Narrative Persuasion*

Kai Barron
WZB Berlin

Tilman Fries
LMU Munich

September 18, 2025

For the current version, click here.

Abstract

We study how one person may shape how another interprets objective information 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: 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 tailor their narratives to the facts. Fourth, competing narratives predictably influence both narrative construction and adoption. Finally, we also analyze belief updating and the limits of persuasion in this setting.

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, AEARCTR-0011565 and AEARCTR-0015403.

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 (2020) 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. They often share narratives using simple stories, metaphors, or anecdotes via word-of-mouth or on social media. Because narratives may also be used as a tool for persuasion, it is crucial to understand how they are communicated and what determines their adoption. Yet, empirical work is scarce. One reason for this is that it is challenging to study the transmission of narratives in field settings. For example, narratives 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].

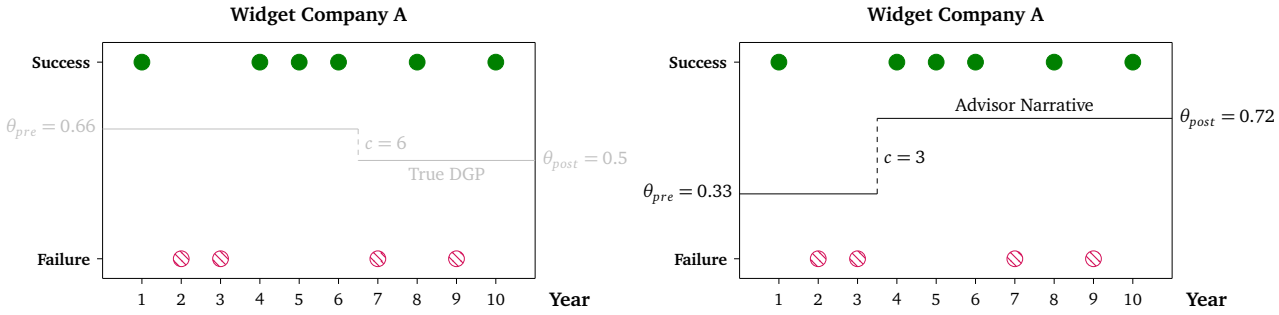
Our study makes five primary contributions to the literature. First, we show that humans are susceptible to narrative persuasion. Narratives are persuasive even if investors know the narrative is constructed by an advisor *(i)* who can tailor it to the public information about the company *ex-post*, *(ii)* who has no private information, *(iii)* who may have 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. Fifth, we examine more closely how investors update their beliefs in response to narratives. We show that, when

presented with a new candidate narrative, the most common reaction by investors is either to fully adopt the advisor's narrative or to stick with their prior beliefs. We also find that persuasion can be attenuated: disclosing the advisor's incentives and providing warnings make investors more skeptical and less willing to adopt the advisor's narrative. Finally, we introduce a versatile experimental framework that can be used as a workhorse for future inquiry. In addition, 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 performance data from a company. This data 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 (θ)—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 fully 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's goal is to estimate the company's probability of success under the current CEO. Therefore, he only cares directly about θ_{post}. The advisor, who may have incentives to bias the investor's beliefs, proposes a narrative—an explanation of the company's performance during the period covered by the historical data. This narrative may shape how the investor interprets the data and influence the beliefs he forms about the company's future performance.

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

Narratives mirror the structure of the DGP. Each narrative consists of a proposal for the three

[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.

parameters: the company's probability of success under the previous CEO, the success probability under the current CEO, and the year in which the CEO changed. The narrative 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 on two key sources of variation:

- **Advisor incentives:** We randomly assign advisors to one of three types—*up, down,* or *aligned. Up-advisors* are incentivized to persuade investors that the company's current probability of success is high, *down-advisors* are incentivized to persuade investors that it is low, and *aligned advisors* are incentivized to induce accurate investor beliefs.

- **Advisor knowledge:** We consider two information scenarios that vary the advisor's knowledge (i.e., their degree of expertise). In the Symmetric scenario, the advisor and investor have identical information: both observe only the historical company performance data. In the Asymmetric scenario, the advisor additionally learns the parameter values of the true data-generating process.

Figure 1 illustrates the core intuition of narrative persuasion in our experiment. The black line in Panel (ii) visualizes a potential narrative that an up-advisor might use to try to persuade an investor. This example highlights a central feature of narrative persuasion. Although the advisor only cares about moving the investor's belief about θ_{post} upwards, 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. 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. As a result, she shifts her assessment of the company's probability of success under the previous CEO downward to improve the fit of her narrative to the data. This highlights a central idea: because the narrative is constrained by the data, the advisor must navigate a trade-off between movement (of their claim) and fit.

With the intuition of our experimental design in place, we now describe our main results in more detail. First, we find that narratives are persuasive: Investors who meet an up-advisor form more optimistic beliefs than those who meet a down-advisor. This holds in both the Asymmetric and Symmetric settings, showing that narratives are persuasive even when advisors have no private information.

Second, we find that a key determinant of a narrative's persuasiveness is its coherence with the public facts (as measured by the empirical fit). When advisors construct better-fitting narratives, investors move their beliefs closer to the advisor's claim. This relationship is causal: when the fit of a narrative is exogenously increased, investors are more likely to believe the narrative.

Third, advisors anticipate the fit-movement tradeoff; we demonstrate that their narratives display a robust *negative correlation* between θ_{post} and θ_{pre}, as the theory predicts. Fourth, the presence of a competing narrative can constrain persuasion: as the fit of a competing narrative increases, advisors increase the fit and reduce bias in their own narratives. This further supports the fit-movement tradeoff in S&S: to be persuasive, advisors must offer a narrative that fits better than the alternative. As the competing narrative fit increases, this limits how ambitious they can be on the movement dimension.

Fifth, we examine how investors update their beliefs when presented with a new narrative. Consistent with the bang-bang adoption rule in S&S, most investors either fully adopt the advisor's narrative or retain their prior beliefs. We also find that belief updating is sensitive to contextual features: when the advisor's incentive type is not disclosed, investors update similarly in response to narratives from aligned and misaligned advisors; however, when the type is explicitly disclosed or investors receive warnings about conflicts of interest, they update significantly less in response to narratives from misaligned advisors. This shift is driven by greater skepticism toward misaligned advisors and greater trust in aligned advisors. In addition, while narrative fit plays a central role in belief updating when incentives are not disclosed, its influence diminishes once disclosures or warnings are introduced.

We also document several further findings that support and extend the core results. We provide direct evidence that the quality of *explanations* matters: *claims* accompanied by high-quality explanations are more persuasive than those paired with low-quality ones. Finally, show that our main results are not driven by participant inattention; they remain robust across a series of exercises that exclude potentially inattentive participants.

Taken together, these results are broadly in line with the predictions and assumptions of S&S. In situations where public information needs to be *interpreted*, narratives can be used as a persuasive tool, with narrative fit playing a key role in determining whether a proposed narrative is believed. This stands 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. Our results do not support this: we show that investors' beliefs are meaningfully shifted in the direction the advisor wishes to bias them. Second, investors could engage in sophisticated strategic reasoning when interpreting the messages they receive from advisors. In relation to this benchmark, the evidence we present from SYMMETRIC demonstrates that persuasion occurs even when both individuals have identical information—i.e., when communication is purely about the interpretation of public information. This suggests that narratives can persuade beyond what is predicted by standard strategic communication models, where persuasion relies on asymmetric information. Together, the results provide novel evidence on the mechanics of narrative persuasion and how individuals update their beliefs in response to a narrative.

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

Our paper contributes to several strands of literature. First, it adds to the recent work in economic theory that has begun to study narratives more formally.[2] One thread of this literature examines 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, 2020; Mailath and Samuelson, 2020; Montiel Olea, Ortoleva, Pai, and Prat, 2022; Schumacher and Thysen, 2022; Jehiel, 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 thread of this literature that relates more closely to our paper analyzes the factors that may influence narrative adoption and the implications for persuasion using narratives (e.g. Froeb, Ganglmair, and Tschantz, 2016; Eliaz and Spiegler, 2020; Schwartzstein and Sunderam, 2021; Ichihashi and Meng, 2021; Jain, 2023; Ispano, 2023; Lang, 2023; Aina, 2024). 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. 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). 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 learn about decision-making in such settings, we test a set of predictions derived from the framework developed by Schwartzstein and Sunderam (2021). Their key assumption is that the receiver will adopt

[2]Currently, there is no 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 Barron and Fries (2024b), we provide a brief overview of the existing literature on narratives and discuss how different conceptualizations of the concept relate to one another. In the current paper, we use the term to refer to *a causal explanation that makes sense of a collection of events*.

a narrative if it explains the data sufficiently well. One central goal of our experiment is to test whether this 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 agents (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 literature by focusing on the *interpretation of facts*. Specifically, advisors in our experiment send messages that not only make claims about the payoff-relevant parameter but also use the payoff-irrelevant parameters to provide an explanation that justifies their claim. This increases the overall fit of the message to the data, making it more convincing. In contrast, in a standard cheap talk framework, sending these additional payoff-irrelevant parameters is typically inconsequential, since strategic considerations make it impossible to achieve informative communication in payoff-irrelevant 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.[3] This literature can be divided into two complementary strands: one set of studies investigates the influence of these constructs within a particular important economic domain (e.g., understanding inflation behavior), while the other focuses on identifying the underlying psychological mechanisms through which these constructs shape behavior. For example, in the first strand, 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. Turning to the second strand, Graeber et al. (2024) explore the relationship between stories and memory, showing that information embedded in a story has a slower memory decay rate than statistics presented without a story context. To further explore the idea that explanatory content affects how individuals process and respond to information, Graeber et al. (2024) investigate the role of qualitative explanations, studying how hearing a verbal explanation affects an individual's willingness to imitate the choice of another individual (for related work from social psychology on

[3]The theoretical work on narratives discussed above has been followed swiftly by 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: Alysandratos, Boukouras, Georganas, and Maniadis (2020); Laudenbach, Weber, and Wohlfart (2021); Andre, Pizzinelli, Roth, and Wohlfart (2022); Gehring, Harm Adema, and Poutvaara (2022); Hillenbrand and Verrina (2022); Barron, Harmgart, Huck, Schneider, and Sutter (2023); Harrs, Berger, and Rockenbach (2023); Hüning, Mechtenberg, and Wang (2022); Andre et al. (2023); Liu and Zhang (2023); Ambuehl and Thysen (2024); Charles and Kendall (2024); Frechette, Vespa, and Yuksel (2024); Graeber, Roth, and Schesch (2024); Graeber, Roth, and Zimmermann (2024); Hagmann, Minson, and Tinsley (2024); Morag and Loewenstein (2025). This empirical work has approached the topic from several different methodological angles.

explanations, see, e.g., Lombrozo, 2006, 2012). Within this literature, the two experiments most closely related to ours are Charles and Kendall (2024), who demonstrate how narratives (represented by DAGs) influence beliefs and study their (nonstrategic) transmission, and Ambuehl and Thysen (2024), who examine how individuals prioritize different features of narratives (such as narrative fit, complexity, or the optimism of the embedded claim) when choosing between competing causal interpretations. One key distinction between our study and this body of empirical work is our focus on narratives in a fully specified strategic context, where one individual seeks to persuade another by shaping their interpretation of facts. Within this setting, we examine both the construction and adoption of narratives, as well as the factors that influence each.

3 Theoretical Framework

This section lays out a theoretical framework in which narratives shape how individuals *interpret* objective data. This framework is based on S&S. We use it to derive predictions for our investor-advisor setting and compare these predictions to those from a rational benchmark.

3.1 Basic Setup

An advisor ("she") and an investor ("he") observe company data $\boldsymbol{h} = (s_1, s_2, \ldots, s_{10})$, which is a vector chronicling company success ($s_t = 1$) and company failure ($s_t = 0$) across a period of ten years. This data is generated by an underlying data-generating process, or true model, $\boldsymbol{m}^T \in \mathcal{M}$. The true model is characterized by three parameters: $c^T \in \{2, 3, \ldots, 8\}$, indicating the final year of the initial CEO's tenure, and $\theta^T_{pre}, \theta^T_{post} \in [0, 1]$, which denote the company's success probabilities under the initial and current CEOs, respectively. The set of feasible models $\mathcal{M} = \{2, 3, \ldots, 8\} \times [0, 1]^2$ includes all possible combinations of these three parameters. Neither agent knows the true model. The advisor sends a narrative $\boldsymbol{m}^A \in \mathcal{M}$ to the investor. The investor then decides whether to adopt the advisor's narrative or retain his default narrative, which is explained below. Based on this decision, he forms an assessment of the current CEO's success probability, $\theta^I_{post} \in [0, 1]$. This assessment affects the payoff of both agents.

Types. The investor's default narrative $\boldsymbol{m}^{I,0}$ serves as an "investor type," drawn from a distribution $f(\boldsymbol{m})$ that has full support on $\mathcal{M}$. The advisor is characterized by an incentive type, φ, drawn with equal probability from $\{\uparrow, \downarrow, \rightarrow\}$. This type determines whether her incentives are aligned or misaligned with the investor's, as we will explain further below. Each agent's type is private information.

Payoffs. The investor's objective is to form an assessment that is as close as possible to the truth. His 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 induce an assessment that is as close as possible to her type-dependent persuasion target τ_φ. For up ($\uparrow$), $\tau_\uparrow = 1$. For down ($\downarrow$), $\tau_\downarrow = 0$. For aligned ($\rightarrow$), $\tau_\rightarrow = \theta^T_{post}$. Hence,

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

This utility is maximized if $\theta^I_{post} = \tau_\varphi$. This means an up-advisor wants an assessment near 1, a down-advisor near 0, and an aligned advisor near the true value θ^T_{post}.

To derive predictions, we first specify how the investor decides whether to adopt the advisor's narrative and how that choice affects his assessment. We then characterize the advisor's optimal narrative, anticipating this adoption rule.

The narrative adoption rule. One key ingredient of S&S's framework is the assumption that, when presented with two different narratives, an agent adopts the one that prevails in a "Bayesian hypothesis test." This narrative, in turn, determines the agent's beliefs and actions.

In a persuasion scenario, the investor faces a decision: to either retain their pre-existing default narrative $\boldsymbol{m}^{I,0}$ or to adopt the narrative $\boldsymbol{m}^{\boldsymbol{A}}$ proposed by the advisor.[4] 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 did. Denote the adopted narrative by $\boldsymbol{m}^{I,1}$ and the likelihood that the model $\boldsymbol{m}$ generated $\boldsymbol{h}$ by $\pi_{\boldsymbol{m}}(\boldsymbol{h})$. Then, the adoption rule is described by the following:

$$\boldsymbol{m}^{I,1}(\boldsymbol{m}^{I,0}, \boldsymbol{m}^A) = \begin{cases} \boldsymbol{m}^A & \text{if } \pi_{\boldsymbol{m}^A}(\boldsymbol{h}) \geq \pi_{\boldsymbol{m}^{I,0}}(\boldsymbol{h}), \\ \boldsymbol{m}^{I,0} & \text{otherwise.} \end{cases} \tag{3}$$

Two features of this adoption 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, S&S consider a variant in which the receiver forms beliefs by averaging across the available narratives, weighting each by its relative fit.[5] We choose to employ the binary structure primarily due to its simplicity, which sharpens the focus on the specific components of persuasion that are of interest to us. Adopting a weighted 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. An investor who adopts narratives according to Equation (3) and who maximizes $U^I(\theta^I_{post}, \theta^T_{post})$ sets his assessment equal to the assessment implied by the adopted narrative: $\theta^I_{post}(\boldsymbol{m}^{I,0}, \boldsymbol{m}^A) = \theta^{I,1}_{post} \in \boldsymbol{m}^{I,1}(\boldsymbol{m}^{I,0}, \boldsymbol{m}^A)$.

[4] Instead of assuming a single 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.

[5] This variant is included as an extension in Appendix E.2 of Schwartzstein and Sunderam (2020).

Having laid out the core model, we now derive the advisor's optimal narrative choice and highlight the key predictions of the S&S-inspired theoretical framework.

The advisor's choice of narrative. The advisor chooses the narrative that maximizes her utility. When doing so, she anticipates that she will only influence the investor's assessment if her narrative beats the default narrative in terms of fit. As the default is the investor's private information, the advisor chooses the narrative that maximizes her expected utility, with the expectation being taken over the distribution of the investor's default narrative:

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

It is useful to benchmark the advisor's narrative against the one that best fits the data, which we refer to as the *data-optimal* narrative and denote by $\boldsymbol{m}^{DO}$.[6] Proposition 1 establishes that whenever the advisor's persuasion target differs from the data-optimal value, she shifts her recommended θ^A_{post} *away from the data-optimal value* and *toward her persuasion target*. She simultaneously adjusts the auxiliary parameters, 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.[7]

Proposition 1. *For an advisor of type 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 her narrative away from the data-optimum, θ^{DO}_{post}, and towards the persuasion target:

$$\begin{aligned}
\theta^A_{post} > \theta^{DO}_{post} \quad &if \quad \theta^{DO}_{post} < \tau_\varphi, \\
\theta^A_{post} < \theta^{DO}_{post} \quad &if \quad \theta^{DO}_{post} > \tau_\varphi, \\
\theta^A_{post} = \theta^{DO}_{post} \quad &if \quad \theta^{DO}_{post} = \tau_\varphi.
\end{aligned}$$

(ii) (Improving coherence) Among all narratives in $\mathcal{M}$, the advisor considers only those with values of c and θ_{pre} that maximize the likelihood function, conditional on θ_{post}. This implies that:

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

The proposition clarifies that, when constructing the narrative, the advisor considers both narrative *fit* and belief *movement*. While the advisor can ensure narrative adoption by sending the data-optimal narrative, this will usually not be her 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

[6] Formally, $\boldsymbol{m}^{DO} \in \arg\max_{\boldsymbol{m}\in\mathcal{M}} \pi_{\boldsymbol{m}}(\boldsymbol{h})$. Note, the data-optimal narrative is not equivalent to the true model.

[7] In Appendix C, we provide the proof of Proposition 1 and a more detailed discussion of S&S's narrative approach.

the targeted movement operates entirely on θ^A_{post}, the advisor will always select the auxiliary parameters θ^A_{pre} and c^A such that she maximizes the likelihood fit *conditional* on θ^A_{post}.

Differences from S&S. Our theoretical setup modifies the S&S framework in several important ways. First, we relax their assumption that the advisor knows the investor's default narrative. Consequently, advisors in our setting face uncertainty about whether their chosen narrative will persuade the investor. This modification aligns closely with our experimental design, where advisors typically do not know the investor's default. Second, we assume that the advisor does not know the true model (in contrast, S&S assume that she does). While this difference is immaterial for the theoretical predictions, it sharpens the contrast with the rational benchmark, where knowledge of the true model is crucial. We discuss this rational benchmark in more detail below. Third, unlike S&S, who allow for uncertainty (about the DGP) beyond what the narrative explains, our environment abstracts from such residual uncertainty. This allows for sharper tests of the adoption rule, without potential contamination caused by belief updating biases unrelated to the core persuasion mechanism.[8] Finally, our distinction between payoff-relevant and auxiliary parameters highlights how advisors strategically choose the payoff-relevant parameter (claim) to move the investor's assessment and select the auxiliary parameters (explanation) to facilitate persuasion. We show that the fit-movement tradeoff—shown by S&S to hold when the set of models is large enough to permit all possible likelihood functions—also applies to our specific setting.

Before turning to the testable predictions, we now compare our framework with a rational benchmark, highlighting where they diverge.

Rational benchmark. In the experiment, we induce a common prior over the data-generating process by truthfully informing participants that the three individual parameters are independently drawn from uniform distributions. This enables us to analyze our setup through the lens of a cheap talk game, thereby establishing a rational benchmark. Appendix D details the model; here, we summarize the main insights. Most fundamentally, under cheap talk, the investor is fully capable of interpreting the data independently. Therefore, in the setting described above, where the advisor and investor have *symmetric* information about the company, cheap talk predicts no influential communication. This is in stark contrast to S&S, where the investor sometimes relies on the advisor's interpretation of the data, implying that communication can be influential even with symmetric information.

In the cheap talk framework, the investor begins by forming a posterior belief about the new CEO's success probability, θ_{post}, by updating using the observed data. He then factors in the advisor's message and equilibrium strategy to see if it conveys additional information about θ_{post}.

[8] In particular, S&S propose a more general adoption rule that can be applied in settings where the narrative does not explain all the uncertainties in a given environment. They define fit as $\int_\omega \pi_{\boldsymbol{m}}(\boldsymbol{h}|\omega)\mathrm{d}\mu_0(\omega)$, where ω denotes uncertain states of the world and $\mu_0(\omega)$ is the prior over ω. In such an environment with residual uncertainty, the *assessment* we consider would be a function of the adopted narrative and the prior; $\theta_{post}(\boldsymbol{m}^{I,1}, \mu_0)$. However, in our environment, without residual uncertainty beyond the parameters included in the narrative, the rule simplifies to the one we specified in (3). Hence, the assessment becomes a function of the adopted model only; $\theta_{post}(\boldsymbol{m}^{I,1})$. Therefore, the S&S adoption rule reduces to our simplified formulation in the setting we consider.

Under symmetric information, however, the advisor cannot transmit any information that the investor doesn't already know, so the investor ends up ignoring the advisor's narrative. By Bayes' rule, his final assessment is equal to $\theta^e_{post} = \frac{\int_{m\in\mathcal{M}} \theta_{post}\pi_m(\boldsymbol{h})\mathrm{d}\boldsymbol{m}}{\int_{m\in\mathcal{M}} \pi_m(\boldsymbol{h})\mathrm{d}\boldsymbol{m}}$. Hence, with symmetric information, the advisor's narrative does not alter the investor's posterior; the company data alone determines his assessment.

When the advisor knows the true data-generating process (which is the case in our ASYMMETRIC treatments), communication may be influential, but cheap talk still does not ascribe a persuasive role to the fit of the auxiliary parameters. The reason for this is that a narrative only persuades an investor if it credibly signals that it came from an aligned advisor. Since a misaligned advisor can costlessly choose auxiliary parameters that yield a high fit, investors cannot distinguish aligned from misaligned narratives based solely on the auxiliary parameters—conditional on a specific θ_{post}, these auxiliary parameters contain no additional information. Appendix D establishes this *irrelevance of auxiliary parameters* result and shows that it remains robust for variants of the cheap talk game that either introduce a subset of honest advisors or relax the investor's rationality in various ways.

3.2 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 the rational benchmark. This will help us to evaluate whether the S&S approach offers a useful additional lens for understanding persuasion in certain contexts.

We 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. Because real-world contexts often allow for persuasion through both interpretation and information transmission, we consider two scenarios: a 'pure interpretation' scenario, in which there is no scope for information transmission, and a 'hybrid' scenario, in which the advisor can both provide and interpret 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. We also discuss how these fit-based predictions contrast with those of the rational benchmark.

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 contrasts with the predictions of the rational benchmark, which conceptualizes communication as a game in which the advisor can signal her private information. Under this rational benchmark, whether the advisor

knows the truth, $\boldsymbol{m}^T$, is of central importance, and persuasion is not possible in the pure interpretation scenario. In contrast, under the S&S assessment rule, it is the fit of the narrative that matters, irrespective of whether the advisor is more informed. Therefore, persuasion is possible even in the pure interpretation scenario.

Prediction 1 (Persuasion)**.** *The narrative sent by the advisor shifts the investor's assessment towards the advisor's persuasion target in:*

(i) the pure interpretation scenario, where it is common knowledge that the advisor has no additional information relative to the investor.

