(\boldsymbol{\tau}|\boldsymbol{m}^A)\,\mathrm{d}\boldsymbol{\tau} \\ &= \psi\bar{\theta} + (1-\psi)\mathbb{E}_\mu(\theta^T_{post}|\boldsymbol{m}^A).\end{aligned} \tag{9}$$

This is equal to the conditional expectation in Equation (7) when we use $\nu(\bar{\theta}) = \bar{\theta}$.

A ($\psi > 0$)—cursed investor makes less extreme assessments about θ^T_{post} than a ($\psi = 0$)—cursed investor. This happens because he learns less about the advisor upon hearing her message. This makes a TTE (including the TTTE) impossible. To illustrate, suppose that $\lambda = 0$, so that there are no intrinsicallly honest types. In a TTE, there are messages m^A which are not sent by the up- or down-advisor. Equation (9) shows that after receiving such a message the investor "shades" any assessment he should optimally make towards the prior expectation. If there now is an interval of assessments that the aligned advisor can induce (which is the case in a TTE), the advisor has an incentive to exaggerate. Instead of sending a message with θ^T_{post}, she will want to send a message with $\theta^A_{post} > \theta^T_{post}$ to overcome the investor's shading towards the prior. In equilibrium the investor, however, fully accounts for exaggeration. Therefore, an equilibrium where the investor can be induced to take any assessment on an *interval* does not exist.

Under cursedness, persuasive partition equilibria (Crawford and Sobel, 1982) can exist. In a partition equilibrium, the receiver can be induced to make any assessment in $\{\theta_1, \theta_2, \ldots, \theta_N\}$, where θ_n increases in n. Because of their incentives, the up-advisor induces θ_N and the down-advisor θ_1. The aligned advisor induces the θ_n that is closest to θ^T_{post}.

As an intermediate case between credulity and cursedness, **a function of the form $\nu(\bar{\theta}, \theta^A_{post}) = \alpha\bar{\theta} + (1-\alpha)\theta^A_{post}$ with $\alpha \in [0,1]$ describes an investor who *shades*.** Such an investor always pulls his assessment of θ^T_{post} towards his prior expectation. The discussion related to cursedness broadly applies here. Whenever $\alpha > 0$, so that there is *some* shading towards the prior, a TTE does not exist. The only persuasive equilibria are of the partition type.

In a partition equilibrium, the presence of auxiliary parameters gives rise to a coordination problem. We can take any partition equilibrium and swap each advisor's strategy over, say, θ^A_{post} and θ^A_{pre}. Combining these re-labeled strategies with a properly adjusted assessment rule constitutes an outcome-equivalent equilibrium. Therefore, coordination becomes harder. The presence of auxiliary parameters does not, however, change the distribution over payoffs that players receive. In this sense, auxiliary parameters remain irrelevant for persuasion.

In an equilibrium in which the investor is not fully rational, the advisor is good at anticipating and using these biases to her advantage. The equilibrium adjustment that follows, however, tends to mute these effects. To illustrate, **suppose that the investor is *credulous* ($\nu = \theta^A_{post}$), but that $\psi(g(\boldsymbol{m}^A))$ is a function with $\psi'(g) > 0$ so that *credulity is fit-based.*** Under this assumption, a

TTTE can exist where the up-advisor's probability of sending $\boldsymbol{m}^A$ is equal to

$$\sigma(\boldsymbol{m}^A|\uparrow) = g(\boldsymbol{m}^A)\frac{2\lambda+1}{1-\lambda}\frac{\max\{\theta^A_{post}-\theta_H,0\}}{\theta_H-(1-\psi(g(\boldsymbol{m}^A)))\bar{\theta}-\psi(g(\boldsymbol{m}^A))\theta^A_{post}}$$

This is an equilibrium where a misaligned senders carefully tailor their messages to fit the data. One can see this in the equation above; σ increases in g, and more strongly so as $\psi'(g)$ increases. Therefore, a misaligned sender is more likely to send a message that, ex-ante, seems likely.

However, the investor makes assessments *as if* he does not react to fit. Whenever he receives a message with $\sigma(\boldsymbol{m}^A|\uparrow) > 0$, he makes assessment θ_H and whenever $\sigma(\boldsymbol{m}^A|\downarrow) > 0$ he makes assessment θ_L: While g might vary in the different messages potentially sent by the up-advisor, equilibrium requires that the up-advisor receives the same payoff (induces the same assessment) for any message she might send. By sending those with a higher g with higher probability, she reduces the plausibility of these messages until they seem as plausible as messages with a lower g. This shows how the equilibrium dynamics mute reactions to fit.

D.7 How can a Cheap Talk Framework Potentially Accommodate Sensitivity to Auxiliary Parameters?

The previous subsection showed that behavioral frictions on the investor side cannot accommodate sensitivity to auxiliary parameters or fit. One may rather look for frictions on the advisor side to provide a strategic rationale. We sketch out a possible strategic argument that could be explored.[51] A misaligned advisor's message strategy in the equilibria discussed above requires her to adjust her message on multiple dimensions. More often than not, she sends messages that *(i)* bias θ^A_{post} in a favored direction and *(ii)* appear plausible ex-ante (increase g). The misaligned advisor might find it difficult to construct messages with both targets in mind. If there are advisors who are differently skilled at tailoring their lies to the data, messages with a bad fit could signal bad tailoring skills (and thus lies).

