Narrative Persuasion

A populated preregistration: Mapping to the main text

Kai Barron
WZB Berlin

Tilman Fries
LMU Munich

April 11, 2024

This is a document reconciling the analysis in our paper *Narrative Persuasion* with our preregistrations. Click here to jump to the paper.

Contents

1 Introduction

The experiments reported in this paper were preregistered in two preregistration documents uploaded to the AEA-Registry under the unique identifiers AEARCTR-0009103 and AEARCTR-0011565 (Barron & Fries, 2022, 2023). We will refer to them as *Preregistration 1* and *Preregistration 2* in the following. Each of the preregistration entries includes a pdf document (which are also attached to the end of this document). Together, these two pdf documents specify: *(i)* the experimental design of all the treatments reported in this paper, *(ii)* the sample size that we planned to collect for each treatment, and *(iii)* our ex-ante hypotheses and plans for the analysis.

The purpose of this document is to provide a map between the preregistration documents and the paper. We view the preregistration documents as serving two main purposes. First, they serve to tie our hands regarding the features of the experimental design and sample size we planned to collect. Second, they provide a detailed snapshot of our ex-ante hypotheses and planned analysis for the experiment. We do not take the approach where a preregistration provides a detailed recipe prescribing in a step-by-step fashion exactly how the final paper will be written. Therefore, the main text of the paper is not a one-to-one mapping from the preregistration documents (aside from anything else, this would have resulted in a much longer and less focused paper). However, equally, we believe that it is crucial to be fully transparent about reporting *all* the pre-registered results, such that the interested reader has access to all of this information and can evaluate the results reported in the paper with full knowledge of all the planned ex-ante analyses (with nothing hidden or suppressed). This allows the reader to evaluate the results with full information and thereby draw informed conclusions about their validity.

Our reasons for not simply following the pre-registration document step-by-step were partially expositional and partially because we gained new insights while working on the project that allowed us to sharpen and improve the focus of the paper (sometimes by including additional analysis that we did not anticipate ex-ante—e.g., analyses suggested by seminar audiences). Our main hypotheses are theory-driven, emanating from the theoretical work of Schwartzstein & Sunderam (2021). This means that the set of hypotheses we are testing is constrained and the core ideas we test remained constant throughout the project.

The deviations in the main text relative to the preregistration include the following. First, we adjusted the names of several treatments. The reason for this was that we converged on names that we thought better captured the essence of some treatments, thereby providing the reader with a better mnemonic for recalling the details of each treatment. This document provides a dictionary of these name changes. Second, we preregistered a set of hypotheses that captured our ex-ante planned analysis. However, in writing the final paper, we have chosen to focus our discussion more and organize it around four theoretical predictions. This implies that we do not report all the results from the preregistered hypothesis tests in the main text. Below, we provide a full description of all of these preregistered hypotheses. In addition, when

we deviate from the preregistered analysis in the main text, we discuss the reasons for doing so here (along with reporting the preregistered analysis). An overarching theme is that the two preregistration documents and the final version of the paper reflect snapshots of our thinking at different points in time. The exposition of the paper is in some ways more mature than that of the pre-analysis plans, as the paper has subsequently been informed by extensive feedback and discussions with colleagues.

This document can be read as tracing the development of our thinking over time and providing a "populated" preregistration plan (Banerjee et al., 2020). Our original preregistration plans can be found in the Appendix at the end of this document.

2 Preregistration 1 (16 March 2022)

In Barron & Fries (2022), we preregister the design and analysis of the ASYMMETRIC, DISCLOSURE, SEQUENTIAL, and PRIVATEDATA treatments (using the treatment names as specified in the preregistration document—see Table 1 for the translation to treatment names in the paper). The preregistration document specifies a set of main hypotheses that investigate whether meeting a misaligned instead of an aligned advisor moves an investor further away from the truth and that compare the investor's distance from the truth when meeting a misaligned advisor between ASYMMETRIC and each of the intervention treatments. The prespecified secondary hypotheses address *persuasion mechanisms*, such as the role of the empirical fit.