(ii) the hybrid scenario, where the advisor knows the true data-generating process, while the investor does 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 than 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, this pattern could also result from endogeneity of investors' priors and advisors' messages. 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)**.** *To measure how closely the investor's final assessment, θ^I_{post}, aligns with the advisor's recommended θ^A_{post}, we calculate the distance $|\theta^I_{post} - \theta^A_{post}|$. Then:*

(i) As the likelihood fit of the advisor's narrative increases, the distance $|\theta^I_{post} - \theta^A_{post}|$ decreases.

(ii) Even if the advisor's recommended θ^A_{post} is held fixed, increasing the narrative's fit (i.e., by adjusting auxiliary parameters) reduces $|\theta^I_{post} - \theta^A_{post}|$.

Advisor behavior

Moving to advisors, we derive predictions for how they construct narratives when they expect the investor to follow an S&S fit-based assessment rule. Consider a data-generating process in which the true model is drawn from a uniform distribution and used to generate a company history. The advisor then observes this history before selecting a narrative. When choosing her narrative, an up-advisor will select a θ^A_{post} that is higher than the data-optimal value, moving it toward her persuasion target. She will support this θ^A_{post} claim by selecting auxiliary parameters that maximize the fit, conditional on θ^A_{post}. For example, the up-advisor may justify sending a high θ^A_{post} by adjusting her assertion about the position of the structural change year, c^A, to artificially increase the apparent success rate in the *post* period. This mechanically lowers the success rate in the *pre* period. To maintain the fit of her narrative to the data, this will induce her to decrease her

stated θ^A_{pre}. Thus, the up-advisor raises θ^A_{post} above its data-optimal value and lowers θ^A_{pre} below its data-optimal value. Conversely, a down-advisor will follow the reverse logic, reducing θ^A_{post} and increasing θ^A_{pre}. This yields a testable prediction: On average, θ^A_{post} will be higher and θ^A_{pre} will be lower for up-advisors than for down-advisors.

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:*

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

In a second test of the fit-movement tradeoff, we derive a prediction for advisor behavior when she knows the investor's default (competing) narrative at the time of constructing her own. This will be the case in one of our treatments. The prediction is that, as the fit of the default narrative increases, the advisor will sacrifice belief movement in order to improve the fit of her own narrative. The reason is intuitive: if she anticipates that the investor will have access to another narrative that fits very well, she needs to send a narrative that fits even better 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)**.** *Suppose the advisor knows the investor's default narrative. Assume also that the investor's default assessment ($\theta^{I,0}_{post}$) is held fixed, while the other components of the default are allowed to vary. Then, as the default narrative's likelihood fit increases, two things happen:*

(i) ***Fit dominance requirement.*** *The fit of the narrative chosen by the advisor increases.*

(ii) ***Less scope for movement.*** *The advisor's claim, θ^A_{post}, becomes less biased—i.e., she moves it further away from her persuasion target.*

We conclude by briefly discussing how persuasion dynamics may change in scenarios similar to those above, but where a particular feature of the environment is altered. In the experiment, we introduce three *intervention treatments*. These interventions induce the investor to form a more carefully considered default, disclose the advisor's incentives, and highlight the possible consequences of incentive misalignment. Following S&S, we model disclosed incentive misalignment by assuming that the investor adopts a narrative only if its fit exceeds that of the default by a margin $x > 0$—that is, the advisor's narrative must fit substantially better to be adopted. Turning to improvements in investors' default formation, we model this as drawing from a default distribution that is more concentrated around the data-optimal narrative. These considerations generate the intuitive prediction that such interventions can immunize investors against persuasion and that incentive disclosure reduces the sensitivity to a better fit. A formal treatment appears in Appendix C.3.

4 Experimental Design

This section describes the experimental implementation of the investor-advisor setup and introduces our main treatment conditions.

4.1 Implementation of the Basic Setup

We begin by outlining how we translated the theoretical setup from Section 3 into an experiment.

Roles. Participants are randomly assigned to the role of either an advisor or an investor. They remain in this role for the entire experiment.

Timing. In each round, the experiment unfolds in four steps. First, the true parameters governing company success ($c^T, \theta^T_{post}, \theta^T_{pre}$) are each drawn from independent uniform distributions. Second, the company data is generated on the basis of these parameters. Third, the advisor observes the company data and chooses a narrative to send to the investor. Fourth, the investor observes the company data and receives the narrative from the advisor. He then reports his assessment of the company's current probability of success.

Choices. The *advisor's choice* consists of proposing a narrative to the investor. This involves choosing an integer value for each of the three parameters: the company's probability of success under the previous CEO (θ^A_{pre}), the success probability under the current CEO (θ^A_{post}), and the year in which the CEO changed (c^A). These three parameters constitute the entire narrative.[9] The *investor's choice* consists of assessing the company's current probability of success (θ^I_{post}). To do this, he chooses an integer between 0 and 100.

Incentives. Participants can earn a bonus payment based on their choices: both players' chances of winning the bonus depend on the investor's choice, but the advisor can try to influence this choice through her own actions.

More specifically, the probability of receiving this bonus is specified by the payoff functions in equations (1) [investor] and (2) [advisor] above. The investor's probability of winning the prize 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}.

Because we are studying persuasion, the advisor's probability of winning depends only on the investor's assessment. Her chance of winning the prize is maximized if the investor's assessment is equal to her (the advisor's) persuasion target. She also suffers a quadratic loss (in probability)

[9] 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 chose an integer between 3 to 9 to indicate the structural break. For expositional clarity, throughout the paper we will continue using the convention that the structural change parameter denotes the *last* year under the old CEO, i.e., that it takes on values between 2 and 8. We recoded all variables in the analysis to be consistent with this "last year under the initial CEO" convention.

as the investor's assessment moves away from her persuasion target.[10] 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).

Core treatment variation: ASYMMETRIC and SYMMETRIC. We implement the basic setup in two core treatments that differ in what the advisor knows about the company when constructing the narrative. In one, the advisor and investor have *identical information*; in the other, the *advisor has more information* than the investor. More specifically, we consider the following two scenarios:

(i) **A Pure Interpretation Scenario.** In SYMMETRIC, the advisor is provided with exactly the same information about the DGP as the investor—neither individual has any private information 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 is 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.

Dynamics. The experiment consists of ten rounds of company assessments, meaning that each advisor and investor evaluates ten companies. In each round, advisors and investors are randomly rematched without identifiers, limiting the possibility of reputation effects. They receive no feedback until the end of the experiment. This minimizes the scope for learning and the potential interdependence of choices across the ten rounds of the experiment.

4.2 The COMPETITION and BELIEF UPDATING Treatments

Building on the core treatments, we implement a set of additional treatments to investigate the underlying behavioral mechanisms and examine how specific changes to the experimental environment influence behavior.

4.2.1 COMPETITION Treatment

The COMPETITION treatment is similar to SYMMETRIC but introduces a robot advisor who also sends a narrative to the investor. Therefore, in this treatment, investors receive two narratives—one from the robot advisor and one from the human advisor. The robot's narrative is determined by a strategy chosen by us, the experimenters, which means we can exogenously vary features of this

[10] As discussed in the introduction and Section 2, an advisor's persuasion target depends on her type. Advisors are randomly assigned to one of three types—*up*, *down*, or *aligned*. *Up-advisors* are incentivized to persuade investors that the company's current probability of success is high; *down-advisors*, that it is low; and *aligned advisors*, to induce accurate investor beliefs. Investors are fully informed about the three advisor types, their incentives, and that they are equally likely to be matched with each type in every round. However, they are not told the specific type of the advisor they are matched with in a given round.

competing narrative. The COMPETITION treatment allows us to cleanly identify how a competing narrative affects the advisor's narrative construction and the investor's narrative adoption.

Each round in COMPETITION proceeds as follows. First, the robot advisor constructs a narrative. Second, the human advisor observes both the company data and the robot advisor's narrative. She then constructs her own narrative to send to the investor. Third, the investor observes the company data and both advisors' narratives. These are labeled as coming from "Advisor A" or "Advisor B," but the investor does not know which label refers to the human and which to the robot. The investor must then make a binary choice between the two narratives: He is incentivized to select the one with the more accurate *claim* (i.e., θ_{post}—value).

We programmed the robot advisors to always make an accurate *claim* about the company's current probability of success—that is, they always send the true θ_{post}-value. However, robot advisors differ in how their *explanation* is constructed—that is, in their choice of the auxiliary parameters (c, θ_{pre}).

Specifically, over the five rounds of the COMPETITION treatment, the robot advisors follow the strategies outlined below:

- In Round 1, the robot sends the true data-generating process $(c^T, \theta_{pre}^T, \theta_{post}^T)$. This allows us to examine how often the investor chooses the human advisor's narrative over the true narrative (i.e., how often the truth is "beaten").

- In Rounds 2–5, the robot continues to send the true θ_{post}-value, but we introduce exogenous variation in the two auxiliary parameters to generate either a HIGH or a LOW narrative fit. Importantly, we introduce this variation in a controlled way by creating pairs of observations with identical company histories, but where one robot advisor chooses LOW fit auxiliary parameters and the other chooses HIGH fit auxiliary parameters.[11] This allows us to examine how exogenous variation in auxiliary parameter fit causally influences narrative construction and adoption.

Appendix B provides a detailed description of the algorithm used to generate the robot narratives. Importantly, participants receive a simplified, high-level explanation of how the robots construct their messages. Specifically, they are told that the robot advisor is "always trying to help" the investor but that "not all robot advisors are equally skilled" at assessing the company. This characterization communicates the robot's strategy in simple words while avoiding a complex and detailed description of the strategy-determining algorithm.

4.2.2 BELIEF UPDATING Treatments

To further investigate the psychological mechanisms of narrative persuasion, we introduce three additional treatments. These treatments build on the SYMMETRIC treatment, introducing two

[11] Specifically, for every realized company history in the COMPETITION treatment, we construct two observation triples, each consisting of a human advisor, a robot advisor, and an investor. Across these two triples, the only feature of the choice architecture that we vary is the fit of the auxiliary parameters chosen by the robot advisor. Crucially, we hold constant (i) the company data, (ii) the incentive type of the human advisor, and (iii) the *claim* (θ_{post}) sent by the robot advisor. This design allows for a clean evaluation of the causal influence of variation in narrative fit on behavior.

key innovations. First, participants' prior beliefs are elicited before they encounter the advisor, allowing us to examine how individuals update their beliefs when they receive a narrative. Second, the treatments are designed to progressively increase the salience of reasons to distrust narratives from misaligned advisors. This enables us to assess the effectiveness of interventions such as *disclosing advisor incentives* or *issuing explicit warnings* about conflicts of interest in mitigating persuasion by misaligned advisors.

These additional treatments vary only the investor's instructions, keeping the advisor's instructions constant and nearly identical to those in the SYMMETRIC treatment, with one important exception: instead of informing advisors that investors "do not know" their incentive type—as in SYMMETRIC—we tell advisors in the BELIEF UPDATING treatments that investors "may or may not" know their incentive type. This allows us to selectively disclose incentives to investors without deceiving advisors.[12]

SYMMETRIC-PRIOR. The SYMMETRIC-PRIOR treatment builds on the SYMMETRIC treatment by asking investors to report their own prior (default) narrative before meeting with an advisor. Specifically, investors first see only the company data and are asked to provide their own assessment of the three parameters they believe generated the data. Thereafter, they receive the advisor's narrative and are required to report a final assessment of θ_{post}. This design allows us to examine belief updating. It also enables us to investigate whether investors become less susceptible to persuasion when considering the data on their own *before* receiving a narrative.

SYMMETRIC-DISCLOSURE. The SYMMETRIC-DISCLOSURE treatment is identical to SYMMETRIC-PRIOR, except that in every round, the advisor's incentives are explicitly disclosed to the investor on their decision screen.

SYMMETRIC-WARNING. The SYMMETRIC-WARNING treatment builds on SYMMETRIC-DISCLOSURE by additionally emphasizing the potential consequences of preference misalignment and the fact that advisors lack additional information. This is achieved in two ways:

(i) Directly before starting the company evaluation task, investors see an information page that explicitly states that misaligned advisors may send biased messages. It also highlights the symmetry of information between advisors and investors, emphasizing that *"[...] advice is based on the same information that you have."*

(ii) On their decision screens, in addition to seeing their advisor's incentives (as in SYMMETRIC-DISCLOSURE), investors receive a "warning" indicating whether their advisor has a conflict of interest and reminding them that advisors do not possess additional information.

[12]To obtain a clean comparison group for the BELIEF UPDATING treatments, we simultaneously ran a SYMMETRIC-REP treatment that is identical to the SYMMETRIC treatment, except for this difference in the advisor instructions. See Section 4.3 for further details.

4.3 Procedures

The experiment was programmed in oTree (Chen, Schonger, and Wickens, 2016) and participants were recruited via the Prolific platform. Appendix A.1 displays summary statistics of participant demographics by treatment. We included several understanding questions that participants were required to answer correctly before proceeding. 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).[13]

Data collection timing (across waves): We collected the data in three waves. In March 2022, we recruited 360 participants (180 advisors and 180 investors) to participate in the Asymmetric treatment. 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. In March 2025, we collected data for the Belief Updating treatments (Symmetric-Prior, Symmetric-Disclosure, and Symmetric-Warning). For each of these treatments, we collected observations from 90 advisors and 90 investors. In this wave, we also replicated our earlier Symmetric treatment by collecting an additional 90 advisor and 90 investor observations—in a treatment we refer to as Symmetric-Rep. This replication offers a contemporaneous benchmark against which we evaluate the Belief Updating treatments.

Group assignment: Participants took part in the experiment in groups of six. Within each group, three participants were randomly assigned to the role of advisor and three to the role of investor. There was one advisor from each incentive condition (up, down, and aligned) within each group of six. Both advisors and investors kept their role for the entire duration of the experiment. In every round of the experiment, each investor was randomly matched with an advisor within their group of six.

Data collection timing (within waves): To avoid issues related to attrition, we took advantage of the sequential nature of our setting by always collecting advisor data before collecting investor data. That is, we first collected three advisor observations for each matching group for all rounds, and only collected the investor data thereafter. For a typical matching group, this meant collecting investor data the day after collecting advisor data. As a result, no participant's experience was interrupted by someone from their group dropping out, nor were they influenced by the decision speed of other participants.[14]

Generating the company data: In the Asymmetric, Symmetric, and Belief Updating treat-

[13]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 three 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, AEARCTR-0011565 and AEARCTR-0015403. Additionally, instructions for all treatments can be found under the following link: https://tilmanfries.github.io/assets/pdf/NarrativePersuasionInstructions.pdf.

[14]Attrition was generally low. Conditional on starting the main task (i.e., the investment or narrative construction task), attrition rates ranged from 3-7% across all data collection waves.

ments, all matched investor–advisor pairs saw company data generated by the same true underlying DGP in each round. We drew ten DGPs—one for each round—before the first session of Asymmetric, and used this same sequence in all of these treatments. In contrast, the Competition treatment uses a different set of DGPs. In Round 1 of Competition, we drew a DGP for each matching group of investors and advisors, so the true DGPs varied across groups. In Rounds 2–5, a new DGP was drawn for each round, but held constant across all advisor-investor pairs within that round.

In each round, the company data was drawn conditional on the true DGP. However, we slightly increased the sophistication of the experimental design across the waves of data collection, in order to improve empirical control in our analysis. In Asymmetric (2022), the historical data was drawn independently for each investor-advisor pair. In the 2023 data collection wave, we opted to instead randomly draw the historical data on the matching group level. This allows us to control for Round×History fixed effects when analyzing data from these later data collection waves. In rounds 2-5 of Competition and in all treatments of the 2025 data collection wave, we drew one historical data set for every two matching groups. This increases the number of observations that we have per realized data set. It also ensures that we have at least two observations for each Advisor Type×History combination, which allows us to control for Round×History fixed effects while also conditioning on advisor type.

Participant payments: 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. In the 2025 collection wave, we increased the participation fee to £4 and the bonus payment to £4.25. 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 to complete the Asymmetric and the Belief-Updating treatments and 30-35 minutes for Symmetric plus Competition.

5 Main Results

In this section, we present our key results. Sections 5.1 and 5.2 test core theoretical predictions for investor and advisor behavior using within-treatment variation in data from the Asymmetric, Symmetric, and Competition treatments. Section 5.3 then turns to examining belief updating and the limits of persuasion using data from our Belief Updating treatments. In Section 6, we discuss the results from additional exercises that serve to extend and provide robustness checks for these core results.

5.1 Investor Behavior (Narrative Adoption)

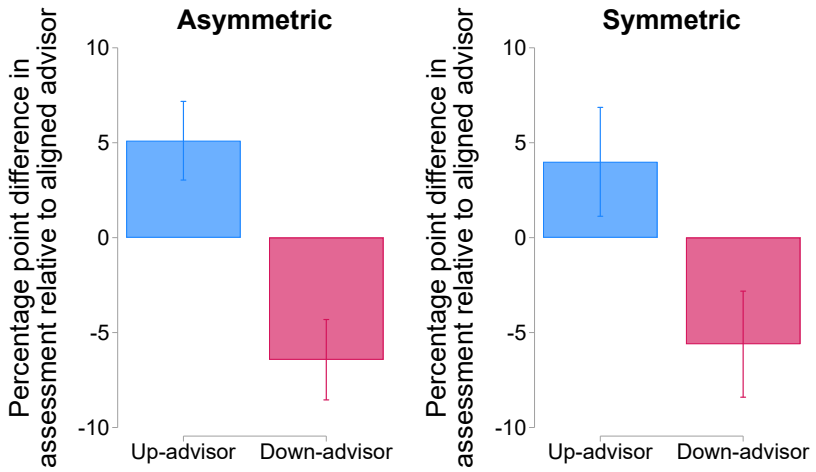
We first discuss the behavior of investors (Predictions 1 and 2), examining whether narratives shift investors' beliefs and evaluating the role of narrative fit.

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 2 addresses this question, showing that investors' beliefs are systematically shifted towards the advisors' persuasion targets. The figure displays the coefficients from a regression of the investor's posterior beliefs, θ^I_{post}, on indicator variables for the up- and down-advisors. 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 (in comparison to those matched with an aligned advisor). Figure A.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. It is consistent with the S&S model but stands in contrast to standard cheap talk models where persuasive equilibria are sustained by an information asymmetry between sender and receiver. In particular, our finding that narratives remain persuasive in Symmetric, where there is no information asymmetry, suggests that persuasion extends beyond the scope of cheap talk.

Result 1 (Persuasion)**.** *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 2: Effect of the advisor type on investor assessments



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 A.4.

Influence of narrative fit: Having shown that narrative persuasion is effective, we examine

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 of the data-optimal (best-fitting) narrative for the relevant history.[15] 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.[16] We use the EPI to investigate the relationship 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. We focus on this exercise first because it provides a clean setting to identify the direct causal effect of fit. Specifically, we ask whether investors choose to follow the human advisor's narrative less often when the fit of the competing robot advisor's narrative is exogenously improved.[17]

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

	(1) $\mathbb{I}$(adopt $\boldsymbol{m}^A$)	(2) $\mathbb{I}$(adopt $\boldsymbol{m}^A$)	(3) $\mathbb{I}$(adopt $\boldsymbol{m}^A$)
$\mathbb{I}$(Robot Narrative Fit = HIGH)	-0.0778***		
	(0.0235)		
EPI of Robot Narrative (cont. fit)		-0.138***	
		(0.0447)	
$\mathbb{I}$(Robot Narrative Fit $\geq$ Human Narrative Fit)			-0.237***
			(0.0379)
Round × History FE	Yes	Yes	Yes
Observations	720	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) 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 (as opposed to LOW). The coefficient shows that the investor is 7.8 pp 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

[15] 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 26 in Section A.3 and also on pages 35-37 in Section A.5.

[16] 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.

[17] Recall 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.

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. In Column (3), we include a binary indicator that equals one if the robot's narrative fits the data better than the human advisor's narrative. The estimation results indicate that the investor is 24 pp less likely to adopt the human narrative if the robot narrative fits better, reinforcing the previous findings. This last result contextualizes the qualitative findings with a quantitative benchmark. According to a strict interpretation of the theoretical framework, investors should always adopt the better-fitting narrative, which would yield a coefficient of -1. The estimated coefficient is less extreme, suggesting that other factors—such as imperfect perception of fit or reliance on additional decision criteria—influence investor behavior. In Section 6, we present results from an additional exercise that quantifies "noisy" narrative adoption and discuss the implications for optimal narrative construction.

In a second exercise on the influence of fit, Table A.5 in Appendix A.2 shows 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. This is consistent with the findings discussed in the previous paragraph, although the effect of fit is not as cleanly identified here since there is a possible endogeneity between the investor's prior and the narrative. However, in our Belief Updating treatments, discussed in Section 5.3, we are able to control for the investor's prior, which allows us to provide further evidence on the influence of fit.

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 the narrative's recommended θ_{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)

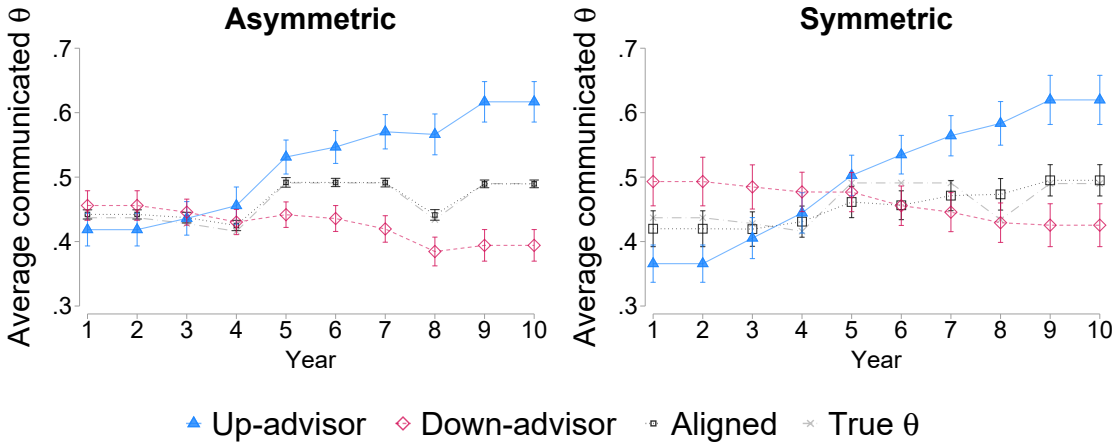
We now turn to advisors with the aim of identifying systematic regularities in the features of the narratives that advisors construct (Predictions 3 and 4).

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 3 provides a visual illustration of how advisors of different types construct narratives. The figure depicts the average narrative transmitted by advisors of each type.[18] 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 solid blue line) construct narratives with a *higher* θ in year 10 than down-advisors (denoted by the dashed 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.

Figure 3: Average narrative communicated by advisors (by advisor type)



Notes: (i) The left panel presents the average narrative of advisors in ASYMMETRIC, while the right panel presents the average narrative of advisors in SYMMETRIC; (ii) Error bars represent 95% confidence intervals that were derived from regressions which cluster standard errors at the advisor level.

To provide further support for this finding, in Table A.6 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 in both treatments, while the opposite is true for down-advisors. Finally, Figure A.2 in Appendix A.3 shows visually that this pattern holds consistently in each of the ten rounds: up-advisors send a higher θ^A_{post} and lower θ^A_{pre} than down-advisors within each round.

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

[18]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 3, we take the average θ for each year across all messages sent by advisors of each type.

with. The S&S framework predicts that when advisors believe 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. As a consequence of being more constrained, they will send a narrative that fits better and is less biased. Here, we test whether these predicted comparative statics are observed in the data.