To be more concrete, consider a game with a sender and a receiver. There is a two-dimensional state $\boldsymbol{\omega} = (\omega_1, \omega_2)$, with $\boldsymbol{\omega} \in \{0,1\}^2$. Nature draws the state from a density function $g(\omega_1, \omega_2)$, with $g(1,1) = g(0,0) = a > b = g(1,0) = g(0,1)$. The state components are positively correlated; $(1,1)$ and $(0,0)$ are more likely than $(1,0)$ and $(0,1)$. The sender observes the state and sends a message $\boldsymbol{m} \in \{0,1\}^2$ to the receiver. The receiver then makes assessment ω^R_2. The sender has utility function $u^S(\omega^R_2) = \omega^R_2$ and the receiver has utility function $u^R(\omega^R_2) = -(\omega_2 - \omega^R_2)^2$. This implies that the receiver wants to guess ω_2 accurately while the sender wants the receiver to guess as high as possible.

We introduce frictions on the sender side by assuming that she has a *tailoring level*-type. The sender is *unskilled* with probability λ. An unskilled sender cannot lie about any dimension of her message; she will always send $\boldsymbol{m} = \boldsymbol{\omega}$. The sender is an *apprentice* with probability α. An apprentice can lie only about the decision-relevant ω_2 but cannot yet adjust ω_1 away from the

[51] To the best of our knowledge, the literature has not explored this model. The model is closely related to narrative approaches to communication since it explores a case where some senders find it impossible to adjust the auxiliary parameters of the messages they send away from the truth. This is related to the idea that crafting convincing narratives is a skill that not everyone may have. Such heterogeneity can provide the receiver with a reason to rely less on messages which appear unconvincing given commonly known data (his prior belief), since these messages might be poorly crafted (false) narratives. True messages do not suffer this as often as true messages typically are coherent sets of payoff-relevant and auxiliary parameters.

truth. Therefore, an apprentice always sends $\boldsymbol{m} = (\omega_1, m_2)$ and can only adjust m_2. The sender is a *master* with probability $1-\lambda-\alpha$. A master can adjust ω_1 and ω_2 away from the truth.

This game can have an equilibrium in which the unskilled sender always tells the truth, the apprentice sends $(1,1)$ if the true state is $(1,0)$ or $(1,1)$, sends $(0,1)$ otherwise and where the master always sends $(1,1)$. In this equilibrium, the receiver's optimal assessment after the different messages is equal to

$$\omega_2^R(0,0) = \omega_2^R(1,0) = 0 < \omega_2^R(0,1) = \frac{(\lambda+\alpha)b}{(\lambda+\alpha)b+\alpha a} < \omega_2^R(1,1) = \frac{a+(1-\lambda)b}{a+(1-\lambda)b+(1-\lambda-\alpha)(a+b)}.$$

The literal meaning of the messages $(0,1)$ and $(1,1)$ is the same. They both imply that $\omega_2 = 1$. However, $(1,1)$ is more plausible than $(0,1)$ ex-ante because the states of the world are positively correlated. It remains more plausible ex-post because the apprentice type is not good at tailoring her lie. The receiver knows this and makes a higher assessment after receiving $(1,1)$ than after receiving $(0,1)$. If the probability of being an apprentice is zero ($\alpha = 0$), then $\omega_2^R(0,1) = \omega_2^R(1,1)$ in any informative equilibrium. This illustrates how heterogeneity in "tailoring skill" can provide a strategic rationale for why message-receivers account for auxiliary parameters when evaluating different messages.

In the equilibrium above, the sender is informed about the state of the world. If she were instead uninformed, no informative equilibrium would exist. So the argument above does not yet explain why the receiver should listen to the sender in such a setting. One way to explain such behavior in an equilibrium framework would weaken the assumption that the receiver can infer the Bayesian belief from the data set (the signal). Different players may instead draw different conclusions from a data set that may become more heterogenous as the data set becomes more complex. One way to think about this is to think of the data set as sending a signal which has a common and an idiosyncratic part. Then, there is scope for persuasion as the receiver can potentially use the sender's message to learn about the idiosyncratic part of her signal.

E Omitted Proofs

E.1 Proof of Proposition 3

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

E.1.1 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 param-

eters 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

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

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

$$\Delta\ell'(\theta_{post}) = \frac{\Delta k}{\theta_{post}} - \frac{\Delta f}{1-\theta_{post}}. \tag{10}$$

Furthermore, as θ_{post} becomes large,

$$\begin{aligned}\lim_{\theta_{post}\to 1} \Delta\ell(\theta_{post}) = & \Delta k(\lim_{\theta_{post}\to 1} ln(\theta_{post}) - ln(\theta'_{pre})) + \Delta f(\lim_{\theta_{post}\to 1} ln(1-\theta_{post}) - ln(1-\theta'_{pre})) + \kappa \\ = & -\Delta k ln(\theta'_{pre}) + \Delta f(\lim_{\theta_{post}\to 1} ln(1-\theta_{post}) - ln(1-\theta'_{pre})) + \kappa\end{aligned} \tag{11}$$

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

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

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 (11) and the discussion afterwards). Furthermore, the derivative in Equation (10) 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 (11) and the discussion afterwards). Furthermore, the derivative in Equation (10) 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 (11) and the discussion afterwards). Furthermore, the derivative in Equation (10) 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

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

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.1.2 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 (11) and the discussion afterwards). Furthermore, the derivative in Equation (10) 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 (11) and the discussion afterwards). Furthermore, the derivative in Equation (10) 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 (11) and the discussion afterwards). Furthermore, the derivative in Equation (10) 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 $\Delta f < 0$.