In the write-up of the paper, we assign relatively more prominence to the mechanisms than to the interventions part. The reason for doing this is that we came to appreciate over time that it was important to first provide empirical evidence regarding the fundamental questions related to narrative persuasion (What are the mechanisms governing persuasion?) before answering the more applied questions (What interventions can we find against harmful persuasion?). This insight was present in much of the helpful feedback we received from colleagues who read early drafts and attended presentations of the paper. This also guided us towards deciding to investigate these underlying mechanisms further using additional treatments—these are the treatments described in Preregistration 2, which we discuss below.

2.1 Mapping Terms and Terminology

Table 1 contains the names that we use to refer to each of the treatments specified in Preregistration 1 in the paper (right column) and in the preregistration document (left column). The names that we use to refer to the treatments in the paper are better at capturing a distinctive *design feature* of each treatment and should provide a more intuitive mnemonic for the reader when considering the full set of treatments described in Preregistration 1 and 2 combined. For example, the term ASYMMETRIC provides a clear and intuitive distinction between this treatment's information environment in comparison to SYMMETRIC (described in Preregistration 2).

Similarly, the updated treatment name, DISCLOSURE, highlights this treatment's distinguishing feature (the disclosure of advisors' incentives to investors), while the name that we used in Preregistration 1 for this treatment, SKEPTICISM, is less suitable as it refers to a possible implication that such disclosure may have (making investors more skeptical).

In the text below, we typically use the treatment names adopted in the paper rather than those in the preregistration document.

Table 1: Names used to refer to treatments in Preregistration 1 and in the paper

Preregistration 1	**Paper**
BASELINE	ASYMMETRIC
SKEPTICISM	DISCLOSURE
SEQUENTIAL	INVESTORPRIOR
PRIVATEDATA	PRIVATEDATA

A further note relating to terminology that is worth keeping in mind is that throughout the paper, we refer to the two participant roles as *advisor* and *investor,* while we refer to them as *sender* and *receiver* in the preregistration. (In the instructions of the experiment, we used *advisor* and *investor*, so the terminology in the paper is consistent with the instructions of the experiment.)

Table 2: Names used to refer to agents in Preregistration 1 and in the paper

Preregistration 1	**Paper**
sender	advisor
receiver	investor

2.2 Preregistered Hypotheses

Hypothesis 1 of Preregistration 1 postulates that investors will be persuaded to shift their assessments by advisors. Specifically, it posits that the distance between the investor's assessment and truth will depend on the incentives of the advisor they meet with.

Preregistration 1: Hypothesis 1. *In ASYMMETRIC, the distance between the investor's assessment and the truth is larger when advisor incentives are misaligned than when advisor incentives are aligned.*

Relation to the results discussion in the paper—The paper discusses the influence of advisor incentives on assessments in the context of Result **??** (Persuasion in pure interpretation and hybrid scenarios). In particular, the left panel of Figure **??** shows that (i) assessments after meeting the aligned advisor are centered around the truth (ii) they are, on average, higher after meeting an up-advisor, (iii) they are, on average, lower after meeting a down-advisor. In the paper, we present results for each advisor type (up, down, aligned) instead of using

the coarser categorization of advisor alignment (aligned, misaligned). This more fine-grained disaggregation of the analysis provides a more detailed understanding of the direction of persuasion than the analysis proposed in the preregistration: Rather than only demonstrating that assessments are further away from the truth when meeting a misaligned advisor (relative to an aligned advisor), we demonstrate that they are higher after meeting an up-advisor and lower after meeting a down-advisor. We report the results of the preregistered test of Hypothesis 1 in Table 3 below. The results support the hypothesis.

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

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

Notes: (i) The dependent variable is the absolute distance between the investor's belief, θ^{I}_{post}, and the true value θ^{T}_{post}, (ii) The sample contains data from all investors in ASYMMETRIC, (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; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

Hypotheses 2-4 of Preregistration 1 each take the following form:

> *When matched with an advisor with misaligned incentives, the distance between the investor's assessment and the truth is smaller in [INTERVENTION] than in ASYMMETRIC.*

The filler term, [INTERVENTION], refers to the DISCLOSURE, INVESTORPRIOR, or PRIVATEDATA treatments, respectively.