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. This is consistent with the idea that the human advisor needs to choose a narrative that fits better than the robot advisor's narrative in order to "beat" it and be persuasive. Columns (5) and (6) provide evidence in support of the second prediction: as the fit of the competing narrative increases, misaligned human advisors choose a θ^A_{post} that is less biased toward their persuasion target. This pattern reflects the constraint imposed by a competing narrative that fits well—fewer higher-fitting alternative narratives remain available, and the advisor is forced to become less ambitious about her intended belief movement.

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

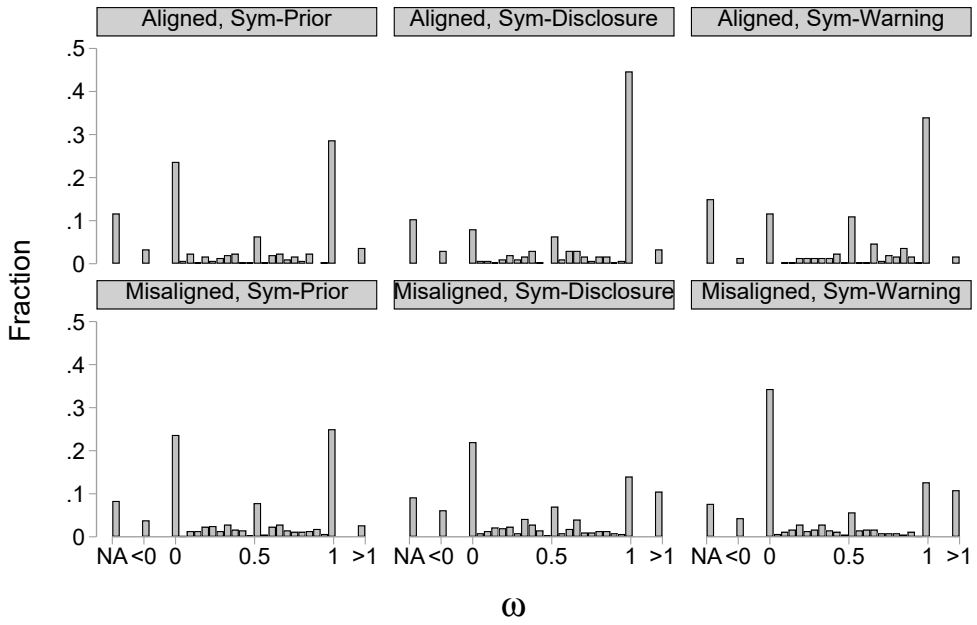
Notes: (i) The dependent variable is either the human advisor's narrative *fit* (EPI^A) or the human advisor's narrative *bias*. We define bias as the negative absolute distance between the advisor's persuasion target (1 for up-advisors and 0 for down-advisors) and the advisor's stated θ^A_{post}: i.e., Bias $= -|\tau_\varphi - \theta^A_{post}|$. This definition assigns higher values to narratives that more closely align with the advisor's persuasion target; (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 five 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; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

These findings 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.[19]

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.*

Figure 4: Distribution of updating weights, ω, by treatment and advisor alignment.



Notes: (i) The figure plots the distributions of belief updating weights by treatment and advisor alignment; (ii) A belief updating weight of $\omega = 0$ indicates that the individual retained their prior belief without updating. A value of $\omega = 1$ signifies full adoption of the advisor's message. ω =NA denotes cases where the investor's prior belief was already equal to the advisor's message. $\omega < 0$ represents instances where the individual updated in the direction opposite to the advisor's message, while $\omega > 1$ indicates that the investor adjusted their belief beyond the advisor's message.

5.3 Belief Updating and the Limits of Persuasion

This section turns to the Belief Updating treatments. Here, we vary features of the context in which persuasion takes place (by introducing interventions that may protect investors) and examine how they influence belief updating and persuasion.

Investor belief updating upon receiving a narrative: To examine belief updating, we define an updating *weight*, ω:

$$\omega = \frac{\theta^{I,1}_{post} - \theta^{I,0}_{post}}{\theta^{A}_{post} - \theta^{I,0}_{post}}.$$

[19]As a consequence of constructing narratives strategically, advisors are able to make them more convincing than the truth nearly half the time. Using Round 1 of Competition, we can test how often the human advisor "beats" the truth—that is, when the investor prefers the advisor's narrative over the true DGP. Because the true DGP need not coincide with the best-fitting narrative under the S&S framework, persuasion away from the truth is possible. We find that advisors construct narratives that investors find more convincing than the truth 49% of the time.

This weight is calculated at the individual belief update level. It captures the degree to which an investor updates their belief as a fraction of the distance between the advisor's θ^{A}_{post} and the investor's own prior, $\theta^{I,0}_{post}$. Therefore, if the investor does not update at all, then $\omega = 0$, and if they fully adopt the advisor's message and report $\theta^{I,1}_{post} = \theta^{A}_{post}$, then $\omega = 1$. If they update partially, then $0 < \omega < 1$.[20]

Figure 4 displays the distribution of updating weights, ω, in the Symmetric-Prior, Symmetric-Disclosure and Symmetric-Warning treatments. The top panel shows updating behavior when investors are matched with an aligned advisor, while the bottom panel reports updating when matched with a misaligned advisor.

The figure delivers two main insights. First, the modal updating weights are $\omega = 0$ and $\omega = 1$, meaning that the most common updating behavior is either to not update at all or to fully adopt the narrative. This behavior is consistent with the bang-bang adoption rule in S&S. We also observe a minority of intermediate updates. Second, belief updating in Symmetric-Prior appears to be similar irrespective of whether the advisor is aligned or misaligned. Thus, when advisors' incentives are undisclosed, investors treat aligned and misaligned advisors similarly when updating their beliefs. This changes in Symmetric-Disclosure and Symmetric-Warning. Here, when incentives are disclosed, investors appear to be more likely to fully adopt the advice of aligned advisors and more likely to disregard the advice of misaligned advisors.

We examine these patterns in updating behavior more formally by employing OLS to estimate the following equation:

$$\omega = \alpha_0 \cdot \mathbb{I}(\text{adv.} = \text{misaligned}) + \alpha_1 \cdot \mathbf{T} + \alpha_2 \cdot \mathbf{T} \times \mathbb{I}(\text{adv.} = \text{misaligned}) + \gamma + \varepsilon. \tag{4}$$

We run this regression on different samples, each including data from Symmetric-Prior and one other treatment (Symmetric-Disclosure or Symmetric-Warning). $\mathbf{T}$ is a treatment indicator that is equal to 0 if the treatment is Symmetric-Prior and 1 otherwise (i.e., for Symmetric-Disclosure or Symmetric-Warning). The indicator function, $\mathbb{I}$(adv. = misaligned), takes a value of 1 for observations where the advisor is misaligned and a value of 0 otherwise. We control for Round×History fixed effects, denoted by γ.

Table 3 presents estimation results from Equation (4) for the following treatment comparisons: Symmetric-Prior & Symmetric-Disclosure and Symmetric-Prior & Symmetric-Warning. Following our preregistration, we run two regressions for each treatment comparison: one that restricts the sample to investors with convex updates ($\omega \in [0, 1]$), and another that includes observations with non-convex updates ($\omega < 0$ and $\omega > 1$), censoring them at zero and one, respectively.

The estimation results provide statistical evidence in support of our earlier discussion. First, we find no significant difference in belief updating when investors are matched with an aligned versus

[20] Note that it is not necessary for ω to lie between 0 and 1. There are three types of exceptions that we need to pay attention to. First, the individual can update *past* the advisor's θ^{A}_{post}. This would result in $\omega > 1$. Second, the individual could update in the *opposite direction* to the advisor's message. This would result in $\omega < 0$. Third, in some instances, the investor's prior will be equal to the advisor's message, i.e., $\theta^{I,0}_{post} = \theta^{A}_{post}$. In such instances, ω is *undefined*. Below, we discuss how we deal with these cases.

misaligned advisor in SYMMETRIC-PRIOR, as indicated by the α_0 coefficient estimates, which are small and insignificant in all specifications. Second, $\alpha_0+\alpha_2$ is negative and statistically significant in all four columns, indicating that investors are more skeptical of advice received from misaligned advisors than from aligned advisors in SYMMETRIC-DISCLOSURE and SYMMETRIC-WARNING. This provides statistical support for the visual evidence presented in Figure 4.

Result 5 (Belief updating: narratives from aligned and misaligned advisors)**.** *Investors update their beliefs less in response to narratives from misaligned in comparison to aligned advisors in SYMMETRIC-DISCLOSURE and SYMMETRIC-WARNING. In SYMMETRIC-PRIOR, investors update similarly in response to narratives from misaligned and aligned advisors.*

Table 3: Belief updating, by treatment and advisor alignment

	DISCLOSURE		WARNING	
	ω (1a)	ω (1b)	ω (2a)	ω (2b)
Misaligned (α_0)	-0.0210	-0.0310	-0.0176	-0.0329
	(0.0402)	(0.0375)	(0.0398)	(0.0374)
T (α_1)	0.198***	0.182***	0.143***	0.133***
	(0.0457)	(0.0448)	(0.0454)	(0.0453)
Misaligned×**T** (α_2)	-0.269***	-0.212***	-0.310***	-0.231***
	(0.0568)	(0.0502)	(0.0521)	(0.0475)
Treatments	Prior & Disc.	Prior & Disc.	Prior & Warn.	Prior & Warn.
Sample	$\omega \in [0,1]$	Censored	$\omega \in [0,1]$	Censored
$p(H_0 : \alpha_0+\alpha_2 \geq 0)$	<.001	<.001	<.001	<.001
Round × History FE	Yes	Yes	Yes	Yes
Observations	1450	1629	1464	1624

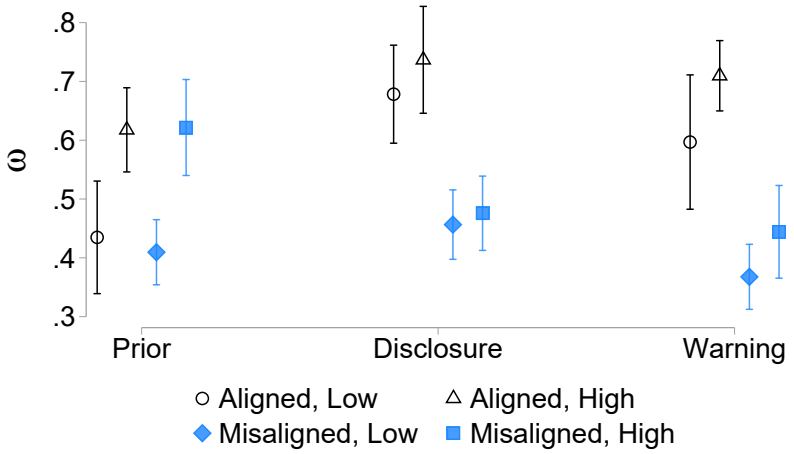
Notes: (i) The dependent variable is the belief updating weight, ω, at the individual update level; (ii) Each column contains data from a pair of treatments—the benchmark treatment, SYM-PRIOR, and another treatment: either SYM-DISCLOSURE or SYM-WARNING; (iii) We consider two ways of dealing with non-convex updates. Following our pre-registration document, our main specification in the (*a) columns excludes non-convex updates. In the (*b) columns, we censor updates where $\omega < 0$ or $\omega > 1$ at zero and one, respectively. We always exclude those where ω is undefined due to the investor's prior being equal to the advisor's message; (iv) Standard errors are clustered at the matching group level and are reported in parenthesis; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

We also consider a difference-in-differences test of whether advisor type disclosures or warnings increase investor skepticism toward narratives from misaligned advisors. The idea is to examine whether disclosures or warnings increase the gap in investor belief updating in response to advice from misaligned versus aligned advisors, by comparing SYMMETRIC-DISCLOSURE and SYMMETRIC-WARNING with SYMMETRIC-PRIOR. The negative and statistically significant coefficient on α_2 indicates that this gap is larger in the presence of disclosures and warnings. This implies that these interventions increase the difference in how investors update their beliefs in response to aligned versus misaligned advisors.

Result 6 (Belief updating: Influence of disclosures and warnings)**.** *Compared to* Symmetric-Prior*, the gap in belief updating between aligned and misaligned advisors is larger in (a)* Symmetric-Disclosure *and (b)* Symmetric-Warning*.*

Finally, we investigate how the fit of the advisor's message influences belief updating. Specifically, we examine whether fit matters for belief updating in each of the three treatments. Figure 5 provides a summary of the results from a regression analysis that includes advisor alignment, treatment, and fit, along with all interactions. We provide further details in Appendix A.4.

Figure 5: Average updating weight by treatment, alignment, and fit.



Notes: (i) The estimates in this figure are obtained from a single pooled regression using data from all three treatments. In Appendix A.4, we also report coefficient estimates from two separate regressions, each using data from Symmetric-Prior and one of the other treatments; (ii) Error bars represent 95% confidence intervals that were derived from regressions that cluster standard errors at the matching group level; (iii) This figure uses the censoring approach for dealing with non-convex ω values; (iv) Following our pre-registration plan, we define *relative fit* as the difference between the EPI of the advisor's narrative and the investor's prior, $\Delta\text{fit} = \text{EPI}^A - \text{EPI}^{I,0}$. We then define a high-fit message as one with a positive relative fit, namely $\mathbb{I}(\Delta\text{fit} > 0)$.

The left-most panel shows that narrative fit plays a crucial role in investor belief updating in the Symmetric-Prior treatment: investors update their beliefs more in response to narratives with a high fit. This supports our earlier findings on the centrality of fit in narrative persuasion.[21] The pattern holds for messages from both aligned (hollow markers) and misaligned advisors (solid markers), suggesting that, in Symmetric-Prior, fit influences investor behavior regardless of advisor type.

In the Symmetric-Disclosure and Symmetric-Warning treatments (middle and right-most panels), the effect of fit is weakened. Higher-fit messages still correspond to higher point estimates of ω, but the increase is significant at the 5% level in only one of four comparisons—misaligned advisors in Symmetric-Warning.

The evidence points to a shift from fit-based to incentive-alignment-based assessments once incentives are disclosed. In both Symmetric-Disclosure and Symmetric-Warning, investors are

[21] A comparison of fit-sensitivity in the Symmetric-Prior treatment and in the Competition treatment (discussed above) suggests that the magnitudes are also quantitatively similar. In Competition, we find that investors are 24 percentage points more likely to adopt a higher- rather than a lower-fitting narrative (see Table 1), which is comparable to the roughly 21 percentage point increase observed in Symmetric-Prior.

presented with salient information about advisor motives. As a result, they place greater weight on incentive alignment than on fit when evaluating advisor narratives. This is consistent with the S&S notion that disclosing that advisors hold misaligned incentives increases the fit threshold for narrative adoption, effectively making receivers more skeptical. In addition, the figure suggests that the reverse also holds: disclosing that advisors are fully aligned reduces skepticism.

Result 7 (Belief updating: Influence of narrative fit)**.** *In Symmetric-Prior, investors update their beliefs by roughly 21 pp more in response to high-fit messages in comparison to low-fit messages, irrespective of advisor alignment. In Symmetric-Disclosure and Symmetric-Warning, the corresponding "fit premium" ranges between 5-10 pp, which is significantly larger than zero in one out of the four possible comparisons.*

Exploring the limits of narrative persuasion: In addition to enabling us to study how individuals update their beliefs when they receive a narrative, the Belief Updating treatments were designed to test the boundaries of narrative persuasion. They do so by beginning with a setting in which the rational model predicts no persuasion—the Symmetric-Rep treatment—and then incrementally introducing elements that heighten the salience of reasons to distrust narratives from misaligned advisors.

Testing the limits of persuasion—To assess how effective persuasion is in each of the treatments, we consider an empirical specification that essentially pools together the two types of misaligned advisors. We do this by defining an *incentive-adjusted* classifier that takes a value of $+1$ for rounds where an investor is matched with an up-advisor, -1 when they are matched with a down-advisor, and 0 when the advisor is aligned. This allows us to compare the reported beliefs of investors matched with misaligned advisors (pooled) to those matched with aligned advisors, thereby estimating the distortion in investors' beliefs arising from advisor incentive misalignment.

$$\mathbb{C}(\text{Adv. type}) = \begin{cases} +1, & \text{if matched with Up-advisor} \\ 0, & \text{if matched with Aligned-advisor} \\ -1, & \text{if matched with Down-advisor} \end{cases} \tag{5}$$

With this classifier in hand, we specify the following regression equation:

$$\theta_{post}^{I,1} = \beta_0 \cdot \mathbb{C}(\text{Adv. type}) + \beta_1 \cdot \mathbf{T} + \beta_2 \cdot [\mathbb{C}(\text{Adv. type}) \times \mathbf{T}] + \gamma + \varepsilon. \tag{6}$$

We estimate this regression on three different samples, each including data from Symmetric-Rep and from one of the other three treatments (Symmetric-Prior, Symmetric-Disclosure, or Symmetric-Warning). To detect any differences in the impact of advisor misalignment by treatment, we interact $\mathbb{C}$(Adv. type) with a treatment indicator, $\mathbf{T}$, which takes a value of 0 for the Symmetric-Rep treatment and 1 otherwise. In addition, we control for Round×History fixed effects, denoted by γ.

The coefficients can be interpreted as follows. The β_0 coefficient captures what we refer to as the *persuasion gap*—the average difference in investor beliefs between those matched with

a misaligned versus an aligned advisor—in the Symmetric-Rep treatment. A positive estimate indicates that misaligned advisors shift investor beliefs toward their persuasion target. The β_2 coefficient captures the extent to which this gap changes across treatment conditions. If the intervention treatments reduce investor susceptibility to persuasion, the estimated β_2 will be negative.

When interpreting β_2 in this way, a key identifying assumption is that advisor behavior remains stable across treatments. Our design is carefully tailored to hold everything constant for advisors across treatments—advisors receive the same instructions, see the same company data, and are randomized into treatments simultaneously within each experimental session. Nevertheless, our data features some chance-driven imbalances in advisors' narratives.[22] To mitigate the impact of this imbalance, we present estimation results from both the full sample and a winsorized sample that excludes the 10% of advisors who sent the most extreme narratives (as measured by the average θ^A_{post} across the 10 rounds of the experiment). This specification ensures an even distribution of advisor narratives across treatments (details are in Appendix A.5). Although this winsorization deviates from our preregistration, we believe it offers valuable insights into investors' responses to the interventions, free from the confounding effects of outlier advisor behavior. Comparing the estimates from the full and winsorized samples allows us to assess the robustness of the results.

Table 4 presents the results from estimating Equation (6) on three different samples, each consisting of Symmetric-Rep and one other treatment: (1) Symmetric-Prior, (2) Symmetric-Disclosure, and (3) Symmetric-Warning. For each pair of treatments, we report the results in two columns: the (*a) columns exclude the 10% most "extreme" advisors, while the (*b) columns use the full sample.

The key coefficient of interest is β_2, which measures how the persuasive influence of misaligned advisors differs in each treatment relative to the benchmark Symmetric-Rep condition. The point estimate is negative in all specifications, suggesting that the interventions reduce persuasion. The (*a) columns present winsorized estimates. Column (1a) indicates that the persuasion gap in Symmetric-Prior decreases by 1.8 percentage points relative to Symmetric-Rep, although this decrease is not statistically significant. Columns (2a) and (3a) show that the gap is reduced by over 5 percentage points in Symmetric-Disclosure and Symmetric-Warning. This reduction is statistically significant and large in magnitude relative to the absolute gap size of approximately 6 percentage points in Symmetric-Rep (as indicated by β_0). The β_2 estimates in the (*b) columns, which use the full sample, are broadly similar but larger in magnitude, reflecting the influence of including extreme narratives. Taken together, the table suggests that encouraging investors to think about the data independently before meeting the advisor may have a small effect in reduc-

[22]For example, nine misaligned advisors in Symmetric-Rep sent consistently extreme narratives—those with an average $\theta^A_{post} \geq 80$ (up-advisor) or ≤ 20 (down-advisor) across the ten rounds. In the intervention treatments, the number of such extreme advisors is lower—four in Symmetric-Prior and Symmetric-Disclosure, and seven in Symmetric-Warning. This imbalance in advisors who systematically sent extreme narratives warrants attention, as such cases can disproportionately affect mean persuasion estimates. This motivates our robustness exercises with winsorizing in Table 4. It is important to note that this is less of a concern in the belief updating analysis because: (i) the updating weight ω is a relative measure and is therefore more robust to more extreme values of θ^A_{post}; and (ii) the majority of the imbalance stems from the Symmetric-Rep treatment, which is not part of the belief updating analysis. Nevertheless, we check the robustness of the belief updating results to winsorizing the sample and find that they remain qualitatively the same and quantitatively very similar. See Appendix A.4 for details.

ing the persuasion gap. Additionally disclosing the advisor's conflict of interest—with or without adding a warning about the potential consequences of their biased incentives—substantially reduces the gap.

Result 8 (Limits of narrative persuasion)**.** *In the baseline Symmetric-Rep treatment, misaligned advisors shift investor assessments toward their persuasion target by approximately 6 percentage points. Adding a private-reasoning stage (Symmetric-Prior) appears to have a small effect in reducing this "persuasion gap" (by approximately 2 percentage points; not statistically significant). Revealing the conflict of interest (Symmetric-Disclosure) or pairing the disclosure with an explicit bias warning (Symmetric-Warning) reduces the gap by approximately 5 percentage points—nearly eliminating the distortion in investors' beliefs caused by misaligned advisors.*

Table 4: The influence of interventions on persuasion (in the Symmetric scenario)

		Prior		Disclosure		Warning	
		$\theta^{I,1}_{post}$ (1a)	$\theta^{I,1}_{post}$ (1b)	$\theta^{I,1}_{post}$ (2a)	$\theta^{I,1}_{post}$ (2b)	$\theta^{I,1}_{post}$ (3a)	$\theta^{I,1}_{post}$ (3b)
$\mathbb{C}$(Adv. type)	(β_0)	5.822***	7.223***	6.025***	7.223***	5.934***	7.223***
		(1.084)	(1.190)	(1.035)	(1.190)	(1.045)	(1.190)
T	(β_1)	0.594	-0.463	0.941	-0.0967	1.313	0.424
		(1.094)	(1.178)	(1.049)	(1.014)	(1.051)	(1.086)
$\mathbb{C}$(Adv. type) × **T**	(β_2)	-1.817	-4.863***	-5.191***	-6.987***	-5.327***	-6.267***
		(1.542)	(1.812)	(1.327)	(1.491)	(1.489)	(1.546)
Incl. treatments		Symm. & Prior	Symm. & Prior	Symm. & Disc.	Symm. & Disc.	Symm. & Warn.	Symm. & Warn.
Incl. advisors		No extreme	All	No extreme	All	No extreme	All
Round × History FE		Yes	Yes	Yes	Yes	Yes	Yes
Observations		1570	1800	1650	1800	1640	1800

Notes: (i) The dependent variable is the investor's posterior belief, $\theta^{I,1}_{post}$, after receiving a narrative from an advisor; (ii) Each column contains data from a pair of treatments—the benchmark treatment, Symmetric-Rep, and another treatment: Sym-Prior, Sym-Disclosure, or Sym-Warning; (iii) The (*a) columns include observations from the winsorized sample (see Appendix A.5 for details) and the (*b) columns include all observations; (iv) Standard errors are clustered at the matching group level and are reported in parenthesis; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

Protection of investors—The results above show that investors are less susceptible to persuasion by misaligned advisors in the three intervention treatments. However, this is not equivalent to showing that investors are better off in these treatments; that depends on whether the quality of investors' decisions improves when they become more skeptical. Although we focus primarily on persuasion, Table 5 also explores investor welfare by examining the effect of the treatments on the distance between the investor's assessment and the truth, $|\theta^{I,1}_{post} - \theta^{T}_{post}|$. We restrict the analysis in the table to only include rounds with misaligned advisors and present the results using the winsorized sample in the (*a) columns and the full sample in the (*b) columns. The treatment effect estimates all yield negative coefficients—though the estimates for Warning are not statistically significant. This suggests that the interventions make investors slightly better off when they meet a misaligned advisor.