E.2 Proof of Proposition 4

To show that θ_H exists and is unique, note that, for all $\boldsymbol{m}^A \in \mathcal{M}^{max}$,

$$\theta_H = \mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\boldsymbol{m}^A) = \mathbb{E}(\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\boldsymbol{m}^A)|\boldsymbol{m}^A \in \mathcal{M}^{max}).$$

By the law of iterated expectations, this is equal to

$$\mathbb{E}(\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\boldsymbol{m}^A)|\boldsymbol{m}^A \in \mathcal{M}^{max}) = \mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\boldsymbol{m}^A \in \mathcal{M}^{max}).$$

Now note that the likelihood of a truth-telling advisor sending $\boldsymbol{m}^A \in \mathcal{M}^{max}$ is equal to $1-G_{\theta_{post}}(\theta_H)$ while that of a strategic up-advisor is equal to 1. Therefore,

$$\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\boldsymbol{m}^A \in \mathcal{M}^{max}) = q(\theta_H)\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \theta_H) + (1-q(\theta_H))\bar{\theta},$$
$$\text{where } q(\theta_H) = \frac{(2\lambda+1)(1-G_{\theta_{post}}(\theta_H))}{(2\lambda+1)(1-G_{\theta_{post}}(\theta_H))+(1-\lambda)}.$$

Now define a function

$$\hat{\theta}_H(\theta_H) \equiv q(\theta_H)\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \theta_H) + (1-q(\theta_H))\bar{\theta}.$$

Since this function maps from $[0,1]$ to $[0,1]$, it has at least one fixed point θ^*_H where $\hat{\theta}_H(\theta^*_H) = \theta^*_H$. Evaluating the function at $\bar{\theta}$ and 1 yields

$$\hat{\theta}_H(\bar{\theta}) > \bar{\theta} \text{ and } \hat{\theta}_H(1) = \bar{\theta} < 1.$$

This indicates that there is at least one fixed point on $(\bar{\theta}, 1)$. To show that it is unique, take the derivative

$$\hat{\theta}'_H(\theta_H) = q'(\theta_H)(\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \theta_H) - \bar{\theta}) + q(\theta_H)\frac{\partial \mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \theta_H)}{\partial \theta_H}.$$

Since $q'(\theta_H) < 0$, $\hat{\theta}'_H(\theta_H) < 1$ if $\frac{\partial \mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \theta_H)}{\partial \theta_H} < 1$. This is the case, as $g_{\theta_{post}}(\theta_H)$ is a mixture distribution of different beta distributions: For a mixture distribution where the conditional expectations of the individual components are $\mathbb{E}_1(\theta^T_{post}|\theta^T_{post} \geq \theta_H)$, $\mathbb{E}_2(\theta^T_{post}|\theta^T_{post} \geq \theta_H)$, …, and where the density functions are weighted by weights w_1, w_2, … which sum up to one we have

$$\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \theta_H) = \sum_i w_i \mathbb{E}_i(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \theta_H) \Rightarrow \frac{\partial \mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \theta_H)}{\partial \theta_H} = \sum_i w_i \frac{\partial \mathbb{E}_i(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \theta_H)}{\partial \theta_H}.$$

As the beta distribution belongs to the family of log-concave distributions, $\frac{\partial \mathbb{E}_i(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \theta_H)}{\partial \theta_H} < 1$,[52] which implies that $\frac{\partial \mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \theta_H)}{\partial \theta_H} < 1$. Therefore, $\hat{\theta}'_H(\theta_H) < 1$, which, together with $\hat{\theta}_H(\bar{\theta}) > \bar{\theta}$ and $\hat{\theta}_H(1) < 1$, implies that $\hat{\theta}_H(\theta_H)$ has a unique fixed point $\theta^*_H \in (\bar{\theta}, 1)$ where $\hat{\theta}_H(\theta^*_H) = \theta^*_H$.

Using an analogous argument, one can show that $\theta_L \in (0, \bar{\theta})$ exists and is unique. Therefore, the equilibrium exists and is unique.

E.3 Proof of Proposition 5

Denote the lowest and highest assessment induced in an equilibrium different from the TTTE by $\dot{\theta}_L$, $\dot{\theta}_H$, the set of messages that induce an assessment $\dot{\theta}_L$ by $\dot{\mathcal{M}}^{min}$ and the set of messages that

[52] See, e.g., Lemma 1 in Harbaugh and Rasmusen (2018).

induce an assessment $\dot{\theta}_H$ by $\dot{\mathcal{M}}^{max}$. In this case, for all $\boldsymbol{m}^A \in \dot{\mathcal{M}}^{max}$,

$$\dot{\theta}_H = \mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\boldsymbol{m}^A) = \mathbb{E}(\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\boldsymbol{m}^A)|\boldsymbol{m}^A \in \dot{\mathcal{M}}^{max}) = \mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\boldsymbol{m}^A \in \dot{\mathcal{M}}^{max}).$$