Relation to the results discussion in the paper—The preregistered tests of these hypotheses are included in the (*a)-columns of Table **??** in the main text of the paper. We do not find evidence in support of any of these hypotheses. When discussing these results, we present additional tests for potential heterogeneous treatment effects in the (*b)-columns of the same table. While we did not preregister this heterogeneity analysis, we find it valuable as it helps to illuminate potential mechanisms behind the null findings.

Hypothesis 5a addresses the role of narrative fit for persuasion (via the empirical plausibility index, or "EPI", measurement of fit, which is defined in the main text and in the preregistration document):

Preregistration 1: Hypothesis 5a. *The distance between the advisor's message and the investor's assessment decreases in the EPI.*

Relation to the results discussion in the paper—We present evidence related to the persuasive role of fit (measured by the EPI) in the discussion of Result **??** (Relevance of narrative fit). In particular, Column (1) of Table **??** reports the preregistered test of the hypothesis, which is in support of the hypothesis. As we came to better appreciate that cleanly testing for the influence of narrative fit was a core insight that the paper could contribute to the literature, we conducted the following additional exercises to provide robustness checks for the role played by narrative fit. First, we used data from the InvestorPrior treatment, where we also collected prior beliefs of investors, to test whether the relationship between the narrative's fit (EPI) and the investor's assessment is robust to controlling for prior beliefs. Second, as specified in Preregistration 2, we collected additional data in our Competition treatment to check whether the influence of narrative fit on persuasion is still observed in a setting where we *exogenously* vary the fit of the message. All of these exercises support the idea that narrative fit matters for persuasion.

Hypothesis 5b is a second sub-hypothesis related to the influence of narrative fit. Its objective is to check for the presence of one potential force that may moderate the relationship between the message's EPI and the investor's assessment: Persuasion may be less effective when the objective data admits multiple different narratives that all fit well. The basic idea is as follows. If the data is consistent with multiple distinct compelling explanations (many narratives fit well), then the investor might already hold on to a well-fitting default narrative before receiving the investor's message and it may be difficult for self-interested advisors to persuade investors to shift their beliefs, since it is harder for the advisor to beat the default explanation. However, when the compelling explanations that the data accommodates are all similar ("close") to one another, then, for a given level of narrative fit, the investor may be persuaded more easily. This may allow the advisor to be more persuasive in shifting the investors' beliefs a little in the direction they want by proposing an explanation that is "close" to the best-fitting one, but adjusted a little to suit the advisor's objectives.

We study the impact of the availability of alternative narratives by examining whether the shape of the EPI function, taken across all possible values of θ_{post}, affects persuasion—measured as the distance between the advisor's message and the investor's assessment, $D^S(\theta^I_{post})$. The EPI function is *single-peaked* in cases where the data entertains only a single relatively salient data-optimal narrative (and all other reasonably compelling narratives are near to this one); it has *multiple peaks* when the data provides room for multiple competing narratives that fit relatively well. We hypothesize that, if the history of company outcomes (i.e., the objective data) can be equally well explained by different narratives, the investor is less easily swayed by the advisor's model. The rationale behind this hypothesis is that when the EPI has multiple peaks, the investor can more easily entertain alternative prior narratives that explain the data similarly well. Therefore, we conjecture that the distance between the advisor's message and the investor's assessment is higher if the EPI has multiple local optima when considering

all possible values of θ_{post}.[1] To adjust for the possibility that different kinds of objective data histories (i.e., those allowing for single-peaked vs multi-peaked EPI functions) will allow advisors to systematically construct higher or lower-fitting narratives, the hypothesis states that the presence of multiple peaks will matter after controlling for the fit of the advisor's message.

Preregistration 1: Hypothesis 5b. *Conditional on the value of the EPI evaluated at the advisor's model, the distance between the advisor's message and the investor's assessment is smaller if the EPI has a single global optimum than if it has multiple local optima.*

Relation to the results discussion in the paper—Table 4 shows the results from the preregistered test of this hypothesis using data from ASYMMETRIC. The results support the hypothesis.