Replication in Asymmetric—So far, the discussion in this section has focused on intervention treatments that explore the limits of persuasion within the Symmetric scenario. We also implemented

a set of intervention treatments in the Asymmetric scenario. These Asymmetric intervention treatments included two interventions similar to those in Symmetric—a Prior treatment and a Disclosure intervention. Additionally, we introduced a third treatment, PrivateData, in which advisors do not observe the company dataset that investors see and therefore cannot perfectly tailor their advice to the data. They may, however, still draw on their expertise (i.e., their knowledge of the true underlying process) when constructing their messages.

Table 5: The influence of interventions on investor welfare (in the Symmetric scenario)

	Prior		Disclosure		Warning	
	$\lvert\theta^{I,1}_{post} - \theta^{T}_{post}\rvert$ (1a)	$\lvert\theta^{I,1}_{post} - \theta^{T}_{post}\rvert$ (1b)	$\lvert\theta^{I,1}_{post} - \theta^{T}_{post}\rvert$ (2a)	$\lvert\theta^{I,1}_{post} - \theta^{T}_{post}\rvert$ (2b)	$\lvert\theta^{I,1}_{post} - \theta^{T}_{post}\rvert$ (3a)	$\lvert\theta^{I,1}_{post} - \theta^{T}_{post}\rvert$ (3b)
T	-3.336***	-3.083***	-1.748*	-2.400**	-0.822	-1.553
	(0.984)	(1.061)	(0.981)	(1.063)	(1.128)	(1.131)
Incl. treatments	Symm. & Prior	Symm. & Prior	Symm. & Disc.	Symm. & Disc.	Symm. & Warn.	Symm. & Warn.
Incl. advisors	No extreme	All	No extreme	All	No extreme	All
Round × History FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1010	1200	1060	1200	1040	1200

Notes: (i) The dependent variable is the investor's distance from the truth, $\lvert\theta^{I,1}_{post} - \theta^{T}_{post}\rvert$, after receiving a narrative from an advisor; (ii) Each column contains data from a pair of treatments—the benchmark treatment, Symmetric-Rep, and another treatment: Sym-Prior, Sym-Disclosure, or Sym-Warning; (iii) The (*a) columns include observations from the winsorized sample (see Appendix A.5 for details) and the (*b) columns include all observations; (iv) Standard errors are clustered at the matching group level and are reported in parenthesis; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

Appendix Section A.6 provides a detailed description of these treatments and presents the results from replicating the analysis above, examining persuasion and investor welfare for these Asymmetric interventions. The main takeaways are: (i) consistent with the Symmetric scenario, the Asymmetric interventions reduce persuasion—the β_2 coefficients are negative for all three treatments in Table A.12 (though not statistically significant for Asymmetric-PrivateData); and (ii) unlike in the Symmetric treatments, the interventions in the Asymmetric scenario do not appear to improve investor welfare on average.

This latter result may reflect fundamental differences in how skepticism toward misaligned advisors operates across the two scenarios. In particular, variation in advisor knowledge and behavior may influence how skepticism affects investor outcomes. Advisors in Asymmetric are *experts*—they know the truth—whereas advisors in Symmetric are no better informed than investors. Moreover, misaligned advisors in Asymmetric tell the truth 25% of the time, despite their incentives. This indicates heterogeneity in advisors' intent to shift investor beliefs away from the truth—it suggests that some advisors display an aversion to lying. We can think of misaligned advisors as falling along a spectrum: from those who aim to help investors move closer to the truth (well-meaning) to those who try to move their beliefs away from the truth (self-interested). Consequently, increased skepticism toward misaligned advisors has different implications in the two settings. In Symmetric, skepticism benefits investors because even well-meaning advisors do not know the truth. In contrast, in Asymmetric, skepticism has two countervailing effects: it helps investors guard against self-interested misaligned advisors but may harm them when it leads to disregarding truthful advice from lying-averse misaligned advisors. These opposing effects appear to offset one another in the Asymmetric scenario.

Taken together, the results from the SYMMETRIC and ASYMMETRIC interventions show that interventions can successfully increase investor skepticism toward misaligned advisors and reduce persuasion. However, the extent to which this benefits investors depends on their ability to make accurate assessments on their own, and on whether increased skepticism causes them to dismiss potentially valuable advice from well-meaning yet misaligned advisors with greater expertise.

6 Extensions and Robustness Exercises

In the previous section, we presented our main results. To explore the robustness and underlying mechanisms of narrative persuasion, we now discuss several additional exercises.

Attention: One concern that can be raised regarding the results above is that it is possible that they are partially driven by inattentive participants—for example, investors who naively follow advisor advice. To investigate this concern more directly, we conduct several exercises. First, we consider the possibility that investors are unthinkingly following advisors' messages. We show that only a small fraction of investors (2% in ASYMMETRIC and 6% in SYMMETRIC) *fully adopts* the advisor's message in all ten rounds—i.e., by setting their final assessment equal to the advisor's suggested value (i.e., $\theta^{I,1}_{\text{post}} = \theta^{A}_{\text{post}}$). We also demonstrate that persuasion using narratives (Result 1) is robust to excluding investors who adopt the advisor's message in eight or more rounds (see Table A.14 in Appendix Section A.7).

Second, we investigate whether our results could be driven by participants who either rushed through the experiment or took unusually long to complete it. We replicate Result 1 using specifications that exclude either the 25% fastest or the 25% slowest participants (based on average decision time across the ten rounds). In both cases, the results remain very similar, indicating that they are not driven by participants who progressed particularly quickly or slowly through the experiment (see Table A.15 in Appendix Section A.7). We provide further details and discussion of participant attention in Appendix Section A.7. Taken together, these exercises suggest that our findings are unlikely to be driven by inattentive participants.

Explanations: 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.

Table 6 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 for whether an explanation was provided or not. Both the prior and posterior distance metrics reflect 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, $\theta^{I,1}_{post}$. 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}.

Table 6: 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 advisor FE	Yes	Yes	Yes
Prior Distance 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) Prior Distance fixed effects control for the prior distance in a nonparametric way by introducing dummy variables for each distinct integer value of prior distance available in the data; (iv) 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); (v) The sample contains data from all investors in EXPLANATION and NOEXPLANATION; (vi) For each investor we have 10 observations—one for each round; (vii) Standard errors are clustered at the investor level and are reported in parentheses; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

As one would expect, 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).[23] We find that the 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}. In contrast, 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,

[23]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:

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

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. This is consistent with the prediction in S&S that improving the auxiliary parameter fit increases persuasion. It is inconsistent with a broad class of (behavioral) cheap talk models predicting that variation in the auxiliary parameters is irrelevant for persuasion. We discuss these treatments, the associated analysis, and the relation to cheap talk in more detail in Appendix Section A.8.

Decision noise: Here, we relax the assumption that decision-makers (advisors and investors) are perfectly consistent when choosing between narratives, acting without error when assessing relative fit. To do this, we use the COMPETITION data to estimate "fit perception 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. In line with our earlier discussion, our model estimates a significant amount of noise in narrative adoption. In addition, we find that advisors appear to anticipate the noise in investors' decisions, accounting for it when constructing their narratives. This exercise suggests that advisors in the experiment do not only consider relative fit but also take into account what would happen if the investor adopted the competing narrative. The intuition is as follows. With noisy adoption, even a higher-fit narrative might be rejected. Advisors, therefore, hedge by increasing their own narrative's fit further as the competing narrative moves away from their persuasion target (since the cost of the alternative being chosen is larger). This increase in fit comes at the expense of narrative bias. Our discrete choice model estimates suggest that this hedging motive helps explain how advisors construct narratives in the experiment. This exercise is discussed in Appendix Section A.9.

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 always fully adopted or not adopted, but still influence the investor's beliefs. For example, an investor who does not fully adopt the narrative sent to him by an advisor 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. To explore this idea, in Appendix A.10, 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.[24] Therefore, while an

[24]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.

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.

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. Because the process generating successes and failures is random, the true DGP is unlikely to coincide exactly with 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 is not present since receivers know that advisors do not know the true DGP; therefore, a narrative that fits perfectly is not a signal of bias. While our experiment was not designed to investigate such skeptical thinking, we explore evidence related to this idea in Appendix A.11. There, we 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.

A closer look at belief updating, fit and incentive disclosure: In this section, we briefly discuss the results from two additional exercises that examine the influence of fit and incentive disclosure on belief updating (discussed in more detail in Appendix A.12).

First, we use a continuous measure of relative fit to replicate the result from Figure 5. Consistent with our earlier findings, belief updating increases with fit, and sensitivity to fit is pronounced in SYMMETRIC-PRIOR but muted in the "disclosed incentives" treatments. The analysis also reveals that the average updating weight placed on the narrative increases gradually as the narrative fit surpasses the fit of the prior; it does not jump discontinuously. This exercise confirms that our belief updating results are robust to replacing the binary fit variable with a continuous measure of relative fit. The findings also align with earlier results suggesting that individuals perceive relative fit with noise, leading to smooth, rather than abrupt changes in the updating weight. These results are consistent with a version of S&S fit-based adoption under noisy fit perception.

Second, we investigate why investors update toward narratives from misaligned advisors even when their incentives are disclosed (i.e., in the SYMMETRIC-DISCLOSURE and SYMMETRIC-WARNING treatments). In Appendix A.12, we provide evidence suggesting that the *apparent self-interest* of advisors plays a role in such cases. Specifically, when a narrative from a misaligned advisor runs counter to the advisor's self-interest (relative to the investor's prior), investors are much more likely to update their beliefs than when the narrative aligns with the advisor's self-

interest. This suggests that the non-zero belief updating observed after disclosing misalignment does not necessarily reflect residual confusion. Instead, investors distinguish between different types of misaligned narratives; they are more cautious when adoption would move them toward the advisor's persuasion target.

7 Concluding Discussion

The discussion above provides empirical evidence showing how narratives can be used as a tool for persuasion. Our analysis is relevant for the class of situations where there exists some public information that can accommodate more than one possible interpretation, allowing some individuals 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. Turning to advisors, a key feature of narrative persuasion is that advisors can construct narratives *ex-post*, tailoring them to the public data. Advisors 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.

These findings underscore the value of broadening our understanding of persuasion beyond expertise-based accounts. Coherent narratives can exert persuasive power even when the sender lacks private information. This form of persuasion is highly relevant in real-world settings where individuals seek to shape how others interpret shared evidence. Such contexts include politicians framing public events, entrepreneurs pitching to investors, and citizens debating which interventions curbed the spread of COVID-19. For such contexts, our findings offer a potential explanation for why people do not simply "listen to the experts" when forming beliefs about the world: narrative fit can be a powerful source of credibility, shaping both which narratives are persuasive and how communicators construct them. Our findings on the limits of persuasion further suggest that narrative fit is particularly relevant in everyday settings in which the incentives of the speaker are undisclosed.

It is important to acknowledge that we consider a stylized setting. The advantage of this is that we are able to provide a clean causal test of theoretical predictions emerging from the S&S framework. However, this controlled approach necessarily abstracts away from other aspects of real-world communication that may also matter for persuasion. In natural settings, senders often select from a less constrained set of narratives or can combine narratives with additional

elements when trying to persuade. For example, they may disclose data alongside a narrative that interprets it, draw connections across multiple variables, or embed emotional appeals. While important, these factors lie beyond the scope of this paper; future work could extend our results by exploring richer environments and additional mechanisms of persuasion.

References

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.

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 (2024a). Narrative persuasion. *WZB Discussion Paper SP II 2023–301r*.

Barron, K. and T. Fries (2024b). Narrative Persuasion: A Brief Introduction. *Encyclopedia of Experimental Social Science*.

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.

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.

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

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.

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.

Fong, M.-J., P.-H. Lin, and T. R. Palfrey (2025, August). Cursed Sequential Equilibrium. *American Economic Review 115*(8), 2616–2658.

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., C. Roth, and F. Zimmermann (2024, October). Stories, statistics, and memory. *The Quarterly Journal of Economics 139*(4), 2181–2225.

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)*.

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.

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

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

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*.

Jehiel, P. (2022). Analogy-based expectation equilibrium and related concepts: Theory, applications, and beyond. *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.

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

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.

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 (2025). Narratives and Valuations. *Management Science 71*(6), 5376–5395.

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

Schwartzstein, J. and A. Sunderam (2020). Using Models to Persuade. *NBER Working Paper No. 26109*.

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

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

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 Additional Results

A.1 Participant Demographics

Tables A.1, A.2, and A.3 below present summary statistics of the demographics of the participants in the experiment. In general, the demographics are fairly balanced within each data collection wave. Accross waves, we observe that our sample is slightly younger in the 2022 (wave 1) sample than in the 2023 (wave 2) and 2025 (wave 3) sample. Furthermore, our 2025 sample is slightly more female than the earlier waves. However, it is important to keep in mind that the tests in our main analysis relies on within-wave variation. Therefore, we do not see these cross-wave imbalances as a threat to our main results.

Table A.1: Demographic characteristics of participants in 2022 (by treatment and role)

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

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) Standard deviations in parenthesis.

The table below displays the 2023 sample characteristics. It contains no information about advisors in the two Explanation treatments. The reason for this is that we re-used the advisor messages from Asym-Prior for the two Explanation treatments. For more details, please refer to the description of the experimental design in Appendix A.8.

Table A.2: Demographic characteristics of participants in 2023 (by treatment and role)

	Symmetric		Explanation	NoExplanation
	Advisor mean/sd	Investor mean/sd	Investor mean/sd	Investor mean/sd
age	41.667	41.783	40.822	40.606
	(13.398)	(14.652)	(13.726)	(12.998)
Gender: Female	0.494	0.500	0.489	0.506
	(0.501)	(0.501)	(0.501)	(0.501)
Gender: Male	0.494	0.489	0.500	0.494
	(0.501)	(0.501)	(0.501)	(0.501)
Gender: Other	0.011	0.011	0.011	0.000
	(0.105)	(0.105)	(0.105)	(0.000)
Edu: Primary school	0.000	0.000	0.006	0.000
	(0.000)	(0.000)	(0.075)	(0.000)
Edu: Secondary school	0.094	0.117	0.122	0.083
	(0.293)	(0.322)	(0.328)	(0.277)
Edu: Higher secondary education	0.200	0.144	0.250	0.206
	(0.401)	(0.353)	(0.434)	(0.405)
Edu: College or university	0.439	0.528	0.389	0.489
	(0.498)	(0.501)	(0.489)	(0.501)
Edu: Post-graduate	0.261	0.211	0.233	0.222
	(0.440)	(0.409)	(0.424)	(0.417)
Edu: Prefer not to say	0.006	0.000	0.000	0.000
	(0.075)	(0.000)	(0.000)	(0.000)
Observations	180	180	180	180

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 Asym-Prior treatment and only recruited new investors; (iii) Standard deviations in parenthesis.

Table A.3: Demographic characteristics of participants in 2025 (by treatment and role)

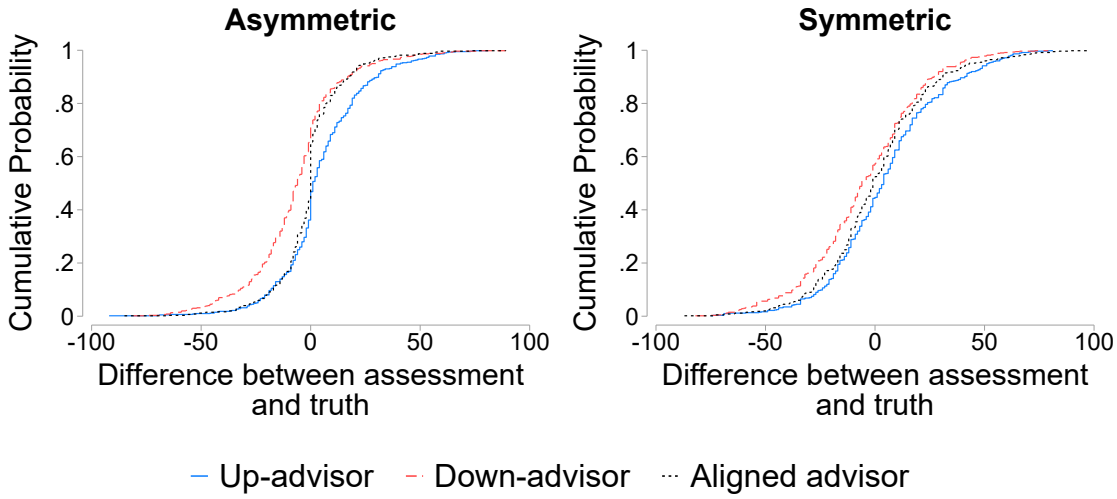
	Symmetric-Rep		Sym-Prior		Sym-Disclosure		Sym-Warning	
	Advisor mean/sd	Investor mean/sd	Advisor mean/sd	Investor mean/sd	Advisor mean/sd	Investor mean/sd	Advisor mean/sd	Investor mean/sd
age	36.922	38.589	40.444	39.044	37.022	41.089	39.533	41.233
	(11.818)	(12.823)	(11.993)	(12.352)	(13.445)	(13.306)	(12.942)	(13.742)
Gender: Female	0.633	0.600	0.622	0.567	0.700	0.656	0.578	0.544
	(0.485)	(0.493)	(0.488)	(0.498)	(0.461)	(0.478)	(0.497)	(0.501)
Gender: Male	0.356	0.389	0.378	0.422	0.289	0.344	0.411	0.444
	(0.481)	(0.490)	(0.488)	(0.497)	(0.456)	(0.478)	(0.495)	(0.500)
Gender: Other	0.011	0.011	0.000	0.011	0.011	0.000	0.011	0.011
	(0.105)	(0.105)	(0.000)	(0.105)	(0.105)	(0.000)	(0.105)	(0.105)
Edu: Primary school	0.011	0.000	0.000	0.000	0.011	0.000	0.000	0.000
	(0.105)	(0.000)	(0.000)	(0.000)	(0.105)	(0.000)	(0.000)	(0.000)
Edu: Secondary school	0.067	0.067	0.133	0.111	0.067	0.100	0.100	0.111
	(0.251)	(0.251)	(0.342)	(0.316)	(0.251)	(0.302)	(0.302)	(0.316)
Edu: Higher secondary education	0.189	0.200	0.156	0.211	0.189	0.244	0.222	0.144
	(0.394)	(0.402)	(0.364)	(0.410)	(0.394)	(0.432)	(0.418)	(0.354)
Edu: College or university	0.489	0.433	0.500	0.478	0.489	0.400	0.422	0.489
	(0.503)	(0.498)	(0.503)	(0.502)	(0.503)	(0.493)	(0.497)	(0.503)
Edu: Post-graduate	0.244	0.300	0.211	0.200	0.244	0.244	0.256	0.233
	(0.432)	(0.461)	(0.410)	(0.402)	(0.432)	(0.432)	(0.439)	(0.425)
Edu: Prefer not to say	0.000	0.000	0.000	0.000	0.000	0.011	0.000	0.022
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.105)	(0.000)	(0.148)
Observations	90	90	90	90	90	90	90	90

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) Standard deviations in parenthesis.

A.2 Investor Behavior

Figure A.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 dotted black, solid blue, and dashed 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 A.1: CDF of distance between investors' beliefs and the truth (by advisor type)



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 A.4 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 A.4 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 A.4: 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 A.5, 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, in our Belief Updating treatments, we are also able to make use of the fact that we elicited investors' prior beliefs before they meet their advisor. 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. The results from these exercises are discussed in Section 5.3 and Appendix Section A.4. In addition, we have one treatment in our Asymmetric context in which we also elicit the investors' prior beliefs before meeting the advisor (our Asym-Prior treatment). Appendix Section A.2 of a previous working paper version (Barron and Fries, 2024a) discusses the results from this treatment (referred to as the InvestorPrior treatment in that version). The results from all of these exercises are consistent, showing that fit matters for narrative adoption.

Table A.5: 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)	-14.62***	-11.14***
	(1.897)	(2.044)
Misaligned advisor = 1	0.691	0.0485
	(0.668)	(1.100)
Treatment	Asymmetric	Symmetric
Round FE	No	Yes
Round × History FE	Yes	No
Observations	1800	1800

A.3 Advisor Behavior

Table A.6 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) regresses 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 A.6: 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

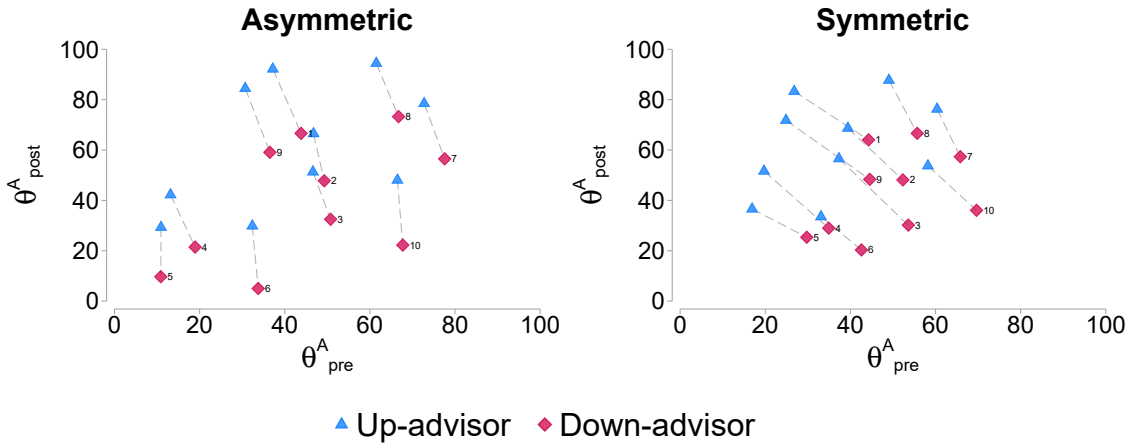
Notes: (i) The dependent variable is either the θ^A_{post} (odd columns) or the θ^A_{pre} (even columns) sent by the advisor, (ii) Columns (1a) and (1b) contain data from all advisors who received the ASYMMETRIC instructions (advisors in DISCLOSURE and INVESTORPRIOR also received the ASYMMETRIC instructions 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 3—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 A.2 shows how up-advisors (blue triangle markers) and down-advisors (red diamond 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 diamond marker. We see that, in each of the ten rounds, the diamond marker is to the south-east of the triangle 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 A.2: Advisor θ_{post} and θ_{pre} reports in each round (by advisor type)



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 triangle markers show the average θ_{post} and θ_{pre} sent by up-advisors in each round, while the red diamond 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 triangle and diamond markers should coincide.

A.4 Belief Updating

In Section 5.3 of the main text, we analyze how the fit of advisors' narrative influences belief updating in the SYMMETRIC-PRIOR, SYMMETRIC-DISCLOSURE, and SYMMETRIC-WARNING treatments. The key results are summarized visually in Figure 5. Here, in this section, we provide further details regarding the preregistered empirical specification and regression output associated with that analysis. We also present the results from several additional exercises that explore the relationship between fit and belief updating in more depth.