Lemma 1 suggests that in any equilibrium, there is a $\dot{\theta}^{\rightarrow}_H \leq \dot{\theta}_H$ such that the aligned advisor finds it optimal to induce $\dot{\theta}_H$ if $\theta^T_{post} \geq \dot{\theta}^{\rightarrow}_H$. Therefore, we can define a function which denotes the investor's assessment conditional on receiving a message $\boldsymbol{m} \in \dot{\mathcal{M}}^{max}$:

$$\begin{aligned}
&\hat{\theta}_H(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H) \equiv \dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}_H) + \dot{r}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}^{\rightarrow}_H) \qquad (12)\\
&+ (1 - \dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H) - \dot{r}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H))\bar{\theta},\\
&\text{where } \dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H) = \frac{3\lambda(1 - G_{\theta_{post}}(\dot{\theta}_H))}{3\lambda(1 - G_{\theta_{post}}(\dot{\theta}_H)) + (1-\lambda)(1 - G_{\theta_{post}}(\dot{\theta}^{\rightarrow}_H)) + (1-\lambda)},\\
&\text{and } \dot{r}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H) = \frac{(1-\lambda)(1 - G_{\theta_{post}}(\dot{\theta}^{\rightarrow}_H))}{3\lambda(1 - G_{\theta_{post}}(\dot{\theta}_H)) + (1-\lambda)(1 - G_{\theta_{post}}(\dot{\theta}^{\rightarrow}_H)) + (1-\lambda)}.
\end{aligned}$$

Claim 1: For any $\dot{\theta}^{\rightarrow}_H \leq \dot{\theta}_H$, there is a unique fixed point $\dot{\theta}^*_H$ such that $\hat{\theta}_H(\dot{\theta}^*_H, \dot{\theta}^{\rightarrow}_H) = \dot{\theta}^*_H$.

Taking the derivative with respect to $\dot{\theta}_H$;

$$\begin{aligned}
&\frac{\partial \hat{\theta}_H(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)}{\partial \dot{\theta}_H} = \frac{\partial \dot{q}}{\partial \dot{\theta}_H}\Big[\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}_H) - \bar{\theta}\Big] + \frac{\partial \dot{r}}{\partial \dot{\theta}_H}\Big[\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}^{\rightarrow}_H) - \bar{\theta}\Big]\\
&+ \dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)\frac{\partial \mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}_H)}{\partial \dot{\theta}_H},\\
&\text{where } \frac{\partial \dot{q}}{\partial \dot{\theta}_H} = -\frac{g_{\theta_{post}}(\dot{\theta}_H)}{1 - G_{\theta_{post}}(\dot{\theta}_H)}\dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)(1 - \dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)) \text{ and } \frac{\partial \dot{r}}{\partial \dot{\theta}_H} = \frac{g_{\theta_{post}}(\dot{\theta}_H)}{1 - G_{\theta_{post}}(\dot{\theta}_H)}\dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)\dot{r}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)
\end{aligned}$$

The sum of the first two derivative terms is negative if

$$\begin{aligned}
&\frac{\partial \dot{q}}{\partial \dot{\theta}_H}\Big[\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}_H) - \bar{\theta}\Big] + \frac{\partial \dot{r}}{\partial \dot{\theta}_H}\Big[\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}^{\rightarrow}_H) - \bar{\theta}\Big] < 0,\\
\Rightarrow &-(1 - \dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H))\Big[\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}_H) - \bar{\theta}\Big] + \dot{r}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)\Big[\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}^{\rightarrow}_H) - \bar{\theta}\Big] < 0.
\end{aligned}$$

Since $\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}_H) \geq \mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}^{\rightarrow}_H)$, a sufficient condition for the inequality to hold is that

$$1 - \dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H) > \dot{r}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H),$$

which holds as $1 - \dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H) - \dot{r}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H) > 0$. Therefore, the sum of the two first terms in the derivative is negative. We further know from the proof of Proposition 4 that $\frac{\partial \mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}_H)}{\partial \dot{\theta}_H} < 1$. Therefore, $\frac{\partial \hat{\theta}_H(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)}{\partial \dot{\theta}_H} < 1$, which suggests a unique fixed point $\dot{\theta}^*_H$ where $\hat{\theta}_H(\dot{\theta}^*_H, \theta^{\rightarrow}_H) = \dot{\theta}^*_H$ for any $\theta^{\rightarrow}_H \leq \dot{\theta}^*_H$.

Claim 2: The fixed point $\dot{\theta}_H^*$ increases in $\dot{\theta}_H^{\rightarrow}$. The fixed point solves

$$h(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow}) = \dot{q}(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow})\mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H) + \dot{r}(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow})\mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^{\rightarrow}) \\ + (1 - \dot{q}(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow}) - \dot{r}(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow}))\bar{\theta} - \dot{\theta}_H \equiv 0. \tag{13}$$

Using the implicit function theorem, the derivative of $\dot{\theta}_H^*$ with respect to $\theta_H^{\rightarrow}$ is given by

$$\frac{\partial \dot{\theta}_H^*}{\partial \theta_H^{\rightarrow}} = \frac{\frac{\partial h(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow})}{\partial \dot{\theta}_H^{\rightarrow}}}{-\frac{\partial h(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow})}{\partial \dot{\theta}_H}}.$$