We decided against discussing this hypothesis and the corresponding results in the paper; discussing it would have required us to introduce an additional concept in the main text (the curvature of the log likelihood) and to provide a discussion of scenarios with multiple vs. single plausible explanations. During the course of working on the project, it became apparent that discussing this hypothesis in the main text would lengthen the paper and distract from the main thread of the study, without meaningfully adding value for the reader. Therefore, we opted to omit discussing this result in the interest of streamlining the paper.

Table 4: The role of multiple plausible explanations.

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

Notes: (i) The dependent variable is the absolute distance between the investor's belief, θ^{I}_{post}, and the sender's value for θ^{A}_{post}, (ii) The sample contains data from all investors in ASYMMETRIC, (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; * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$.

We have two hypotheses, 6a and 6b, on narrative construction. These hypotheses are based on the observation that the incentives of advisors, in combination with the movement-fit trade-

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

off, will lead (i) up-advisors to increase θ_{post} and decrease θ_{pre}, while (ii) down-advisors will decrease θ_{post} and increase θ_{pre}. The hypotheses are as follows:

Preregistration 1: Hypothesis 6a. *The distance between the advisor's message and the truth of the post report, $D^T(\theta^A_{post})$, is larger for misaligned advisors than for aligned advisors.*

Preregistration 1: Hypothesis 6b. *The distance between the advisor's message and the truth of the pre report, $D^T(\theta^A_{pre})$, is larger for misaligned advisors than for aligned advisors.*

Relation to the results discussion in the paper—In the main text of the paper, the discussion of Result **??** (Fit-movement tradeoff in narrative construction) provides evidence related to these hypotheses. There, Figure **??** and Table **??** provide evidence that up-advisors send a higher θ_{post} and lower θ_{pre} than aligned advisors (and vice versa for down-advisors). In this discussion in the main text, we deviate a little from reporting the precise preregistered analysis because it seemed more informative to provide a more fine-grained disaggregation of the analysis by advisor incentive type (up, down, aligned) instead of advisor alignment (aligned, misaligned). [The rationale is similar to our discussion above of a similar expositional decision for Hypothesis 1.] This allows us to provide a more thorough test of some of the theoretical predictions. For example, we can test the prediction that up-advisors *increase* θ_{post} and *decrease* θ_{pre} rather than only testing whether they move these parameters *away from* the truth.

Table 5 reports the results of the preregistered tests of both hypotheses. The results support both hypotheses.

Table 5: Distance from the truth of narratives proposed by misaligned vs aligned advisors

	(1) $\lvert\theta^A_{post} - \theta^T_{post}\rvert$	(2) $\lvert\theta^A_{pre} - \theta^T_{pre}\rvert$
Misaligned advisor = 1	12.72***	6.492***
	(0.702)	(0.660)
Dep. var. aligned adv. mean	1.478	1.929
Round FE	Yes	Yes
Observations	3600	3600

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

The hypothesis section in Preregistration 1 ends with two hypotheses for ASYMMETRIC that investigate the narrative construction of aligned advisors, and the role of truth-telling preferences in narrative construction. The aligned advisor knows that the investor will compare the narrative she sends to the objective data to assess how convincing it is. In the absence of truth-telling preferences, the aligned advisor has no interest in reporting the true model, but

rather wants to send a message that: (i) fits the data well, and (ii) induces a belief that is close to the truth. If the message that fits the data best induces a belief in the investor that is "close" to the truth, θ^{T}_{post}, the advisor may wish to send this data-optimal narrative to the investor. This logic suggests that when the exogenous variation in the historical data is such that that true model does not actually fit the data well—i.e., the data-optimal value θ^{DO}_{post} is far from the true value θ^{T}_{post}—aligned advisors will send a message that contains a θ^{A}_{post} value that is further from θ^{DO}_{post}. In other words, we hypothesize that the average aligned advisor will follow a strategy that involves sending a θ^{A}_{post} that is a weighted average of the truth, θ^{T}_{post}, and the data-optimal narrative, θ^{DO}_{post}.