To assess whether investors respond to the narrative fit differently in the three treatments, we estimate a regression that examines whether the belief updating weights, ω, differ systematically according to treatment, narrative fit, and advisor alignment. To evaluate this, we distinguish between messages with high and low relative fit, using the empirical fit index (EPI) to define *relative fit* as the difference between the EPI of the advisor's narrative and the investor's prior: $\Delta\text{fit} = \text{EPI}^A - \text{EPI}^{I,0}$. Using this, we augment equation 4 to incorporate fit and estimate the following specification:

$$\begin{aligned}\omega = {} & \alpha_0 \cdot I_1 + \alpha_1 \cdot \mathbf{T} + \alpha_2 \cdot I_2 \\ & + \alpha_3 \cdot (I_1 \times \mathbf{T}) + \alpha_4 \cdot (I_1 \times I_2) + \alpha_5 \cdot (\mathbf{T} \times I_2) \\ & + \alpha_6 \cdot (I_1 \times \mathbf{T} \times I_2) \\ & + \gamma + \varepsilon \end{aligned} \tag{7}$$

where:

- $I_1 = \mathbb{I}(\text{adv.} = \text{misaligned})$ is an indicator for whether the advisor is misaligned.
- $I_2 = \mathbb{I}(\Delta\text{fit} > 0)$ is an indicator variable for relative fit.

- **T** is a treatment indicator.

The results are reported in Table A.7. These results correspond to those displayed in Figure 5, which are discussed in the main text. One exception to this is that Figure 5 is generated using a single regression that pools the data from the Symmetric-Prior, Symmetric-Disclosure, and Symmetric-Warning treatments. In contrast, Table A.7 follows our preregistration and reports regressions that contain pairs of treatments. The results are consistent across the two approaches. The reason for this choice is simply that it is unwieldy to report a table with even more rows, while the single pooled regression seems most appropriate for generating the figure.

Table A.7: Belief updating, by treatment, advisor alignment, and fit

		Disclosure		Warning	
		ω	ω	ω	ω
Misaligned	(α_0)	-0.0176	-0.0310	-0.00757	-0.0220
		(0.0524)	(0.0516)	(0.0517)	(0.0510)
T	(α_1)	0.246***	0.240***	0.181**	0.161**
		(0.0634)	(0.0659)	(0.0775)	(0.0784)
$\mathbb{I}(\Delta\text{EPI} > 0)$	(α_2)	0.190***	0.172***	0.196***	0.184***
		(0.0584)	(0.0598)	(0.0581)	(0.0603)
Misaligned×**T**	(α_3)	-0.249***	-0.190**	-0.303***	-0.214**
		(0.0762)	(0.0738)	(0.0836)	(0.0821)
Misaligned×$\mathbb{I}(\Delta\text{EPI} > 0)$	(α_4)	0.0389	0.0408	0.0243	0.0169
		(0.0769)	(0.0751)	(0.0782)	(0.0770)
T × $\mathbb{I}(\Delta\text{EPI} > 0)$	(α_5)	-0.105	-0.119	-0.0908	-0.0698
		(0.0853)	(0.0931)	(0.0960)	(0.0960)
Misaligned×**T** × $\mathbb{I}(\Delta\text{EPI} > 0)$	(α_6)	-0.0722	-0.0796	-0.0261	-0.0427
		(0.115)	(0.117)	(0.110)	(0.110)
Treatments		Prior & Disc.	Prior & Disc.	Prior & Warn.	Prior & Warn.
Sample		$\omega \in [0,1]$	Censored	$\omega \in [0,1]$	Censored
$p(H_0 : \alpha_2 + \alpha_4 \leq 0)$		<.001	<.001	<.001	<.001
$p(H_0 : \alpha_2 + \alpha_5 \leq 0)$		.094	.228	.076	.055
$p(H_0 : \alpha_2 + \alpha_4 + \alpha_5 + \alpha_6 \leq 0)$		.103	.374	.006	.016
Round × History FE		Yes	Yes	Yes	Yes
Observations		1450	1629	1464	1624

Notes: (i) The dependent variable is the belief updating weight, ω, at the individual update level; (ii) Columns (1) and (2) use data from the Symmetric-Prior and Symmetric-Disclosure treatments, while columns (3) and (4) use data from Symmetric-Prior and Symmetric-Warning; (iii) We consider two ways to deal with non-convex updates. Following our pre-registration document, our main specification in columns (1) and (3) excludes non-convex updates. In columns (2) and (4), we censor updates where $\omega < 0$ or $\omega > 1$. We always exclude those where ω is undefined due to the investor's prior being equal to the advisor's message; (iv) Standard errors are clustered at the matching group level and are reported in parenthesis; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

Robustness to excluding "extreme" advisors— As argued in Section 5.3, it is worthwhile to check the robustness of the results from the Belief Updating treatments by using a winsorized sample that omits advisors who consistently send extreme messages (see Appendix A.5 for further details on the winsorization procedure). We report these robustness checks here.

Table A.8 replicates Table 3 from the main text using the winsorized sample that excludes extreme advisors. The coefficient estimates and significance test results remain similar when focusing on this sample. Therefore, the results are robust to winsorizing the sample.

Table A.8: Belief updating, by treatment and advisor alignment (winsorized sample)

		Disclosure		Warning	
		ω	ω	ω	ω
Misaligned	(α_0)	-0.0195	-0.0332	-0.00858	-0.0273
		(0.0472)	(0.0433)	(0.0461)	(0.0422)
T	(α_1)	0.197***	0.176***	0.138***	0.127***
		(0.0490)	(0.0476)	(0.0485)	(0.0478)
Misaligned×**T**	(α_2)	-0.280***	-0.216***	-0.311***	-0.219***
		(0.0644)	(0.0569)	(0.0578)	(0.0512)
Treatments		Prior & Disc.	Prior & Disc.	Prior & Warn.	Prior & Warn.
Sample		$\omega \in [0,1]$	Censored	$\omega \in [0,1]$	Censored
$p(H_0 : \alpha_0 + \alpha_2 \geq 0)$		<.001	<.001	<.001	<.001
Round × History FE		Yes	Yes	Yes	Yes
Observations		1286	1446	1283	1431

Notes: (i) The dependent variable is the belief updating weight, ω, at the individual update level; (ii) Each column contains data from a pair of treatments—the benchmark treatment, Sym-Prior, and another treatment: either Sym-Disclosure or Sym-Warning; (iii) We consider two ways to deal with non-convex updates. Following our pre-registration document, our main specification in the (*a) columns excludes non-convex updates. In the (*b) columns, we censor updates where $\omega < 0$ or $\omega > 1$ at zero and one, respectively. We always exclude those where ω is undefined due to the investor's prior being equal to the advisor's message; (iv) Standard errors are clustered at the matching group level and are reported in parenthesis; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

Table A.9 replicates Table A.7 from the main text using the winsorized sample that excludes extreme advisors. The coefficient estimates and significance test results remain similar when focusing on this sample. Therefore, the results are robust to winsorizing the sample.

Table A.9: Belief updating, by treatment, advisor alignment, and fit (winsorized sample)

		Disclosure		Warning	
		ω	ω	ω	ω
Misaligned	(α_0)	-0.0359	-0.0601	-0.0124	-0.0391
		(0.0681)	(0.0666)	(0.0652)	(0.0640)
T	(α_1)	0.239***	0.223***	0.166*	0.147*
		(0.0692)	(0.0729)	(0.0852)	(0.0861)
$\mathbb{I}(\Delta\text{EPI} > 0)$	(α_2)	0.143**	0.127*	0.158**	0.145**
		(0.0658)	(0.0690)	(0.0658)	(0.0681)
Misaligned×**T**	(α_3)	-0.251***	-0.170**	-0.294***	-0.181*
		(0.0881)	(0.0849)	(0.0938)	(0.0906)
Misaligned×$\mathbb{I}(\Delta\text{EPI} > 0)$	(α_4)	0.0765	0.0910	0.0502	0.0607
		(0.0875)	(0.0866)	(0.0881)	(0.0869)
T × $\mathbb{I}(\Delta\text{EPI} > 0)$	(α_5)	-0.0701	-0.0819	-0.0528	-0.0370
		(0.0893)	(0.101)	(0.100)	(0.101)
Misaligned×**T** × $\mathbb{I}(\Delta\text{EPI} > 0)$	(α_6)	-0.0892	-0.122	-0.0476	-0.0870
		(0.124)	(0.128)	(0.119)	(0.118)
Treatments		Prior & Disc.	Prior & Disc.	Prior & Warn.	Prior & Warn.
Sample		$\omega \in [0,1]$	Censored	$\omega \in [0,1]$	Censored
$p(H_0 : \alpha_2 + \alpha_4 \leq 0)$		<.001	<.001	<.001	<.001
$p(H_0 : \alpha_2 + \alpha_5 \leq 0)$		.135	.273	.078	.069
$p(H_0 : \alpha_2 + \alpha_4 + \alpha_5 + \alpha_6 \leq 0)$		.083	.388	.008	.034
Round × History FE		Yes	Yes	Yes	Yes
Observations		1286	1446	1283	1431

Notes: (i) The dependent variable is the belief updating weight, ω, at the individual update level; (ii) Columns (1) and (2) use data from the Symmetric-Prior and Symmetric-Disclosure treatments, while columns (3) and (4) use data from Symmetric-Prior and Symmetric-Warning; (iii) We consider two ways to deal with non-convex updates. Following our pre-registration document, our main specification in columns (1) and (3) excludes non-convex updates. In columns (2) and (4), we censor updates where $\omega < 0$ or $\omega > 1$. We always exclude those where ω is undefined due to the investor's prior being equal to the advisor's message; (iv) Standard errors are clustered at the matching group level and are reported in parenthesis; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

A.5 Limits of Narrative Persuasion (Symmetric scenario)

As discussed in Section 5.3, Table 4 presents results from specification (6), using the whole sample and a winsorized sample. Here, we explain the winsorizing procedure and discuss reasons for doing so.

First, our starting point for considering adjusting for advisor behavior is the observation that advisor messages are differentially extreme across treatments. This is easily illustrated by a regression of the form

$$\theta^A_{post} = \beta_0 \cdot \mathbb{C}(\text{Adv. type}) + \beta_1 \cdot \mathbf{T} + \beta_2 \cdot [\mathbb{C}(\text{Adv. type}) \times \mathbf{T}] + \gamma + \varepsilon. \quad (8)$$

The right-hand side of this regression is equivalent to specification (6) but the outcome variable is now equal to the advisor's θ^A_{post}. Such a regression measures the average message bias of a misaligned advisor relative to aligned in Symmetric-Rep (β_0) and how this message bias changes between treatments (β_2). Note that we would expect the treatment difference to be equal to zero—all advisors received the same instructions and data, and were randomly assigned to treatment within-session. This is not what we see in the data; Table A.10 illustrates how misaligned advisors in Symmetric-Rep send more extreme messages than in the intervention treatments.

Table A.10: Average message bias across treatments in the full sample

		(1) θ^A_{post}	(2) θ^A_{post}	(3) θ^A_{post}
$\mathbb{C}$(Adv. type)	(β_0)	13.12***	13.12***	13.12***
		(1.780)	(1.780)	(1.780)
T	(β_1)	-3.581*	-0.830	1.157
		(1.882)	(1.509)	(1.621)
$\mathbb{C}$(Adv. type) × **T**	(β_2)	-4.948*	-5.453**	-2.613
		(2.805)	(2.419)	(2.363)
Sample		Symm. & Prior	Symm. & Disc.	Symm. & Warn.
Incl. advisors		All	All	All
Round × History FE		Yes	Yes	Yes
Observations		1800	1800	1800

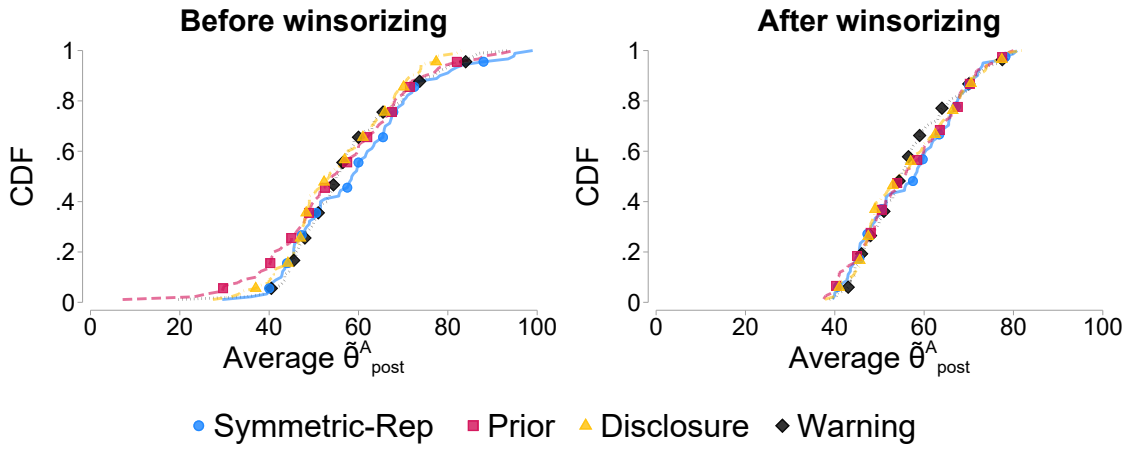
Notes: (i) The dependent variable is the advisor's θ^A_{post} contained in the narrative; (ii) Each column contains data from a pair of treatments—the benchmark treatment, Symmetric-Rep, and another treatment: Sym-Prior, Sym-Disclosure, or Sym-Warning; (iii) Observations from the whole sample are used in the regressions; (iv) Standard errors are clustered at the matching group level and are reported in parenthesis; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

Note that this effect can be driven by more misaligned advisors in Symmetric-Rep sending a θ^A_{post} closer to their persuasion target or by fewer misaligned advisors sending a θ^A_{post} further away from their persuasion target. To identify both types of deviations, we define an incentive-adjusted θ^A_{post} which is equal to θ^A_{post} for an up-advisor or an aligned advisor or equal to $1-\theta^A_{post}$ for a down-advisor. We denote this normalized θ^A_{post} by $\widetilde{\theta}^A_{post}$. We then average these $\widetilde{\theta}^A_{post}$ on the advisor level across 10 rounds. This is important, as single extreme narratives may be driven by extreme data,

while we aim to identify advisors with a stable proclivity for extreme narratives *regardless* of the data. An misaligned advisor with a large average $\widetilde{\theta}^{A}_{post}$ thus consistently sends a narrative close to the persuasion target and a misaligned advisor with a small average $\widetilde{\theta}^{A}_{post}$ consistently sends something that is far away.

Figure A.3 shows how winsorizing yields more similar distributions. Pre winsorizing, advisors in PRIOR tend to be less extreme and advisors in SYMMETRIC-REP more extreme. These differences are largely evened out in the winsorized sample that drops 5% of advisors with the lowest average $\widetilde{\theta}^{A}_{post}$ and 5% with the highest average $\widetilde{\theta}^{A}_{post}$.

Figure A.3: Consequence of winsorizing for the distribution of advisor narratives



Notes: The panels show the cumulative distribution of the average $\tilde{\theta}^{A}_{post}$. The markers on each CDF are located at the median value within each decile of the distribution.

This procedure has the intended effect of removing differences in the average message bias across treatments. Table A.11 displays estimation results of specification (8) using the winsorized sample. We observe that the treatment differences are estimated to be small and insignificant. Because of this, we prefer the winsorized treatment effect estimates reported in Table 4. They are more reflective of changes in investor behavior.

Table A.11: Average message bias across treatments in the winsorized sample

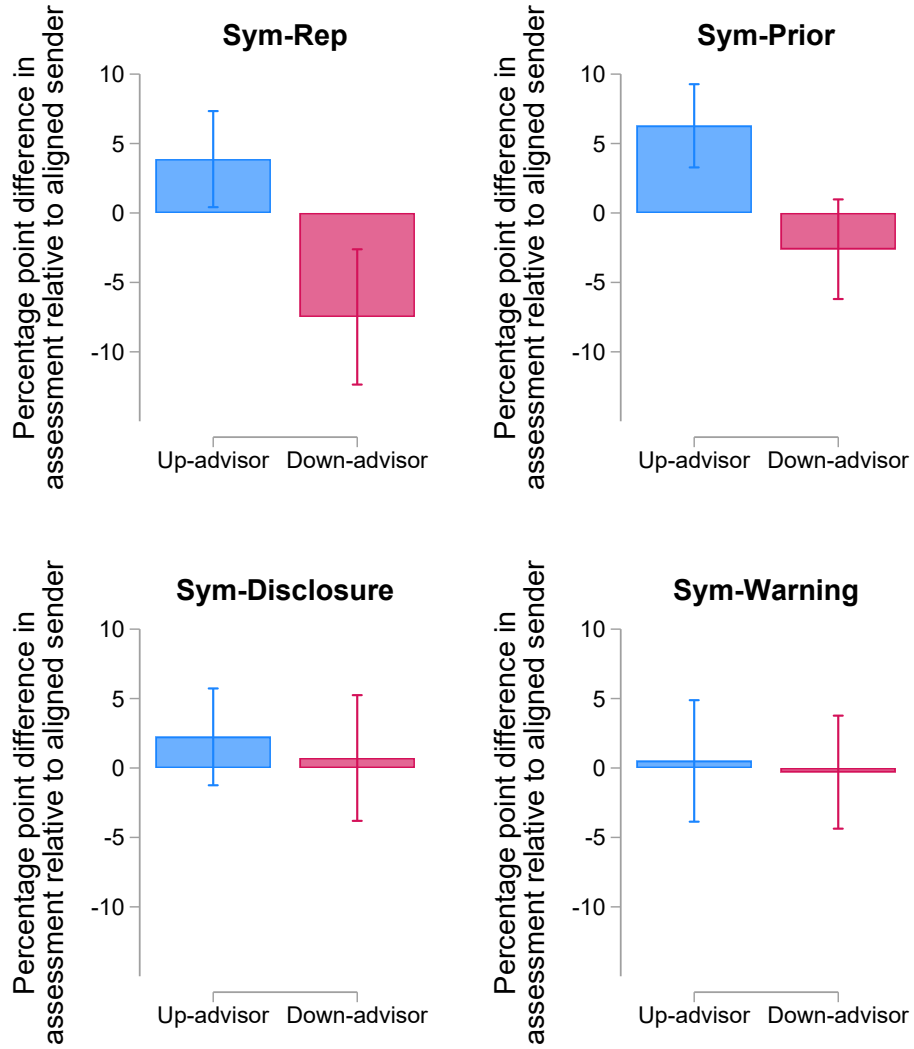
		(1) θ^A_{post}	(2) θ^A_{post}	(3) θ^A_{post}
$\mathbb{C}$(Adv. type)	(β_0)	10.44***	10.71***	10.71***
		(1.407)	(1.364)	(1.364)
T	(β_1)	-1.336	1.646	2.028
		(1.424)	(1.381)	(1.377)
$\mathbb{C}$(Adv. type) × **T**	(β_2)	0.340	-2.076	-1.354
		(1.992)	(1.982)	(1.923)
Sample		Symm. & Prior	Symm. & Disc.	Symm. & Warn.
Incl. advisors		No extreme	No extreme	No extreme
Round × History FE		Yes	Yes	Yes
Observations		1570	1650	1640

Notes: (i) The dependent variable is the advisor's θ^A_{post} contained in the narrative; (ii) Each column contains data from a pair of treatments—the benchmark treatment, Symmetric-Rep, and another treatment: Sym-Prior, Sym-Disclosure, or Sym-Warning; (iii) Only observations present in the winsorized sample are used in the regressions; (iv) Standard errors are clustered at the matching group level and are reported in parenthesis; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

A.5.1 Additional Figure

We also double-checked the average persuasion results using the same approach that we used in the analysis underlying Result 1. The figure below replicates Figure 2 using data from the Belief Updating treatments and Symmetric-Rep. These results broadly cohere with previous findings. First, the visual evidence confirms that persuasion in Symmetric-Rep is comparable to the patterns found in the initial Symmetric treatment (as displayed in Figure 2). Second, persuasion becomes slightly more muted in when moving from Symmetric-Rep to Symmetric-Prior and substantially more muted when moving to Symmetric-Disclosure and Symmetric-Warning.

Figure A.4: Effect of the advisor type on investor assessments in the SYMMETRIC intervention treatments



Notes: This figure was created in the same way as Figure 2, using the data from the winsorized sample of the SYMMETRIC intervention treatments.

A.6 Limits of Narrative Persuasion (Asymmetric scenario)

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. In section 5.3 on the main text, we discuss a set of treatments that examine these questions in the Symmetric scenario. To also explore these issues in the Asymmetric scenario, we conducted three additional treatments. In each of these treatments, we intervened on a specific feature of the choice environment to evaluate whether this reduced persuasion and helped to protect investors.

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

Asym-Prior— This treatment builds on Asymmetric, additionally eliciting the investor's prior belief about the data-generating process before providing the advisor's narrative. This intervention is the asymmetric analogue of the Symmetric-Prior intervention.

Asym-Disclosure— This treatment builds on Asymmetric but discloses the advisor's incentives to the investor in every round of the experiment. This intervention is equivalent to the Symmetric-Disclosure intervention. However, since it builds on Asymmetric, which does not elicit a prior, we also do not elicit a prior in Asym-Disclosure.

Both our Asym-Disclosure and Asym-Prior 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.

Asym-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).[25] The advisor, therefore, knows the true underlying parameters of the data-generating process and is still able to try to persuade the investor by

[25]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.

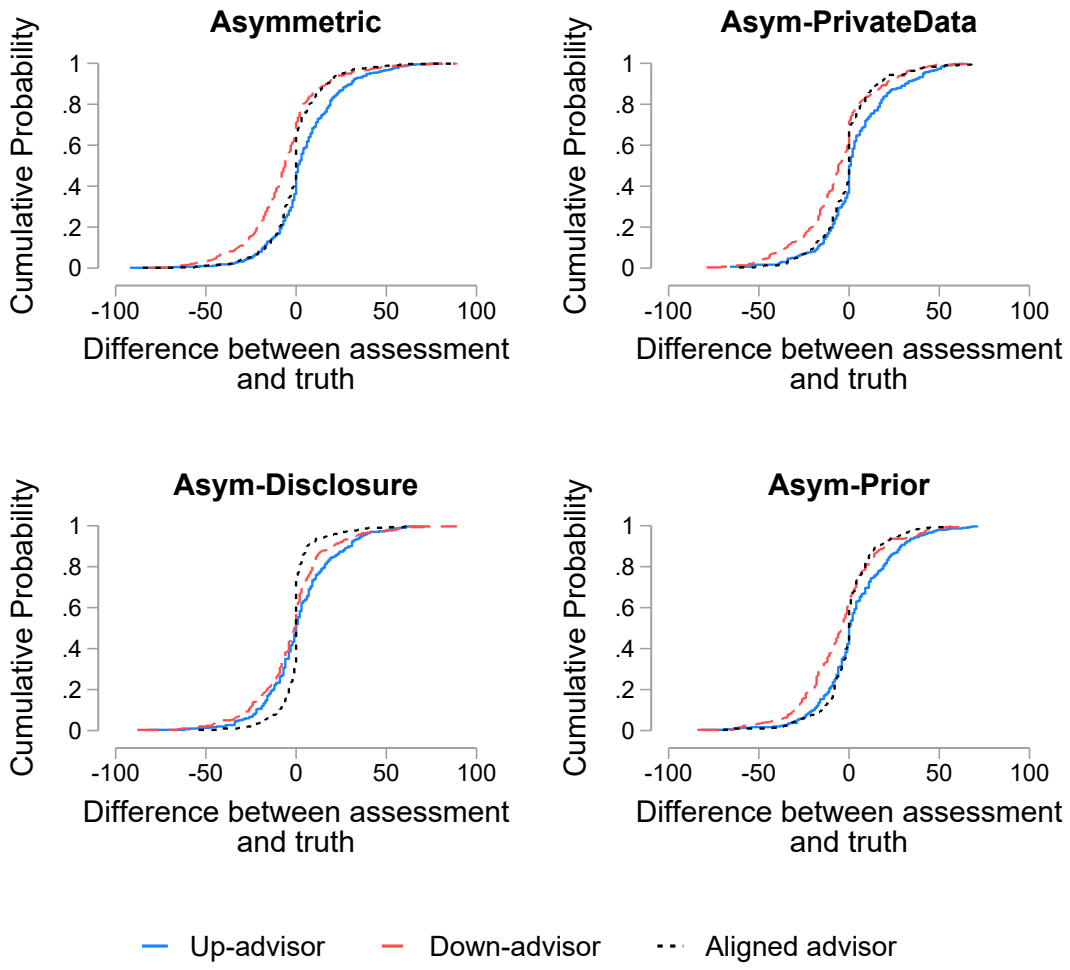
sending an inaccurate message. However, she is unable to precisely tailor the message to the data that the investor observes (she can only tailor it to her expectation of the data). This may make it more difficult for the advisor to send a message that is both deceptive and persuasive.