The results from Claim 1 now suggest that the denominator above is positive. Therefore, $\frac{\partial \dot{\theta}_H^*}{\partial \theta_H^{\rightarrow}}$ is positive if and only if the numerator is positive. This is the case if

$$\frac{\partial h(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow})}{\partial \dot{\theta}_H^{\rightarrow}} = \frac{\partial \dot{q}}{\partial \dot{\theta}_H^{\rightarrow}}\Big[\mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H) - \bar{\theta}\Big] + \frac{\partial \dot{r}}{\partial \dot{\theta}_H^{\rightarrow}}\Big[\mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^{\rightarrow}) - \bar{\theta}\Big] \\ + \dot{r}(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow})\frac{\partial \mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^{\rightarrow})}{\partial \dot{\theta}_H^{\rightarrow}} \geq 0,$$

where $\frac{\partial \dot{q}}{\partial \dot{\theta}_H^{\rightarrow}} = \frac{g_{\theta_{post}}(\dot{\theta}_H^{\rightarrow})}{1 - G_{\theta_{post}}(\dot{\theta}_H^{\rightarrow})}\dot{q}(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow})\dot{r}(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow})$, $\frac{\partial \dot{r}}{\partial \dot{\theta}_H^{\rightarrow}} = -\frac{g_{\theta_{post}}(\dot{\theta}_H^{\rightarrow})}{1 - G_{\theta_{post}}(\dot{\theta}_H^{\rightarrow})}\dot{r}(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow})(1 - \dot{r}(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow}))$,

and $\frac{\partial \mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^{\rightarrow})}{\partial \dot{\theta}_H^{\rightarrow}} = \frac{g_{\theta_{post}}(\dot{\theta}_H^{\rightarrow})}{1 - G_{\theta_{post}}(\dot{\theta}_H^{\rightarrow})}\Big[\mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^{\rightarrow}) - \dot{\theta}_H^{\rightarrow}\Big]$.

Plugging in and simplifying the inequality above yields

$$\dot{q}(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow})\Big[\mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H) - \bar{\theta}\Big] - (1 - \dot{r}(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow}))\Big[\mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^{\rightarrow}) - \bar{\theta}\Big] \\ + \mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^{\rightarrow}) - \dot{\theta}_H^{\rightarrow} \geq 0.$$

Note that we can rearrange Equation (12) to

$$\dot{q}(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow})\Big[\mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H) - \bar{\theta}\Big] = \dot{\theta}_H - \bar{\theta} - \dot{r}(\dot{\theta}_H, \dot{\theta}_H^{\rightarrow})\Big[\mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^{\rightarrow}) - \bar{\theta}\Big],$$

which we can use to simplify the inequality to

$$\dot{\theta}_H - \bar{\theta} - \Big[\mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^{\rightarrow}) - \bar{\theta}\Big] + \mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^{\rightarrow}) - \dot{\theta}_H^{\rightarrow} \geq 0 \\ \Rightarrow \dot{\theta}_H - \dot{\theta}_H^{\rightarrow} \geq 0.$$

Therefore, $\frac{\partial \dot{\theta}_H^*}{\partial \theta_H^{\rightarrow}} \geq 0$ if $\dot{\theta}_H^{\rightarrow} \leq \dot{\theta}_H^*$.

Claim 3: For any $\dot{\theta}_L^{\rightarrow} \geq \dot{\theta}_L$, there is a unique fixed point $\dot{\theta}_L^*$ such that $\hat{\theta}_L(\dot{\theta}_L^*, \dot{\theta}_L^{\rightarrow}) = \dot{\theta}_L^*$. This follows from analogous arguments as in Claim 1.

Claim 4: The fixed point $\dot{\theta}_L^*$ increases in $\dot{\theta}_L^{\rightarrow}$. This follows from analogous arguments as in Claim 2.

Claim 5: The TTTE is a most informative equilibrium. Consider any equilibrium of the game. As

before, use $\dot{\theta}_H^{\rightarrow}$ to denote the threshold value such that the aligned advisor sends a message in $\mathcal{M}^{max}$ if $\theta_{post}^T \geq \dot{\theta}_H^{\rightarrow}$. Similarly, use $\dot{\theta}_L^{\rightarrow}$ to denote the threshold value such that the aligned advisor sends a message in $\mathcal{M}^{min}$ if $\theta_{post}^T \leq \dot{\theta}_L^{\rightarrow}$. Given these thresholds, denote the maximum assessment that the investor can be induced to make by $\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow})$ as defined in Equation (13) and the minimum assessment that the investor can be induced to make by $\dot{\theta}_L^*(\dot{\theta}_L^{\rightarrow})$. Note that both of these thresholds are unique and functions of $\dot{\theta}_H^{\rightarrow}$ and $\dot{\theta}_L^{\rightarrow}$. In such an equilibrium, the expected squared error of the assessment conditional on meeting the up-advisor is given by

$$\int_0^1 \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post}) \, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post}$$

and the error conditional on meeting the down-advisor is

$$\int_0^1 \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_L^*(\dot{\theta}_L^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post}) \, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post}.$$

Conditional on meeting the honest advisor, the expected squared error is given by

$$\begin{aligned}
&\int_{\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow})}^1 \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post}) \, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \\
&+ \int_0^{\dot{\theta}_L^*(\dot{\theta}_L^{\rightarrow})} \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_L^*(\dot{\theta}_L^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post}) \, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \\
&+ \int_{\dot{\theta}_L^*(\dot{\theta}_L^{\rightarrow})}^{\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow})} \int_0^1 \sum_{c=2}^{8} (\theta_{post}^I(c, \theta_{pre}, \theta_{post}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post}) \, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \\
&\geq \int_{\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow})}^1 \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post}) \, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \\
&+ \int_0^{\dot{\theta}_L^*(\dot{\theta}_L^{\rightarrow})} \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_L^*(\dot{\theta}_L^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post}) \, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post}.
\end{aligned}$$

Finally, the expected squared error conditional on meeting the aligned advisor is given by

$$\begin{aligned}
&\int_{\dot{\theta}_H^{\rightarrow}}^1 \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post}) \, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \\
&+ \int_0^{\dot{\theta}_L^{\rightarrow}} \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_L^*(\dot{\theta}_L^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post}) \, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \\
&+ \int_{\dot{\theta}_L^{\rightarrow}}^{\dot{\theta}_H^{\rightarrow}} \int_0^1 \sum_{c=2}^{8} (\theta_{post}^I(c, \theta_{pre}, \theta_{post}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post}) \, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \\
&\geq \int_{\dot{\theta}_H^{\rightarrow}}^1 \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post}) \, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \\
&+ \int_0^{\dot{\theta}_L^{\rightarrow}} \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_L^*(\dot{\theta}_L^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post}) \, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post}.
\end{aligned}$$

Combining the various errors, we can define a function $\tilde{L}(\dot{\theta}_H^{\rightarrow}, \dot{\theta}_L^{\rightarrow})$ which provides a lower bound for the expected squared error in any equilibrium:

$$\begin{aligned}
\tilde{L}(\dot{\theta}_H^{\rightarrow}, \dot{\theta}_L^{\rightarrow}) \equiv \frac{1-\lambda}{3}\Big[& \int_0^1 \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post})\, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} + \\
& \int_0^1 \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_L^*(\dot{\theta}_L^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post})\, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \Big] \\
& + \lambda \Big[\int_{\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow})}^1 \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post})\, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \\
& + \int_0^{\dot{\theta}_L^*(\dot{\theta}_L^{\rightarrow})} \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_L^*(\dot{\theta}_L^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post})\, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \Big] \\
& + \frac{1-\lambda}{3}\Big[\int_{\dot{\theta}_H^{\rightarrow}}^1 \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post})\, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \\
& + \int_0^{\dot{\theta}_L^{\rightarrow}} \int_0^1 \sum_{c=2}^{8} (\dot{\theta}_L^*(\dot{\theta}_L^{\rightarrow}) - \theta_{post})^2 g(c, \theta_{pre}, \theta_{post})\, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \Big].
\end{aligned}$$

Taking the derivative with respect to $\dot{\theta}_H^{\rightarrow}$ brings

$$\begin{aligned}
\frac{\partial \tilde{L}}{\partial \dot{\theta}_H^{\rightarrow}} = \dot{\theta}_H^{*'}(\dot{\theta}_H^{\rightarrow}) \Big\{ & \frac{1-\lambda}{3}\Big[\int_0^1 \int_0^1 \sum_{c=2}^{8} 2(\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow}) - \theta_{post}) g(c, \theta_{pre}, \theta_{post})\, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \Big] \\
& + \lambda \Big[\int_{\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow})}^1 \int_0^1 \sum_{c=2}^{8} 2(\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow}) - \theta_{post}) g(c, \theta_{pre}, \theta_{post})\, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \Big] \\
& + \frac{1-\lambda}{3}\Big[\int_{\dot{\theta}_H^{\rightarrow}}^1 \int_0^1 \sum_{c=2}^{8} 2(\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow}) - \theta_{post}) g(c, \theta_{pre}, \theta_{post})\, \mathrm{d}\theta_{pre} \mathrm{d}\theta_{post} \Big] \Big\} \\
& - \frac{1-\lambda}{3} \int_0^1 \sum_{c=2}^{8} 2(\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow}) - \dot{\theta}_H^{\rightarrow}) g(c, \theta_{pre}, \dot{\theta}_H^{\rightarrow})\, \mathrm{d}\theta_{pre}.
\end{aligned}$$

Consider the term in curly brackets, which we can rewrite as (we write $\dot{\theta}_H^*$ instead of $\dot{\theta}_H^*(\dot{\theta}_H^{\rightarrow})$ for ease of notation)

$$\begin{aligned}
& = 2\Big[\frac{1-\lambda}{3}(\dot{\theta}_H^* - \bar{\theta}) + \lambda(1 - G_{\theta_{post}}(\dot{\theta}_H^*))(\dot{\theta}_H^* - \mathbb{E}[\theta_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^*]) \\
& + \frac{1-\lambda}{3}(1 - G_{\theta_{post}}(\dot{\theta}_H^{\rightarrow}))(\dot{\theta}_H^* - \mathbb{E}[\theta_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^{\rightarrow}]) \Big] \\
& = 2\Big[\Big(\frac{1-\lambda}{3} + \lambda(1 - G_{\theta_{post}}(\dot{\theta}_H^*) + \frac{1-\lambda}{3}(1 - G_{\theta_{post}}(\dot{\theta}_H^{\rightarrow})) \Big) \dot{\theta}_H^* \\
& - \frac{1-\lambda}{3}\bar{\theta} - \lambda(1 - G_{\theta_{post}}(\dot{\theta}_H^*)) \mathbb{E}[\theta_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^*] - \frac{1-\lambda}{3}(1 - G_{\theta_{post}}(\dot{\theta}_H^{\rightarrow})) \mathbb{E}[\theta_{post}^T | \theta_{post}^T \geq \dot{\theta}_H^{\rightarrow}] \Big] \\
& = 2\Big(\frac{1-\lambda}{3} + (\lambda + \frac{1-\lambda}{3})(1 - G_{\theta_{post}}(\dot{\theta}_H^*)) \Big) \Big[\dot{\theta}_H^* - \mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \boldsymbol{m}^A \in \dot{\mathcal{M}}_{max}) \Big].
\end{aligned}$$