Procedures: We recruited 180 participants (90 advisors and 90 investors) per treatment via the Prolific platform in March 2022. Participants received a participation fee of £3.50 and could receive an additional bonus payment of £3.75.

Results

Limits of Persuasion— We begin by visualizing the distribution of investors' beliefs in each of the three treatments, alongside the benchmark Asymmetric treatment. Figure A.5 displays the cumulative distribution functions (CDFs) of the difference between investors' beliefs, θ^I_{post}, and the true value, θ^T_{post}. The dotted black, solid blue, and dashed 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 Asymmetric "intervention" treatments: Asym-PrivateData (top-right), Asym-Disclosure (bottom-left), and Asym-InvestorPrior (bottom-right). The figure shows that in the three treatments aside from Asym-Disclosure, investors appear to report higher beliefs when matched with an up-advisor in comparison to when matched with a down-advisor. In Asym-Disclosure, investors seem 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 A.5: 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.

To provide a more rigorous empirical test of the influence of each intervention on the effectiveness of persuasion, we replicate the analysis used in Table 4 in the main text, which examined interventions in the SYMMETRIC scenario. Table A.12 reports the results for the interventions in the ASYMMETRIC scenario. For transparency reasons we repeat the same winsorizing procedure we conducted in the SYMMETRIC scenario (see Appendix A.5) also here although there were no substantial imbalances in narratives across treatments so that the estimates are not sensitive to winsorizing. As in the SYMMETRIC scenario, we observe negative coefficient estimates for β_2 across all columns. However, in columns (3*), the coefficient is not statistically significant, suggesting that PRIVATEDATA may not have reduced persuasion. The other columns show larger, statistically significant reduction in persuasion in ASYM-PRIOR and ASYM-DISCLOSURE treatment. The point estimates are also of comparable size to the point estimates we obtained in the SYMMETRIC scenario, suggesting that the interventions reduce persuasions in similar ways in both scenarios.

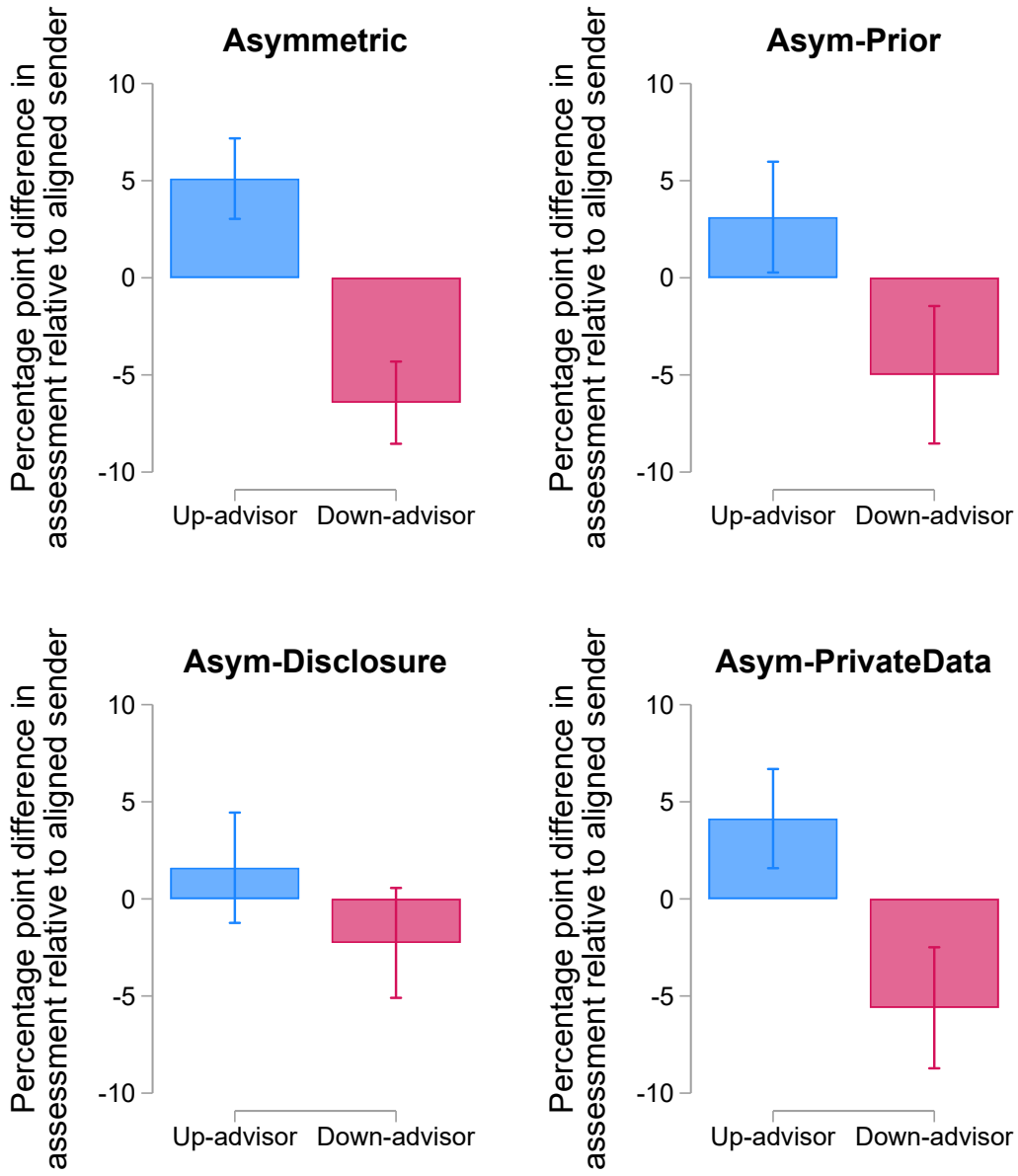
Table A.12: The influence of interventions on persuasion (in the ASYMMETRIC scenario)

		PRIOR		DISCLOSURE		PRIVATEDATA	
		$\theta^{I,1}_{post}$ (1a)	$\theta^{I,1}_{post}$ (1b)	$\theta^{I,1}_{post}$ (2a)	$\theta^{I,1}_{post}$ (2b)	$\theta^{I,1}_{post}$ (3a)	$\theta^{I,1}_{post}$ (3b)
$\mathbb{C}$(Adv. type)	(β_0)	5.842***	5.768***	5.842***	5.768***	5.842***	5.768***
		(0.546)	(0.598)	(0.546)	(0.598)	(0.546)	(0.598)
T	(β_1)	-0.0365	0.0917	0.904	0.631	-0.565	-0.459
		(0.727)	(0.846)	(0.795)	(0.863)	(0.831)	(0.849)
$\mathbb{C}$(Adv. type) × **T**	(β_2)	-1.715*	-1.709*	-4.103***	-3.833***	-1.016	-0.898
		(1.010)	(0.989)	(0.964)	(0.981)	(1.014)	(0.998)
Incl. treatments		Asymm. & Prior	Asymm. & Prior	Asymm. & Disc.	Asymm. & Disc.	Asymm. & PrivDat.	Asymm. & PrivDat.
Incl. advisors		No extreme	All	No extreme	All	No extreme	All
Round × History FE		Yes	Yes	Yes	Yes	Yes	Yes
Observations		2390	2700	2480	2700	2410	2700

Notes: (i) The dependent variable is the investor's posterior belief, $\theta^{I,1}_{post}$, after receiving a narrative from an advisor; (ii) Each column contains data from a pair of treatments—the benchmark treatment, ASYMMETRIC, and another treatment: ASYM-PRIOR, ASYM-DISCLOSURE, or ASYM-WARNING; (iii) The (*a) columns include observations from a winsorized sample, using the same procedures as described in Appendix A.5 applied to the ASYMMETRIC setting, and the (*b) columns include all observations; (iv) Standard errors are clustered at the matching group level and are reported in parenthesis; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

We also replicate Figure 2 for the different intervention treatments in Figure A.6.

Figure A.6: Effect of the advisor type on investor assessments in the ASYMMETRIC intervention treatments



Notes: This figure was created in the same way as Figure 2, using the data from the ASYMMETRIC intervention treatments.

Protection of Investors— 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.13 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.

As we see in Figure A.5, investors appear to display more skepticism of advisors' messages in DISCLOSURE, relative to the other treatments. 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).

Table A.13: 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

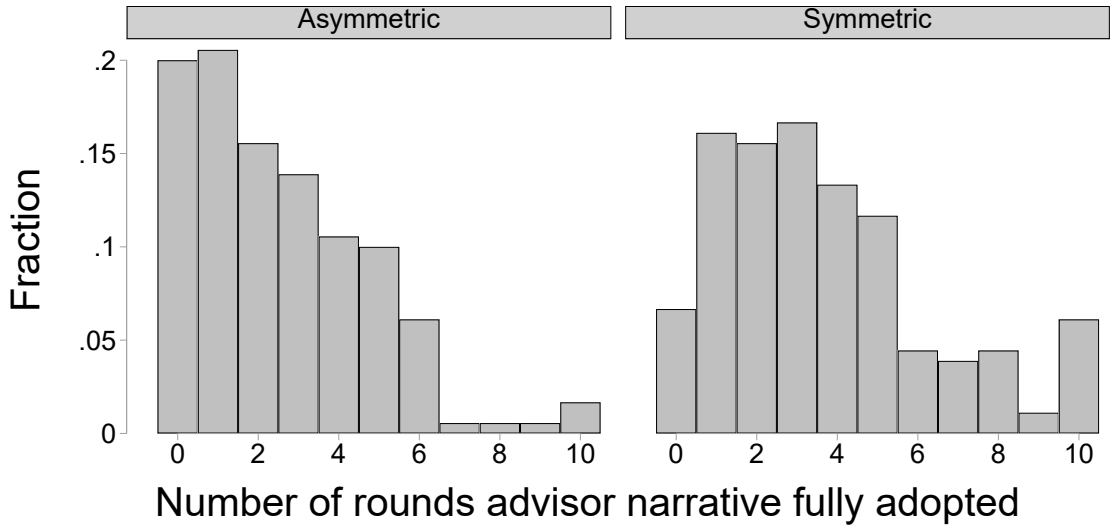
(i) The dependent variable is the distance between the true θ^T_{post} parameter and the corresponding belief held by the investor θ^I_{post}, (ii) Each column uses data from the ASYMMETRIC treatment as well as the relevant treatment mentioned in the column header, (iii) The value of the constant is 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) ** $p < 0.05$, *** $p < 0.01$.

A.7 Attention

This section examines whether our results could be driven by inattentive participants. To investigate this potential concern, we conduct several exercises.

First, we consider the possibility that investors are unthinkingly following advisors' messages. We capture this form of naive adoption by counting the number of times each investor *fully adopts* the advisor's message—that is, sets their final assessment equal to the advisor's suggested value (i.e., $\theta^{I,1}_{\text{post}} = \theta^{A}_{\text{post}}$). Since there are ten rounds, each investor receives a *full adoption score* between 0 and 10. Figure A.7 reports the distribution of full adoption scores for the Asymmetric and Symmetric treatments. It is clear from the figure that only a minority of participants naively follow the advisor in all ten rounds (approximately 6% in Symmetric and less than 2% in Asymmetric). To check whether our main results are robust to the exclusion of such participants, we replicate Figure 2, which documents susceptibility to persuasion in Asymmetric and Symmetric. Specifically, we conduct one analysis that removes individuals who fully adopted ten times, and another that removes all individuals who fully adopted eight or more times. The results, reported in Table A.14 and Figure A.8, remain highly consistent with our main findings.

Figure A.7: Histogram of number of rounds each investor fully adopts, by treatment



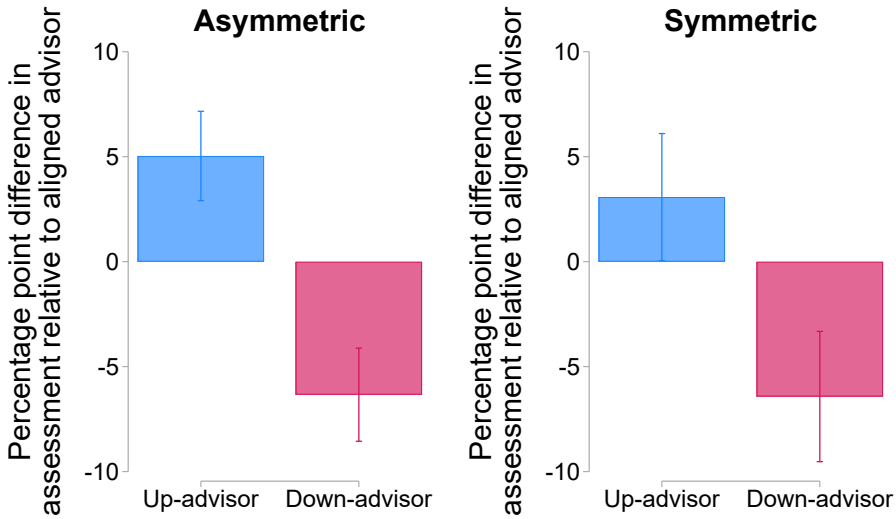
Notes: (i) To construct this histogram, for each individual, we calculate the number of rounds (from 10) that they fully adopt the message received from the advisor. We then plot the distribution over individual investors, (ii) This measure is conservative as it may overestimate full adoption by including cases in which the investor's own prior assessment coincides with the advisor's message.

Table A.14: Investor assessment by advisor type and treatment, among full adoption subsamples

	(1) $\theta^{I,1}_{post}$	(2) $\theta^{I,1}_{post}$	(3) $\theta^{I,1}_{post}$	(4) $\theta^{I,1}_{post}$	(5) $\theta^{I,1}_{post}$	(6) $\theta^{I,1}_{post}$
Up-advisor	5.105***	5.049***	5.035***	3.993***	3.266**	3.077*
	(1.057)	(1.072)	(1.086)	(1.463)	(1.494)	(1.544)
Down-advisor	-6.430***	-6.376***	-6.335***	-5.612***	-5.923***	-6.423***
	(1.080)	(1.112)	(1.131)	(1.425)	(1.559)	(1.583)
Treatment	Asymmetric	Asymmetric	Asymmetric	Symmetric	Symmetric	Symmetric
Sample	Whole	Fully adopts < 10	Fully adopts < 8	Whole	Fully adopts < 10	Fully adopts < 8
Round FE	Yes	Yes	Yes	No	No	No
Round × History FE	No	No	No	Yes	Yes	Yes
Observations	1800	1770	1750	1800	1684	1577

Notes: (i) The dependent variable is the investor's assessment (ii) The sample used in columns (1)-(3) contains data from investors in ASYMMETRIC, while Column (4)-(6) 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 A.8: Effect of the advisor type on investor assessments, for advisors fully adopting less than eight times



Notes: The figure plots the coefficient estimates and 95% confidence intervals of columns (3) and (6) of Table A.14.

Second, we investigate whether our results could be driven by investors who rushed through the experiment. Such participants may have paid less attention to the task. For completeness, we also consider the opposite possibility—that very slow decision times may reflect lower decision quality, for example due to distraction. Looking at decision times, we find that the average participant spent 35 seconds on each company assessment. This already suggests that most participants took time to think carefully about their decisions. To directly test the robustness of our main results, we replicate Figure 2 and Table 1 using two alternative specifications that exclude either the 25% fastest or the 25% slowest participants, based on average decision time across the ten rounds.[26] The results, reported in Table A.15 and Figure A.9, remain highly consistent with our

[26]In ASYMMETRIC, the software did not record the decision times of eight investors in at least one round of the experiment. We, consequently, drop these eight investors from the analysis.

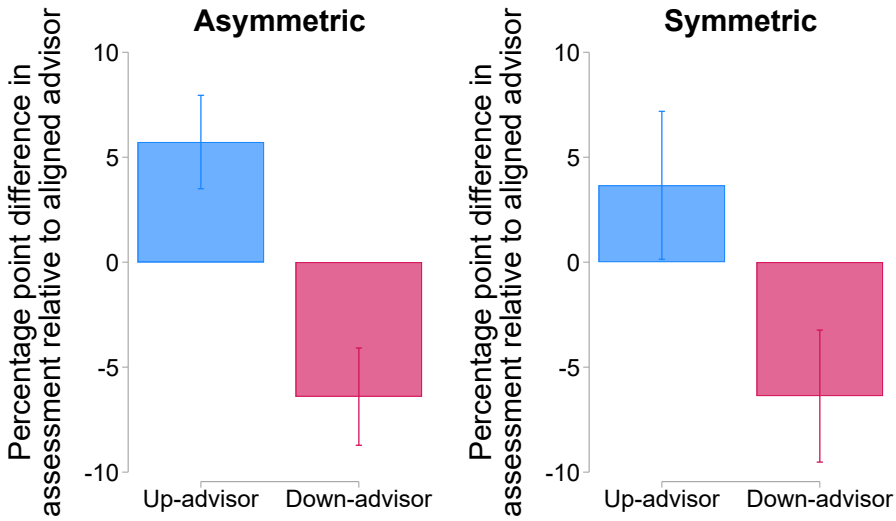
main findings.[27]

Table A.15: Investor assessment by advisor type, treatment, and decision time

	(1) $\theta^{I,1}_{post}$	(2) $\theta^{I,1}_{post}$	(3) $\theta^{I,1}_{post}$	(4) $\theta^{I,1}_{post}$	(5) $\theta^{I,1}_{post}$	(6) $\theta^{I,1}_{post}$
Up-advisor	5.221***	5.723***	5.225***	3.993***	3.662**	3.714**
	(1.071)	(1.135)	(1.173)	(1.463)	(1.798)	(1.786)
Down-advisor	-6.362***	-6.406***	-6.171***	-5.612***	-6.378***	-4.824**
	(1.069)	(1.182)	(1.109)	(1.425)	(1.605)	(1.867)
Treatment	Asymmetric	Asymmetric	Asymmetric	Symmetric	Symmetric	Symmetric
Sample	Whole	w/o 25% fastest	w/o 25% slowest	Whole	w/o 25% fastest	w/o 25% slowest
Round FE	Yes	Yes	Yes	No	No	No
Round × History FE	No	No	No	Yes	Yes	Yes
Observations	1720	1290	1290	1800	1320	1300

Notes: (i) The dependent variable is the investor's assessment (ii) The sample used in columns (1)-(3) contains data from investors in ASYMMETRIC, while Column (4)-(6) 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 A.9: Effect of the advisor type on investor assessments, excluding 25% fastest investors



Notes: The figure plots the coefficient estimates and 95% confidence intervals of columns (2) and (5) of Table A.15.

A.8 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

[27]We also consider two alternative measures of whether participants rushed through the task or progressed unusually slowly: (i) the time spent on each individual company assessment, rather than the average across all ten rounds, and (ii) the time taken to read the instructions. Excluding the 25% fastest or 25% slowest participants based on these metrics yields results that remain highly consistent with our main findings. In the interest of space, we do not report these results here, but they are available upon request.

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}.[28] 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.

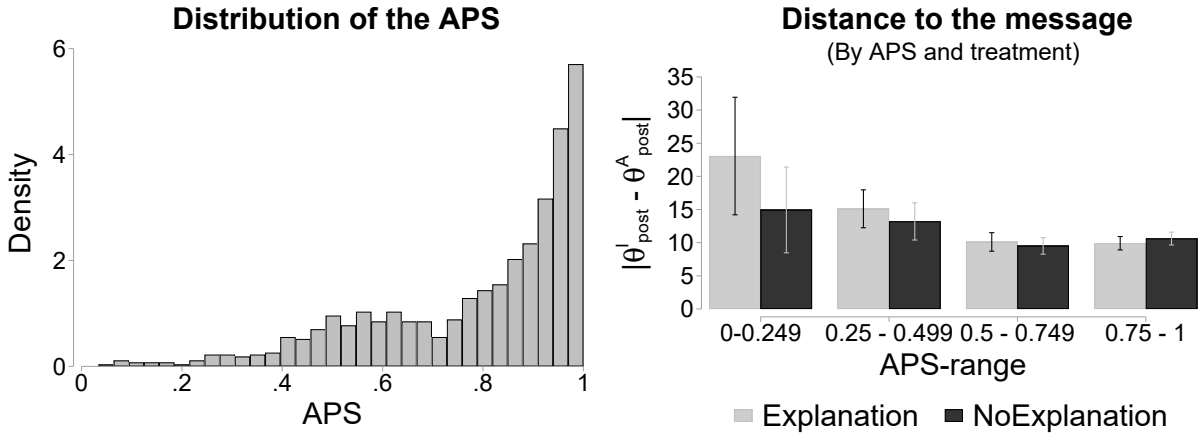
Main Results: Table 6 in the main text reports our main results from the "explanation" treatments, and Section 6 contains a discussion of these results. Essentially, we find that the quality of the explanation matters—claims supported by a good explanation are more persuasive than claims supported by a bad explanation.

Additional Evidence: 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 A.10, 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 6. They further suggest that, in the data, investors are particularly skeptical

[28] 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 Asym-Prior 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.

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 A.10: Distribution of the APS and relation between the APS and the investor's distance to the message (by treatment)



Notes: The left panel presents the distribution of the Auxiliary Parameter Score (APS) and the right panel 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 that cluster standard errors at the investor level.

Relationship to Cheap Talk: The evidence above is inconsistent with a broad class of cheap talk equilibria that 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. See Appendix D for a formal treatment (in particular, Proposition 7 and Subsection D.6). 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. To summarize, the persuasion we observe here, and also in other parts of 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.

A.9 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.[29]

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. 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

[29]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.

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.[30] 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.[31] 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.16 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 EPI on adoption is continuous.[32] 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.[33] 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 parame-

[30] We present parameter estimates using the whole message space below in the "Additional Evidence" subsection. 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.

[31] We used Monte Carlo experiments to confirm that this procedure reliably identifies the true underlying parameter values. See the "Additional Evidence" subsection below for details.

[32] 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.16) is enough to generate a discontinuous jump.

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

ter 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.16: 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.11 consists of three panels that each display:

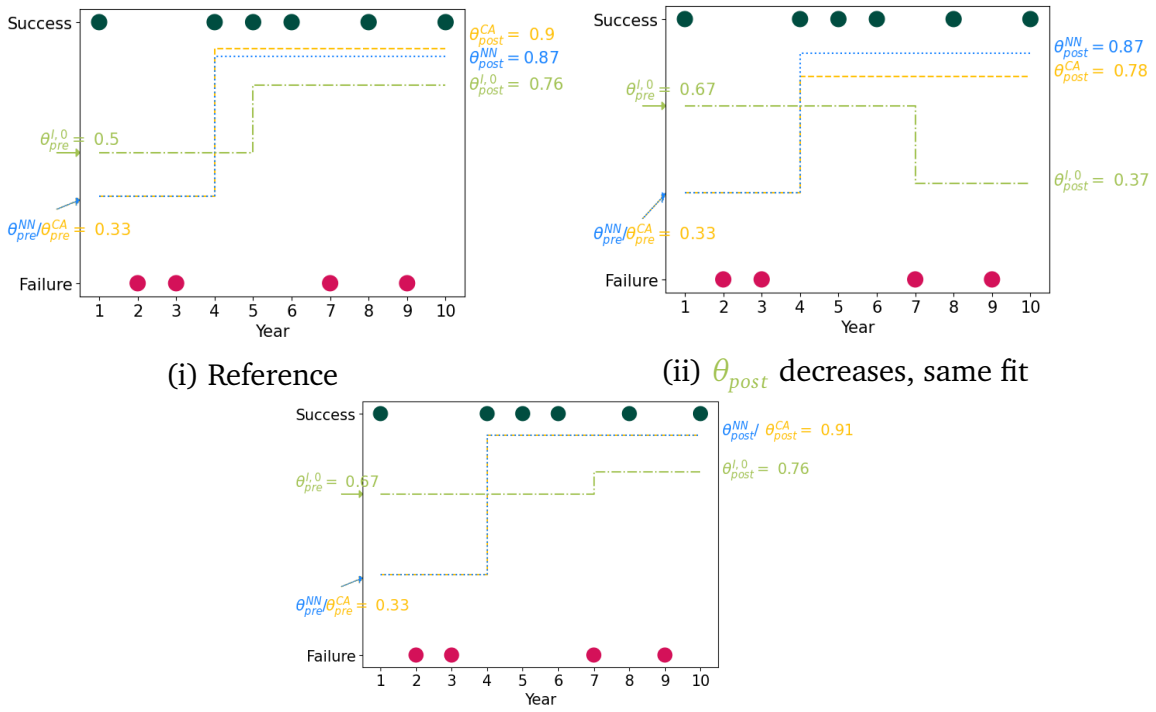
(a) a realization of historical company data (which is constant across all panels),

(b) a competing narrative (in green),

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

When moving from Panel *(i)* to Panel *(ii)* in the figure, the empirical fit of the default model is held constant. Consequently, $\boldsymbol{m}^{NN}$ 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 $\boldsymbol{m}^{NN}$.

Figure A.11: 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.

Additional evidence: In this section, we provide two additional pieces of evidence to supplement the analysis of Section A.9. First, Table A.17 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.16, 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 previously.

Table A.17: 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 A.18 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 A.18: 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. * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

A.10 Shaping the Interpretation of Data: Direct Evidence

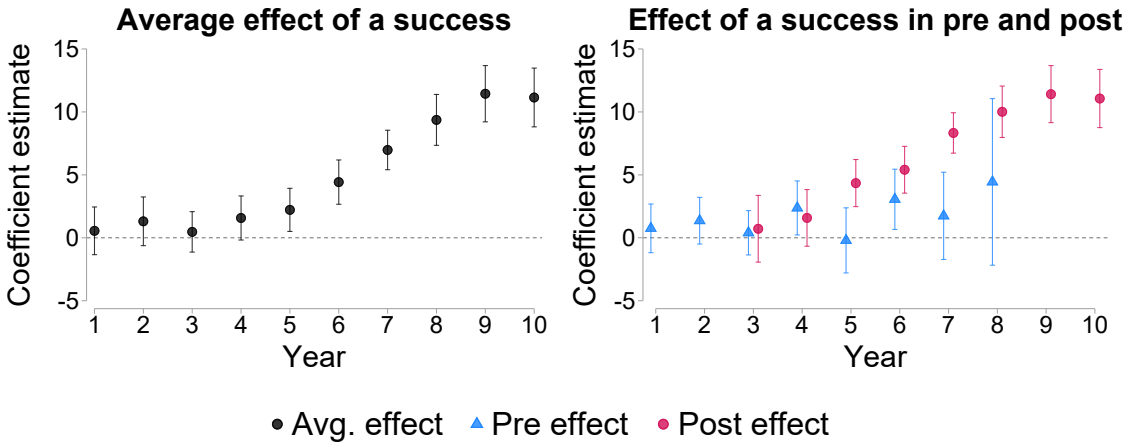
While much of the analysis above has focused on evaluating the impact of the advisor's narrative on the investor's assessment, it is also informative to examine how the investor uses the historical data directly to form his assessment. In particular, we can ask whether more recent successful years in 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 A.12 plots the β-coefficient estimates, using data from ASYMMETRIC. 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 A.12: Effect of company success on assessments, by year



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 A.12 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 . This result is consistent with the idea that the advisor influences which years in the data the investor deems relevant when making his assessment.

We statistically test for the presence of such shaping in Table A.19. This table reports regression estimates of the equation

$$\theta^{I,1}_{post} = \beta_1 s_1 + \beta_2 s_2 + \beta_9 s_9 + \beta_{10} s_{10} + \alpha_0 s_{flex} + \alpha_1 s_{flex,post} + \rho + \varepsilon.$$

In this equation, s_{flex} is the sum of successes between year 3 - 8. These are the years where the advisor can flexibly choose where to split them into *pre* and *post*. The specification also includes $s_{flex,post}$ which sums the successes between year 3 - 8 that the advisor attributed to *post*. This last term is our term of interest. A positive coefficient estimate would indicate that investors put more weight on success and failure of those years in the flexible period that the advisor attributed to *post*. The estimation results indicate that this is indeed the case in ASYMMETRIC and also in SYMMETRIC, SYMMETRIC-REP, and SYMMETRIC-PRIOR. While the α_1 coefficient generally becomes smaller when additionally controlling for θ^A_{post}, it still remains positive and significant. This provides some indication that the narrative shapes assessments beyond θ^A_{post}, through the asserted structural break.

Table A.19: Shaping how investors interpret data

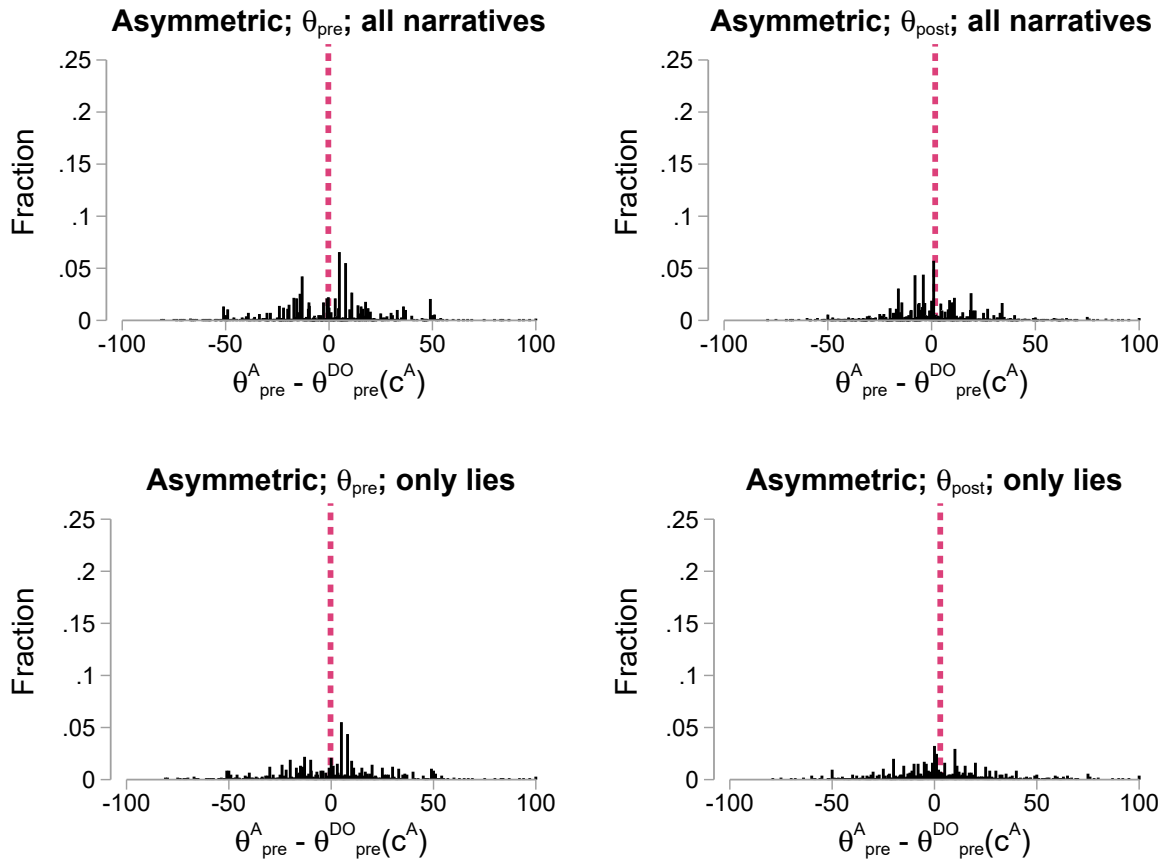
	ASYMMETRIC		SYMMETRIC		SYM-REP		SYM-PRIOR	
	$\theta^{I,1}_{post}$	$\theta^{I,1}_{post}$	$\theta^{I,1}_{post}$	$\theta^{I,1}_{post}$	$\theta^{I,1}_{post}$	$\theta^{I,1}_{post}$	$\theta^{I,1}_{post}$	$\theta^{I,1}_{post}$
s_1	0.529	0.642	-0.937	0.363	-2.594	-1.944	1.319	1.156
	(0.980)	(0.846)	(1.029)	(0.848)	(1.876)	(1.409)	(1.330)	(1.131)
s_2	1.339	0.588	-0.258	-0.168	1.012	0.855	-2.875**	-1.139
	(0.992)	(0.935)	(1.094)	(0.893)	(1.801)	(1.385)	(1.277)	(0.933)
s_9	11.31***	9.871***	16.49***	10.11***	17.44***	12.39***	14.85***	8.948***
	(1.164)	(0.982)	(1.416)	(1.126)	(2.387)	(1.956)	(1.819)	(1.683)
s_{10}	11.57***	10.18***	21.37***	9.973***	29.33***	14.98***	23.33***	14.33***
	(1.203)	(1.087)	(1.612)	(1.273)	(2.814)	(2.459)	(2.382)	(1.931)
s_{flex}	2.027***	2.435***	1.705***	1.470***	1.141	1.396	2.044***	1.949***
	(0.564)	(0.482)	(0.580)	(0.440)	(1.109)	(0.902)	(0.658)	(0.619)
$s_{flex,post}$	3.346***	1.970***	3.681***	1.304***	4.129***	1.270*	3.059***	1.191**
	(0.489)	(0.375)	(0.415)	(0.419)	(0.758)	(0.624)	(0.514)	(0.531)
θ^A_{post}		0.434***		0.489***		0.487***		0.399***
		(0.0343)		(0.0291)		(0.0369)		(0.0364)
Round FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1800	1800	1800	1800	900	900	900	900

Notes: (i) The dependent variable is the investor's assessment, $\theta^{I,1}_{post}$; (ii) The different column headers indicate the different treatments on which the regression was estimated on (iii) Standard errors clustered at the matching group level are reported in parenthesis; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

A.11 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.

Figure A.13: Histograms of the difference between θ^A and the conditionally data-optimal parameter (ASYMMETRIC).



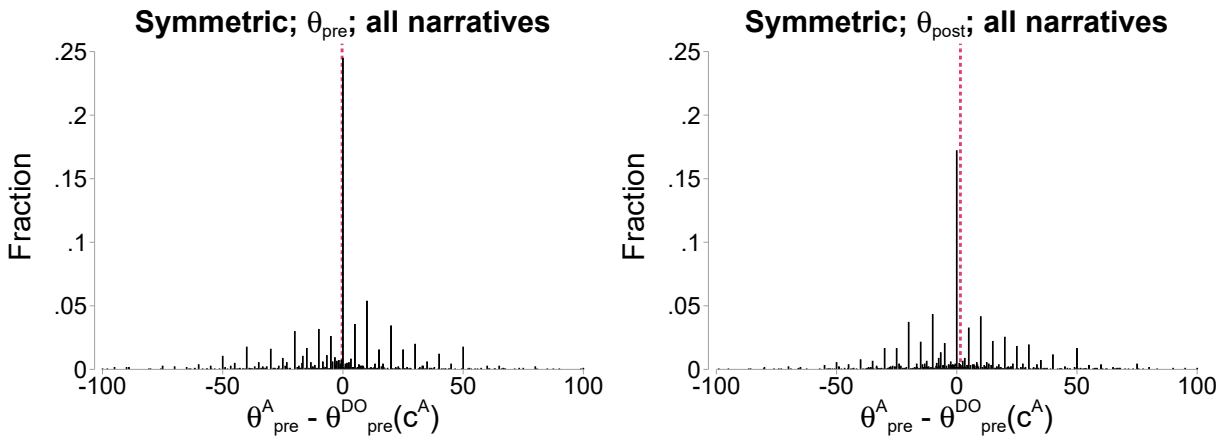
Notes: The figures use data from the ASYMMETRIC treatment. The top two panels use data of all advisors, while the bottom two panels only use data of advisors who lie on at least one dimension of the narrative. The red vertical lines plot the mean difference.

Figure A.13 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 A.13 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 A.14 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 A.14: Histograms of the difference between θ^A and the conditionally data-optimal parameter (SYMMETRIC).



Notes: The figures use data from all advisors in the SYMMETRIC treatment.

A.12 A Closer Look at Belief Updating, Fit and Incentive Disclosure

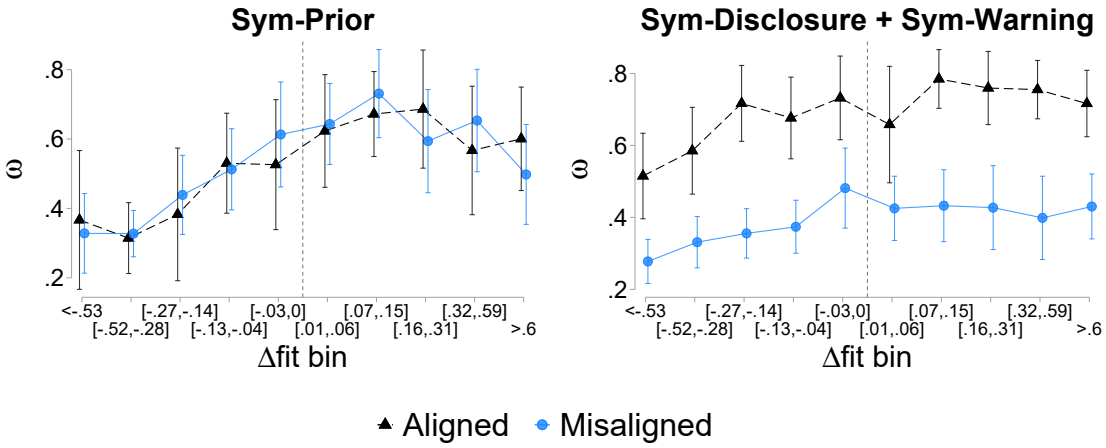
In this section, we present the results from two additional exercises that take a closer look at how narrative fit and incentive disclosure influence belief updating. First, we use a continuous measure of relative fit to replicate the result portrayed in Figure 5 of the main text—namely, that sensitivity to fit is pronounced in Symmetric-Prior but more muted in the "disclosed incentives" treatments. Second, we explore why investors update their beliefs in response to narratives from misaligned advisors even when their incentives are disclosed, providing evidence that the *apparent self-interest* of a narrative influences investor reactions when advisor incentives are disclosed.

How disclosing incentives changes updating: fit & incentive disclosure. In the main text, we show that investors in Symmetric-Disclosure and Symmetric-Warning react little to better-fitting narratives, while those in Symmetric-Prior react strongly. Here, we provide a more detailed comparison of belief updating with and without disclosed incentives. For concision, we pool the data from Symmetric-Disclosure and Symmetric-Warning into a single "disclosed incentives" condition, as both share this feature and yield comparable behavior.

We begin by replicating this result from Figure 5 using a continuous measure of relative fit, showing that sensitivity to fit is pronounced in Symmetric-Prior but muted in the "disclosed incentives" treatments. To do this, we categorize Δfit into ten bins. The bin sizes were chosen to ensure roughly equal numbers of observations per bin, with two adjacent bins sharing a boundary at Δfit$= 0$. This is the point where the narrative shifts from fitting slightly worse than the default to fitting slightly better. We then regress ω on bin indicators, doing this separately by advisor alignment and incentive disclosure condition.

Figure A.15 presents the results. In all four scenarios, we see that the average ω increases gradually with the relative fit measure, plateauing somewhat for Δfit > 0. We also observe no discontinuity at the Δfit $= 0$ threshold. Given that we know from earlier results that a substantial fraction of updates are bang-bang (i.e., either $\omega = 0$ or $\omega = 1$), the fact that the average ω changes gradually around the threshold suggests that perceptions of relative fit may be noisy.

Figure A.15: Average updating weight by treatment, alignment, and fit.



Notes: (i) The left panel includes data from investors in Sym-Prior and the right panel includes data from investors in Sym-Disclosure and Sym-Warning; (ii) Both panels only include observations from investors with convex updates; (iii) The estimates were derived using regressions of ω on bin indicators; (iv) We first rounded Δfit to two digits after the dot before binning; (v) Error bars are 95% confidence intervals clustered at the matching group level.

Why do investors update toward misaligned advisors—even when incentives are disclosed? Here, we examine why investors update toward misaligned advisors even when their incentives are disclosed. To investigate this, we conduct an exercise aimed at identifying cases where such updating occurs. The main idea is as follows. If investors' defaults are noisy—sometimes underestimating and sometimes overestimating θ^T_{post}—then there can be cases where the investor's default is actually closer to the advisor's persuasion target than the advisor's narrative. In such cases, adopting the narrative would move the investor *further away* from the advisor's target, potentially reducing the investor's skepticism toward messages in this type of scenario.

We investigate whether this is the case by tracing how ω changes as a function of the difference between the narrative and the investor's prior, $\Delta\theta_{post} = \theta^A_{post} - \theta^{I,0}_{post}$. We refer to this as *apparent narrative self-interest*—whether the advisor's message appears to serve their own persuasion goal. Since we pool up- and down-advisors in this analysis, we invert the difference for down-advisors to create a normalized measure, $\Delta\widetilde{\theta}_{post}$. Positive values of $\Delta\widetilde{\theta}_{post}$ indicate that the misaligned advisor's recommendation is closer to their persuasion target than the investor's prior; negative values indicate that the recommendation lies on the opposite side of the investor's prior from the persuasion goal the advisor is incentivized to pursue.

We then partition $\Delta\widetilde{\theta}_{post}$ into ten bins of roughly equal size. Since we are particularly interested in the categories adjacent to zero, where a misaligned advisor's narrative switches from going against their private interests to advancing their private interests (relative to the investor's prior), we ensure that there are two bins adjacent to zero—one bin with an upper limit of $\Delta\widetilde{\theta}_{post} = -1$ and one bin with a lower limit of $\Delta\widetilde{\theta}_{post} = 1$.[34] Using this partition, we regress ω on bin indicators, separately by disclosed incentives condition and advisor alignment.

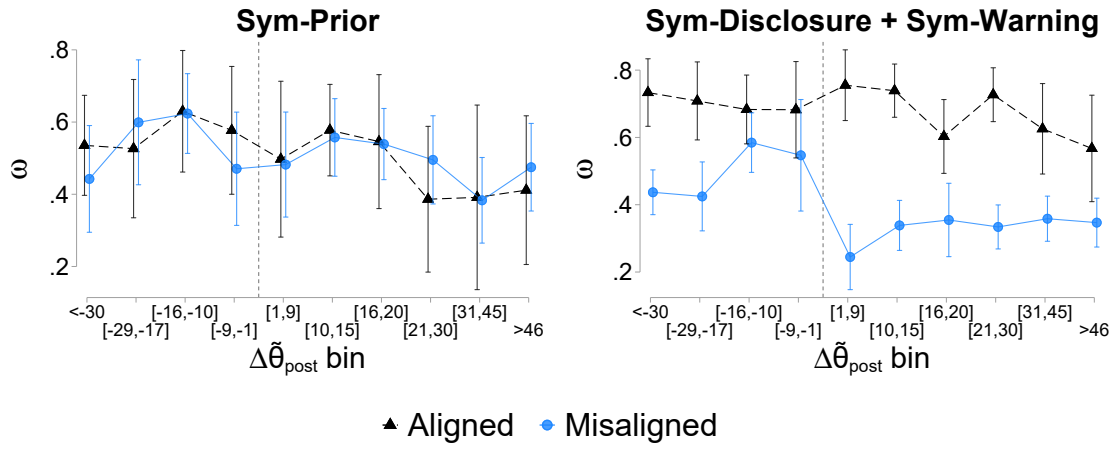
Figure A.16 reports the results. The figure shows a striking difference in belief updating between Symmetric-Prior and the other two treatment conditions. In Symmetric-Prior, investors update very similarly in response to narratives from aligned and misaligned advisors. Their updating also does not appear to react strongly to variation in the apparent self-interest of the narratives—i.e., they update similarly across different $\Delta\widetilde{\theta}_{post}$-categories. Specifically, most point estimates hover around $\omega = 0.5$ or slightly above, with the exception of those associated with very high $\Delta\widetilde{\theta}_{post}$ values (i.e., above 20), where there is a slight drop in ω.

In contrast, in the Symmetric-Disclosure and Symmetric-Warning treatments, the most notable difference emerges in how investors respond to narratives from misaligned advisors. There is a sharp discontinuity around zero: investors are willing to move toward a misaligned advisor's narrative when doing so runs counter to the advisor's interests, but they are much less willing to update when updating aligns with the advisor's incentives. This asymmetry is absent in the Symmetric-Prior treatment, where the advisor's type is unknown.

These results shed light on how investors become more skeptical of narratives from advisors with a known conflict of interest when the message appears to serve the advisor's own interests. Yet when a conflicted advisor offers a narrative that runs counter to their own incentives, investors are willing to trust it—at least to some extent. Hence, part of the observed updating towards an advisor with disclosed misalignment is explained by investors updating more when adopting the misaligned advisor's narrative goes against the advisor's interests. The average ω for misaligned advisors with disclosed incentives is 0.5 when they send a narrative that goes against their private interest (i.e., when $\Delta\widetilde{\theta}_{post} < 0$), but drops to 0.33 when the narrative aligns with their private interest (i.e., when $\Delta\widetilde{\theta}_{post} > 0$).

[34] Since ω is undefined when $\Delta\widetilde{\theta}_{post} = 0$, the analysis does not include these observations.

Figure A.16: Updating weights by treatment, alignment, and narrative self-interest ($\Delta\widetilde{\theta}_{post}$).



Notes: (i) The left panel includes data from investors in Sym-Prior and the right panel includes data from investors in Sym-Disclosure and Sym-Warning; (ii) Both panels only include observations from investors with convex updates; (iii) The estimates were derived using regressions of ω on bin indicators; (iv) Error bars are 95% confidence intervals clustered at the matching group level.

B Additional Details on the Competition Treatment

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.

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.

C Additional Discussion of S&S's Narrative Approach

C.1 Notation

Throughout the discussion, we will use several notational shortcuts. Define by

$$k_{pre}(c) \equiv \sum_{t=1}^{c} s_t, \quad f_{pre}(c) \equiv c - k_{pre}(c), \quad k_{post}(c) \equiv \sum_{t=c}^{10} s_t, \text{ and } f_{post}(c) \equiv 10 - c - k_{post}(c)$$

the numbers of successes and failures in the *pre* and *post* period for a given c.