Since, in equilibrium, $\dot{\theta}_H^* = \mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \boldsymbol{m}^A \in \dot{\mathcal{M}}_{max})$, this term is equal to zero. Therefore, the

derivative simplifies to

$$\frac{\partial \tilde{L}}{\partial \dot{\theta}_H^{\rightarrow}} = -\frac{1-\lambda}{3} \int_0^1 \sum_{c=2}^{8} 2(\dot{\theta}_H^* - \dot{\theta}_H^{\rightarrow}) g(c, \theta_{pre}, \dot{\theta}_H^{\rightarrow}) \, \mathrm{d}\theta_{pre}.$$

This term is negative whenever $\dot{\theta}_H^* > \dot{\theta}_H^{\rightarrow}$. Therefore, the expected squared error is minimized when $\dot{\theta}_H^* = \dot{\theta}_H^{\rightarrow}$. A similar argument shows that the expected squared error is minimized when $\dot{\theta}_L^* = \dot{\theta}_L^{\rightarrow}$. Since these conditions hold in the TTTE, there is no equilibrium with a lower expected squared error than the TTTE. We conclude that the TTTE is a most informative equilibrium.

Claim 6: Any most informative equilibrium is a TTE characterized by the same thresholds θ_L and θ_H that characterize the TTTE. Claim 5 suggests that $\dot{\theta}_H^* = \dot{\theta}_H^{\rightarrow}$ and $\dot{\theta}_L^* = \dot{\theta}_L^{\rightarrow}$ in any most informative equilibrium. Claims 1 and 3 suggest that $\dot{\theta}_H^*$ and $\dot{\theta}_L^*$ are unique and so any most informative equilibrium has the same thresholds as the TTTE. Finally, note that, in the TTTE, the aligned/honest advisor induce θ_{post}^T if $\theta_{post}^T \in (\theta_L, \theta_H)$. Therefore, they must also be able to induce θ_{post}^T if $\theta_{post}^T \in (\theta_L, \theta_H)$ in any most informative equilibrium, as the expected error would otherwise be larger. Therefore, any most informative equilibrium is a TTE characterized by the same thresholds θ_L and θ_H that characterize the TTTE.

E.4 Proof of Proposition 6

Proposition 5 suggests that any most informative equilibrium is a TTE. We will show that the properties above hold in any most informative TTE.

Claim 1: In any most informative equilibrium, the aligned advisor sends $\theta_{post}^A = \theta_{post}^T$ if $\theta_{post}^T \in (\theta_L, \theta_H)$. Corollary 1 suggests that in any TTE, the aligned advisor induces the investor to make assessment θ_{post}^T if $\theta_{post}^T \in (\theta_L, \theta_H)$. Suppose by contradiction that in a most informative equilibrium, the aligned advisor sends a message $\boldsymbol{m}' = (c', \theta'_{pre}, \theta'_{post})$ with $\theta'_{post} \neq \theta_{post}^T \in (\theta_L, \theta_H)$ with positive probability. There can be three cases. First and second, $\boldsymbol{m}'$ can either induce θ_L or θ_H; this leads to an immediate contradiction with Corollary 1 since $\theta_{post}^T \in (\theta_L, \theta_H)$ but $\theta_{post}^I \neq \theta_{post}^T$. Third, if $\boldsymbol{m}'$ does not induce θ_L or θ_H, then Corollary 1 suggests that $\boldsymbol{m}'$ is only sent by the honest/aligned advisor with positive probability in equilibrium. Therefore, the advisor's optimal assessment after receiving $\boldsymbol{m}'$ is equal to

$$\mathbb{E}_\mu(\widetilde{\theta}_{post}^T | \boldsymbol{m}') = \frac{3\lambda}{3\lambda + (1-\lambda)} \theta'_{post} + \frac{(1-\lambda)}{3\lambda + (1-\lambda)} \theta_{post}^T \neq \theta_{post}^T,$$

which also contradicts Corollary 1. Therefore, the aligned advisor sends $\theta_{post}^A = \theta_{post}^T$ if $\theta_{post}^T \in (\theta_L, \theta_H)$.

Claim 2: In any most informative equilibrium, the misaligned advisors do not send messages with $\theta_{post}^A \in (\theta_L, \theta_H)$. Since $\theta_H > \theta_L$ in any TTE, Corollary 1 implies that the up- and down-advisor never send the same message with positive probability in equilibrium since they induce different actions. Consider the up-advisor and suppose by contradiction that in a most informative equilibrium, the up-advisor sends a message $\boldsymbol{m}' = (c', \theta'_{pre}, \theta'_{post})$ with $\theta_{post}^A \in (\theta_L, \theta_H)$ with positive probability. Consider the case where $\boldsymbol{m}^T = \boldsymbol{m}'$ and an honest advisor who sends $\boldsymbol{m}'$. In the most informative TTE, whenever $\theta_{post}^T \in (\theta_L, \theta_H)$, the honest/aligned advisor induce assessment θ_{post}^T. This suggests that, upon receiving $\boldsymbol{m}'$, the investor makes assessment $\theta'_{post} < \theta_H$. But then it is no

longer optimal for the up-advisor to send $\boldsymbol{m}'$ with positive probability, which is a contradiction. Therefore, the up-advisor does not send messages with $\theta^A_{post} \in (\theta_L, \theta_H)$. A similar argument can be made for the down-advisor. Therefore, the misaligned advisors do not send messages with $\theta^A_{post} \in (\theta_L, \theta_H)$.

Claim 3: In any most informative equilibrium, the up-advisor only sends messages with $\theta^A_{post} \geq \theta_H$ and the down-advisor only sends messages with $\theta^A_{post} \leq \theta_L$. Since $\theta_H > \theta_L$ in any TTE, Corollary 1 implies that the up- and down-advisor never send the same message with positive probability in equilibrium since they induce different actions. Claim 2 now suggests that one case we need to rule is the case where the up-advisor sends a message $\boldsymbol{m}'$ with $\theta'_{post} \leq \theta_H$ with positive probability that is not sent by the down-advisor with positive probability. Consider a most informative equilibrium where this is the case and note that in this equilibrium, $\boldsymbol{m}'$ induces action θ_H and is sent with positive probability by the up-advisor, the honest advisor if $\boldsymbol{m}^T = \boldsymbol{m}'$, or the aligned advisor if $\theta^T_{post} \geq \theta_H$. We can always perturb this equilibrium by reducing the probability that the up-advisor or the aligned advisor send $\boldsymbol{m}'$ after observing $\theta^T_{post} \geq \theta_H$ to zero and increasing the probability that the down-advisor sends $\boldsymbol{m}'$ by such an amount that it becomes optimal for the investor to make assessment θ_L after hearing $\boldsymbol{m}'$. The resulting thresholds of the new equilibrium will now be more extreme than those of the original equilibrium (i.e., θ_H increases and θ_L decreases), which contradicts the fact that the original equilibrium is most informative. Therefore, the up-advisor only sends messages with $\theta^A_{post} \geq \theta_H$. We can use a similar argument to show that the down-advisor only sends messages with $\theta^A_{post} \leq \theta_L$.

Claim 4: In any most informative equilibrium, the aligned investor sends $\theta^A_{post} \geq \theta_H$ if $\theta^T_{post} \geq \theta_H$ and $\theta^A_{post} \leq \theta_L$ if $\theta^T_{post} \leq \theta_H$. Corollary 1 suggests that the aligned investor induces θ_H whenever $\theta^T_{post} \geq \theta_H$. In the most informative equilibrium, the aligned advisor's message must pool in that case with the up-advisor's message. Claim 3 suggests that the up-advisor only sends messages with $\theta^A_{post} \geq \theta_H$. Therefore, the aligned investor sends $\theta^A_{post} \geq \theta_H$ if $\theta^T_{post} \geq \theta_H$. A similar argument can be made for the case where $\theta^T_{post} \leq \theta_L$.

Claim 5: In any most informative equilibrium, the investor's assessment is equal to θ^A_{post} if $\theta^A_{post} \in (\theta_L, \theta_H)$. This follows from Claims 1, 2, and 3.

Claim 6: In any most informative equilibrium, the investor's assessment is equal to θ_H if $\theta^A_{post} \geq \theta_H$. Suppose by contradiction that there is a message $\boldsymbol{m}'$ with a $\theta'_{post} \geq \theta_H$ such that the investor's assessment is strictly smaller than θ_H. The structure of the TTE then suggests that the investor's assessment upon receiving $\boldsymbol{m}'$ is smaller than θ_H. Therefore, by utility maximization, the up-advisor does not send $\boldsymbol{m}'$ with positive probability and the aligned advisor does not send $\boldsymbol{m}'$ if $\theta^T_{post} \geq \theta_H$. Claims 1 and 4 further suggest that the aligned advisor also does not send $\boldsymbol{m}'$ if $\theta^T_{post} < \theta_H$. However, the honest advisor sends $\boldsymbol{m}'$ with positive probability. Therefore, the investor's assessment upon receiving $\boldsymbol{m}'$ is equal to θ'_{post}. This leads to a contradiction if $\theta'_{post} > \theta_H$. Therefore, the investor's assessment is equal to θ_H if $\theta^A_{post} \geq \theta_H$.

Claim 7: In any most informative equilibrium, the investor's assessment is equal to θ_L if $\theta^A_{post} \leq \theta_L$. This follows from analogous arguments as made in Claim 6.

E.5 Proof of Proposition 7

Consider any equilibrium in which the aligned advisor follows an honest strategy and denote the maximum assessment by $\dot{\theta}_H$ and the set of messages that induce the maximum assessment by $\dot{\mathcal{M}}^{max}$. It follows that

$$\dot{\theta}_H = \mathbb{E}_\mu(\widetilde{\theta}_{post}|\boldsymbol{m}^A) = \mathbb{E}_\mu(\widetilde{\theta}_{post}|\boldsymbol{m}^{\boldsymbol{A}} \in \dot{\mathcal{M}}^{max}).$$

Now define a function

$$\hat{\theta}_H(\dot{\theta}_H) \equiv q(\dot{\theta}_H)\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}_H) + (1 - q(\dot{\theta}_H))\bar{\theta},$$
$$\text{where } q(\dot{\theta}_H) = \frac{(2\lambda+1)(1 - G_{\theta_{post}}(\dot{\theta}_H))}{(2\lambda+1)(1 - G_{\theta_{post}}(\dot{\theta}_H)) + (1-\lambda)}.$$

Using analogous arguments as used in the proof of Proposition 4 shows that this function has a unique fixed point. Therefore, $\dot{\theta}_H$ is unique (and a similarly defined $\dot{\theta}_L$ is also unique). Both thresholds are also identical to the thresholds of the TTTE. It then follows that any equilibrium in which the aligned advisor follows an honest strategy is as informative as the TTTE and, therefore, most informative.