The likelihood function is equal to

$$\pi_{\boldsymbol{m}} = \theta_{pre}^{k_{pre}(c)}(1-\theta_{pre})^{f_{pre}(c)}\theta_{post}^{k_{post}(c)}(1-\theta_{post})^{f_{post}(c)}.$$

C.2 Proof of Proposition 1

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

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

denotes the cdf of the default narrative likelihood fit and $\tilde{F}(l) = \int_{-\infty}^{l} \tilde{f}(s)\,\mathrm{d}s$ is the cdf. This distribution has full support on $[0, \pi_{\boldsymbol{m}^{DO}}]$.

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

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

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

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

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

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

Since U^{φ} is strictly concave with an optimum at the persuasion target τ_{φ}, Part *(i)* follows.

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

$$\pi^C_{\theta_{post}} \equiv \pi_{(\hat{c}(\theta_{post}), \hat{\theta}_{pre}(\theta_{post}), \theta_{post})}.$$

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:

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

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 $\theta_{pre} \in [0,1]$ such that $\pi_{c,\theta_{pre},\theta_{post}} = \bar{\pi}$, where $\bar{\pi} \in [0, \pi^C_{\theta_{post}}]$. If $\bar{\pi} = \pi^C_{\theta_{post}}$, the claim directly follows as the message $(\hat{c}(\theta_{post}), \hat{\theta}_{pre}(\theta_{post}), \theta_{post})$ induces likelihood value $\bar{\pi}$. Now consider $\bar{\pi} < \pi^C_{\theta_{post}}$. We know that

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

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

$$\pi_{(\hat{c}(\theta_{post}), t, \theta_{post})} = t^{k_{pre}(\hat{c}(\theta_{post}))}(1-t)^{f_{pre}(\hat{c}(\theta_{post}))}\theta_{post}^{k_{post}(\hat{c}(\theta_{post}))}(1-\theta_{post})^{f_{post}(\hat{c}(\theta_{post}))}.$$

Observe that π is continuous in t. We consider two cases.

(i) If $k_{pre} > 0$, and $t = 0$, $\pi_{(\hat{c}(\theta_{post}), t, \theta_{post})} = 0$. If $t = \hat{\theta}_{pre}(\theta_{post})$, $\pi_{(\hat{c}(\theta_{post}), t, \theta_{post})} = \pi^C_{\theta_{post}}$. By the intermediate value theorem, there is at least one value $t \in (0, \hat{\theta}_{pre}(\theta_{post}))$ such that $\pi_{(\hat{c}(\theta_{post}), t, \theta_{post})} = \bar{\pi}$.

(ii) If $f_{pre} > 0$, and $t = 1$, $\pi_{(\hat{c}(\theta_{post}), t, \theta_{post})} = 0$. If $t = \hat{\theta}_{pre}(\theta_{post})$, $\pi_{(\hat{c}(\theta_{post}), t, \theta_{post})} = \pi^C_{\theta_{post}}$. By the intermediate value theorem, there is at least one value $t \in (\hat{\theta}_{pre}(\theta_{post}), 1)$ such that $\pi_{(\hat{c}(\theta_{post}), t, \theta_{post})} = \bar{\pi}$.

One case always applies as at least one of k_{pre} or f_{pre} is strictly positive. We conclude that we can always fix θ_{post} and find auxiliary parameter values that induce any likelihood fit on $[0, \pi^C_{\theta_{post}}]$.

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 τ_ϕ. For the proof, we introduce a correspondence, which, for a given likelihood value $\bar{\pi}$, returns all messages whose fit is equal to that value:

$$\dot{\mathcal{M}}(\bar{\pi}) = \{ \boldsymbol{m} \in \mathcal{M} | \pi_{\boldsymbol{m}} = \bar{\pi} \}.$$

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

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

Note that any alternative model in $\dot{\mathcal{M}}(\bar{\pi})$ only changes the value of $U^\varphi(\cdot)$ in the first term of the utility function, while the values of all other utility components remain fixed. Therefore, choosing the message that maximizes utility for a given level of fit $\bar{\pi}$ 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 $(\tau_\phi - \theta^A_{post})^2$.

Claim 3: Suppose that $\theta^*_{post} \neq \tau_\phi$. Then, $\boldsymbol{m}^* \in \mathcal{C}$. Take a $\theta^*_{post} \neq \tau_\phi$ and suppose by contradiction that $\boldsymbol{m}^* \notin \mathcal{C}$. 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 $\pi(\boldsymbol{m}') = \pi(\boldsymbol{m}^*)$ 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 $\boldsymbol{m}' = (c', \theta'_{pre}, \tau_\phi)$ and $\boldsymbol{m}'' = (c'', \theta''_{pre}, \tau_\phi)$ with $\pi_{\boldsymbol{m}'} > \pi_{\boldsymbol{m}''}$. The advisor prefers sending $\boldsymbol{m}'$ over sending $\boldsymbol{m}''$. Denote by $\Delta\tilde{F}$ the difference $\tilde{F}(\pi_{\boldsymbol{m}'}) - \tilde{F}(\pi_{\boldsymbol{m}''})$. For notational brevity we will also use $\tilde{F}'' \equiv \tilde{F}(\pi_{\boldsymbol{m}''})$, $\pi' \equiv \pi_{\boldsymbol{m}'}$, $\pi'' \equiv \pi_{\boldsymbol{m}''}$, and $\pi^{I,0} \equiv \pi_{\boldsymbol{m}^{I,0}}$. We can then denote the expected utility of the sender from sending $\boldsymbol{m}'$ as

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

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 there is always a positive likelihood that an investor who follows message $\boldsymbol{m}'$ but not message $\boldsymbol{m}''$ has a default narrative with $\theta^{I,0}_{post} \neq \tau_\phi$, which implies that $U^\varphi(\tau_\phi) > \mathbb{E}[U^\varphi(\theta^{I,0}_{post})|\pi^{I,0} \in (\pi'', \pi')]$. 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.

Claim 4 implies that, if $\theta^A_{post} = \tau_\varphi$, then $\boldsymbol{m}^A \in \mathcal{C}$. By combining this insight with Claim 3, Part (ii) 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 narrative likelihood fits (that can be derived from F, see the proof of Proposition 1) by $\tilde{F}(\ell)$. We will refer to the set of narratives identified in Proposition 1 that the advisor may send as the "likelihood frontier".

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 (Fit-movement tradeoff in narrative construction). 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 τ_φ and that, *(ii)*, the advisor sends the data-optimal auxiliary parameters conditional on θ^A_{post}. 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 2. *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.,

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

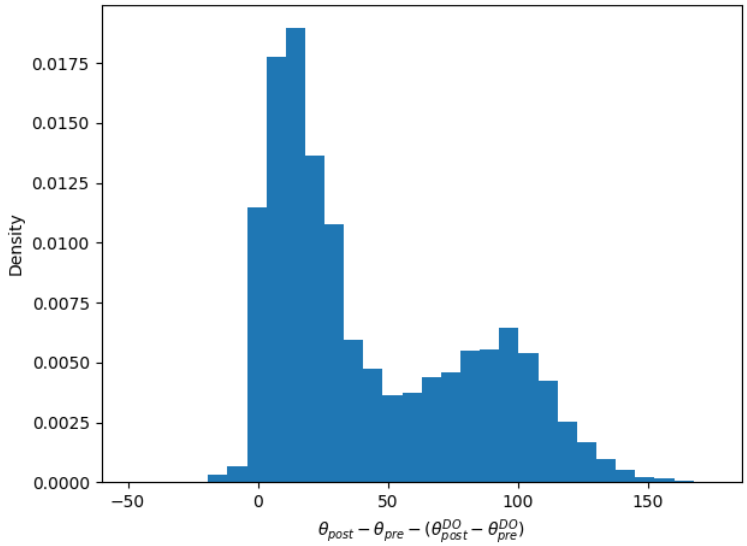
Proof. See Appendix E. □

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,[35] 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 DGP 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 predicted 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.

[35] 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 (Responding to a competing narrative). 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 $\pi^{min} \geq 0$ to $\pi_{\boldsymbol{m}^{DO}}$. Suppose that the fit of the investor's default narrative, $\pi_{\boldsymbol{m}^{I,0}}$, is smaller than π^{min}. Then, the advisor can send a message inducing the persuasion target τ_φ. Instead, if $\pi_{\boldsymbol{m}^{I,0}} \in (\pi^{min}, \pi_{\boldsymbol{m}^{DO}})$, the advisor will always send a message with fit $\pi_{\boldsymbol{m}^{I,0}}$: First, the advisor is always weakly better of by getting the investor to adopt her message; considering that she sends a message with a weakly higher fit than the default narrative is without loss. 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 $\pi_{\boldsymbol{m}^{I,0}}$. 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 $\pi_{\boldsymbol{m}^{I,0'}} < \pi_{\boldsymbol{m}^{I,0''}}$. 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 $\boldsymbol{m}^{I,0''}$.

Intervention treatment predictions. We briefly discuss how to map the intervention treatments into the formal framework.

Disclosing incentives— As discussed by S&S, the simple fit-based narrative adoption rule can easily be adapted to allow for the investor to be *skeptical*. 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:

$$\pi_{\boldsymbol{m}^A} \geq \pi_{\boldsymbol{m}^{I,0}} + s,$$

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. This is the mechanism we had in mind when designing Symmetric-Disclosure and Symmetric-Warning. Both treatments, discussed in detail in Section 4, disclose the advisor's incentives to the investor in each round of the experiment on their decision screen. We consequently expect that persuasion decreases in these treatments relative to the baseline non-disclosure case.

Eliciting a prior— In the narrative persuasion framework, the advisor can only convince the investor to adopt a narrative when it fits better than the investor's own prior interpretation of the data (i.e., his default narrative). To encourage investors to improve the fit of their default narrative, we implement the Symmetric-Prior treatment (discussed in detail in Section 4) 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, as in Symmetric, in Symmetric-Prior 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}. We propose the following way of thinking about the effect of asking the investor for a prior assessment of the data before receiving a narrative. 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 $\tilde{f}(\pi)$. Eliciting a prior first encourages a more carefully formed default narrative, which changes the default narrative density to g and the corresponding distribution of likelihood values to $\tilde{g}$. We propose that, because of "careful thinking", the density g is more concentrated around narratives close to the data-optimal narrative than f, resulting in a fit density $\tilde{g}$ that first-order stochastically dominates $\tilde{f}$. Consequently, the investor becomes less likely to adopt the advisor's narrative and persuasion decreases.

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. Parts D.1 to D.4 discuss the cheap talk predictions in the ASYMMETRIC scenario of our setup, where the advisor knows the true data-generating process. Part D.5 turns to analyzing the SYMMETRIC scenario, where the advisor sends a message without knowing the true data-generating process.

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 $\boldsymbol{m}^T$. This case corresponds to the ASYMMETRIC scenario in our experiment. We comment on the case where the advisor does not observe $\boldsymbol{m}^T$ (corresponding to the SYMMETRIC scenario in our experiment) further below in Section D.5. The investor does not observe $\boldsymbol{m}^T$, but it is common knowledge that $\boldsymbol{m}^T$ is distributed according to $G(\boldsymbol{m}^T)$. 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;

$$U^{\varphi}(\theta^I_{post}) = \begin{cases} 1-(1-\theta^I_{post})^2 & \text{if } \varphi = \uparrow, \\ 1-(0-\theta^I_{post})^2 & \text{if } \varphi = \downarrow, \\ 1-(\theta^T_{post}-\theta^I_{post})^2 & \text{if } \varphi = \rightarrow. \end{cases} \tag{9}$$

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 τ_φ.[36] 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

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

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.

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 $\mu(\boldsymbol{\tau})$.[37] 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

$$g_{\theta^T_{post}}(\theta^T_{post}) = \sum_{c=2}^{8} \frac{\int_0^1 \pi_{(c^T, \theta^T_{pre}, \theta^T_{post})} \, \mathrm{d}\theta^T_{pre}}{\sum_{c=2}^{8} \int_0^1 \int_0^1 \pi_{(c^T, \theta^T_{pre}, \theta^T_{post})} \, \mathrm{d}\theta^T_{pre} \mathrm{d}\theta^T_{post}}. \tag{10}$$

This expression gives the pdf of the marginal distribution of θ^T_{post} given g. In the equation above, $\pi_{(c^T, \theta^T_{pre}, \theta^T_{post})} = \theta_{pre}^{Tk_{pre}(c^T)}(1-\theta^T_{pre})^{f_{pre}(c^T)} \theta_{post}^{Tk_{post}(c)}(1-\theta^T_{post})^{f_{post}(c^T)}$ 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 that $q(\theta|k+1, f+1) \equiv \theta^k(1-\theta)^f / B(k+1, f+1)$ is the density function of the beta distribution with shape parameters $k+1$ and $f+1$. Substituting the likelihood terms out of Equation (10), the marginal density of θ^T_{post} becomes

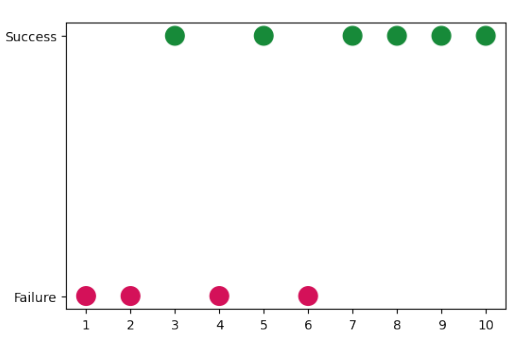
$$g_{\theta^T_{post}}(\theta^T_{post}) = \sum_{c=2}^{8} w_c \, q(\theta^T_{post} | k_{post}(c^T) + 1, f_{post}(c^T) + 1),$$
$$\text{where } w_c \equiv \frac{B(k_{pre}(c^T)+1, f_{pre}(c^T)+1) B(k_{post}(c^T)+1, f_{post}(c^T)+1)}{\sum_{c'=2}^{8} B(k_{pre}(c')+1, f_{pre}(c')+1) B(k_{post}(c')+1, f_{post}(c')+1)}.$$

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.

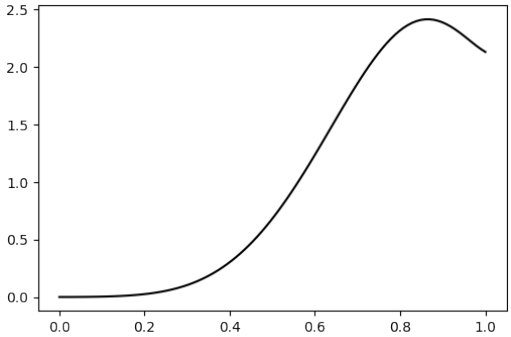
In the text below, whenever we refer to G, we refer to a distribution that is derived in the way described above and whose marginal pdf with respect to θ^T_{post} is given above. We will also

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

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



Historical data



Belief density over θ_{post}

generally assume that G has full support on $\mathcal{M}$, which is true in all but a few special cases.[38] 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.

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 $\sigma(\boldsymbol{m}^A|\tau)$ the probability that an advisor with type τ 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 $\boldsymbol{m} \in \mathcal{M}^{max}$.

[38] 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.

(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};

$$\mathbb{E}_\mu(\widetilde{\theta}_{post}|\boldsymbol{m}^A) = p(\boldsymbol{m}^A)\theta^A_{post} + (1-p(\boldsymbol{m}^A))\bar{\theta}, \tag{11}$$
$$\text{where } p(\boldsymbol{m}^A) = \frac{(2\lambda+1)g(\boldsymbol{m}^A)}{(2\lambda+1)g(\boldsymbol{m}^A) + (1-\lambda)\sigma(\boldsymbol{m}^A|\uparrow)}.$$

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 up-advisor's optimal strategy

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

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 3. *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 4. *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 5. *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 6. *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 the following irrelevance result, that the investor's assessment is independent of the auxiliary parameters.

Proposition 7. *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.*

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 5 does not hold and we would need to enrich the statement 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 5.

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) and uncorrelated with $\boldsymbol{m}^T$. Therefore, when the investor updates $\mu(\boldsymbol{\tau}|\boldsymbol{m})$, he does no longer update over $\boldsymbol{m}^T$. The advisor cannot tell him anything he does not already know. Therefore, in the SYMMETRIC scenario of our experiment, no persuasive cheap talk equilibrium exists.

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 $\mu^\psi(\boldsymbol{\tau}|\boldsymbol{m}^A)$. 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 $\boldsymbol{m}^A$:

$$\mathbb{E}_{\mu^\psi}(\widetilde{\theta}^T_{post}|\boldsymbol{m}^A) = \psi\, \nu(\cdot) + (1-\psi)\mathbb{E}_\mu(\theta^T_{post}|\boldsymbol{m}^A). \tag{12}$$

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

$$\theta_H = \psi\theta^A_{post} + (1-\psi)\mathbb{E}_\mu(\theta^T_{post}|\boldsymbol{m}^A). \tag{13}$$

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 (11). When we solve Equation (13) 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.[39]

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

[39]Since ours is a sequential game, we adopt the Cursed Sequential Equilibrium (CSE) solution concept developed by Fong, Lin, and Palfrey (2025). Eyster and Rabin (2005) develop a cursed equilibrium concept for simultaneous games.

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 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 $\psi-$cursed equilibrium, advisor and investor maximize their utility and the investor uses $\psi-$*cursed* Bayes' rule to update beliefs whenever possible.

Consider an $\psi-$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 (τ). His assessment will then be equal to his expectation of θ^T_{post} given this updated belief:

$$\begin{aligned}\mathbb{E}_{\mu^{\psi}}(\widetilde{\theta}^T_{post}|\boldsymbol{m}^A) &= \int_{\tau\in\mathcal{T}}\theta^T_{post}\mu^{\psi}(\tau|\boldsymbol{m}^A)\,\mathrm{d}\tau \\ &= \psi\bar{\theta} + (1-\psi)\mathbb{E}_{\mu}(\theta^T_{post}|\boldsymbol{m}^A).\end{aligned} \tag{14}$$

This is equal to the conditional expectation in Equation (12) 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 (14) 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.[40] 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

[40]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.

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_2^R. The sender has utility function $u^S(\omega_2^R) = \omega_2^R$ and the receiver has utility function $u^R(\omega_2^R) = -(\omega_2 - \omega_2^R)^2$. 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 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 2

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

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

and will denote differences in the number of successes in *post* under the structural change parameters c' and c^{DO} by $\Delta k = k'_{post} - k''_{post}$ and $\Delta f = f'_{post} - f''_{post}$. Denote the log likelihood function of $\boldsymbol{m}$ given a specific dataset by $\ell(\boldsymbol{m})$ and define a function that returns the log likelihood fit difference between messages $\boldsymbol{m}'$ and $\boldsymbol{m}''$ 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 $\ell(\boldsymbol{m}^{DO})$, $\Delta\ell(\theta^{DO}_{post}) < 0$. The derivative is equal to

$$\Delta\ell'(\theta_{post}) = \frac{\Delta k}{\theta_{post}} - \frac{\Delta f}{1-\theta_{post}}. \tag{15}$$

Furthermore, as θ_{post} becomes large,

$$\begin{aligned}
\lim_{\theta_{post}\to 1} \Delta\ell(\theta_{post}) = & \Delta k (\lim_{\theta_{post}\to 1} ln(\theta_{post}) - ln(\theta'_{pre})) + \Delta f (\lim_{\theta_{post}\to 1} ln(1-\theta_{post}) - ln(1-\theta'_{pre})) + \kappa \\
= & -\Delta k ln(\theta'_{pre}) + \Delta f (\lim_{\theta_{post}\to 1} ln(1-\theta_{post}) - ln(1-\theta'_{pre})) + \kappa
\end{aligned} \tag{16}$$

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

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

When does this condition hold? Define a function

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

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

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

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

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

Case 1: $\Delta k > 0$, $\Delta f = 0$. As $\theta_{post} \to 1$, $\Delta\ell(\theta_{post}) > 0$ (see Equation (16) and the discussion afterwards). Furthermore, the derivative in Equation (15) 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 (16) and the discussion afterwards). Furthermore, the derivative in Equation (15) 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 (16) and the discussion afterwards). Furthermore, the derivative in Equation (15) 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 (16) and the discussion afterwards). Furthermore, the derivative in Equation (15) 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 (16) and the discussion

afterwards). Furthermore, the derivative in Equation (15) 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 (16) and the discussion afterwards). Furthermore, the derivative in Equation (15) 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 3

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 $w_1, w_2, \ldots$ 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, $\frac{\partial \mathbb{E}_i(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \theta_H)}{\partial \theta_H} < 1$,[41] 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 4

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 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 $\boldsymbol{m} \in \dot{\mathcal{M}}^{max}$:

$$\begin{aligned}
&\hat{\theta}_H(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H) \equiv \dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}_H) + \dot{r}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}^{\rightarrow}_H) \qquad (17)\\
&+ (1 - \dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H) - \dot{r}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H))\bar{\theta},\\
&\text{where } \dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H) = \frac{3\lambda(1 - G_{\theta_{post}}(\dot{\theta}_H))}{3\lambda(1 - G_{\theta_{post}}(\dot{\theta}_H)) + (1-\lambda)(1 - G_{\theta_{post}}(\dot{\theta}^{\rightarrow}_H)) + (1-\lambda)},\\
&\text{and } \dot{r}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H) = \frac{(1-\lambda)(1 - G_{\theta_{post}}(\dot{\theta}^{\rightarrow}_H))}{3\lambda(1 - G_{\theta_{post}}(\dot{\theta}_H)) + (1-\lambda)(1 - G_{\theta_{post}}(\dot{\theta}^{\rightarrow}_H)) + (1-\lambda)}.
\end{aligned}$$

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 $\dot{\theta}_H$;

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

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

The sum of the first two derivative terms is negative if

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

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 3 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 $\dot{\theta}^{\rightarrow}_H$. The fixed point solves

$$\begin{aligned} h(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H) = \dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}_H) + \dot{r}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H)\mathbb{E}_\mu(\widetilde{\theta}^T_{post}|\theta^T_{post} \geq \dot{\theta}^{\rightarrow}_H) \\ + (1-\dot{q}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H) - \dot{r}(\dot{\theta}_H, \dot{\theta}^{\rightarrow}_H))\bar{\theta} - \dot{\theta}_H \equiv 0. \end{aligned} \tag{18}$$

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

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

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

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

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

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

Plugging in and simplifying the inequality above yields

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

Note that we can rearrange Equation (17) to

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

which we can use to simplify the inequality to

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

Therefore, $\frac{\partial \dot{\theta}^*_H}{\partial \dot{\theta}_H^{\rightarrow}} \geq 0$ 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 $\hat{\dot{\theta}}_L(\dot{\theta}^*_L, \dot{\theta}_L^{\rightarrow}) = \dot{\theta}^*_L$. 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 $\theta^T_{post} \geq \dot{\theta}_H^{\rightarrow}$. Similarly, use $\dot{\theta}_L^{\rightarrow}$ to denote the threshold value such that the aligned advisor sends a message in $\mathcal{M}^{min}$ if $\theta^T_{post} \leq \dot{\theta}_L^{\rightarrow}$. 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 (18) 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

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

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

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

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)

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

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 5

Proposition 4 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

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

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

Claim 2: In any most informative equilibrium, the misaligned advisors do not send messages with $\theta^{A}_{post} \in (\theta_L, \theta_H)$. 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 $\theta^{A}_{post} \in (\theta_L, \theta_H)$ 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 $\theta^{T}_{post} \in (\theta_L, \theta_H)$, the honest/aligned advisor induce assessment θ^{T}_{post}. 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 6

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

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

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 3 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.