Preregistration 1: Hypothesis 7a. *The distance between the data-optimal model and the aligned advisor's message, $|\theta^{A}_{post} - \theta^{DO}_{post}|$, increases in the distance between the truth and the data optimal message, $|\theta^{T}_{post} - \theta^{DO}_{post}|$.*

For the misaligned advisors, the true model should not play a role unless truth-telling preferences influence the narratives they construct. Misaligned advisors face monetary incentives to draw the investor's belief away from the truth. They are constrained only by the investor's information set (i.e.,the historical data) and their own truth-telling preferences. If they hold no truth-telling preferences, they will completely disregard the truth and it will play no role in influencing the narrative they construct. In the following hypothesis we check (a) whether truth-telling preferences influence misaligned advisors, and (b) whether the size of this influence (pull towards the truth) is smaller than it is for aligned advisors.

Preregistration 1 Hypothesis 7b. *The distance between the data-optimal model and the misaligned advisor's message is governed to a lesser extent by the size of $|\theta^{T}_{post} - \theta^{DO}_{post}|$ than in the aligned advisor's message.*

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

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

In the equation above, $\mathbb{I}$(Misaligned) is an indicator function which takes a value of 1 if the advisor's incentives are misaligned, ρ_r are round fixed effects and ε is an error term.[2] Table 6 reports the results. We test Hypothesis 7a by examining β_2. Since β_2 is statistically greater than 0, we find evidence in support of Hypothesis 7a. Specifically, the aligned advisor's message

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

is biased away from the data-optimal model towards the truth. This indicates that aligned advisors are motivated both by their monetary incentives and also by truth-telling preferences. The magnitude of this coefficient suggests that truth-telling is the dominant approach adopted by aligned advisors.

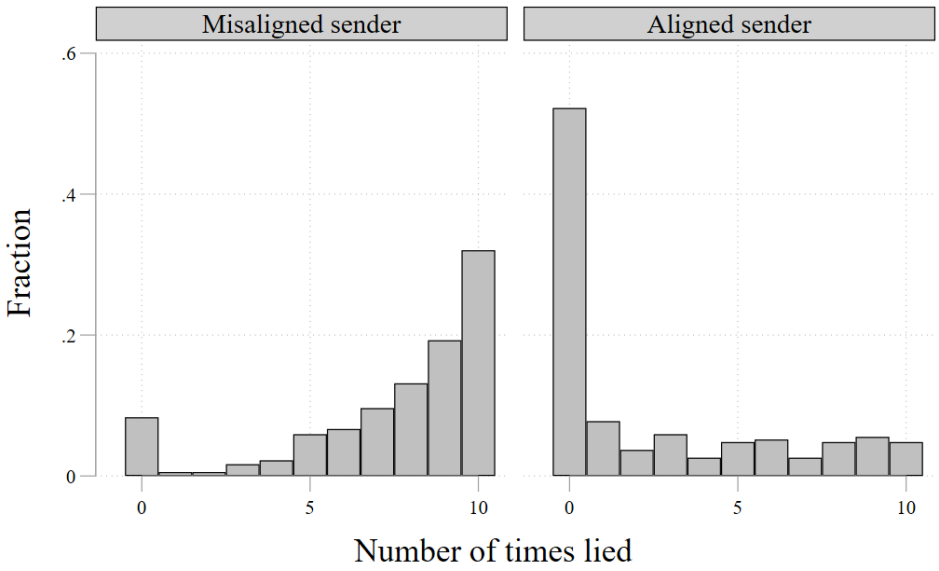
Table 6: The influence of the truth on advisor narratives

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

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

In support of Hypothesis 7b, we find that misaligned advisors respond less strongly to the truth than aligned advisors ($\beta_3 < 0$). However, misaligned advisors do not ignore the truth completely—on average, they do still adjust their narratives towards the truth, even though they are not incentivized to do so ($\beta_2 + \beta_3 > 0$). One potential explanation for this is that (some) advisors hold truth-telling preferences that are sufficiently strong to induce them to tell the truth in some rounds. We find support for this when we calculate the number of rounds in which each advisor lied, as displayed in Figure 1. We see that while the vast majority of misaligned advisors lied in more than five rounds, fewer than 40% lied in all ten rounds. This suggests that a majority of advisors hold some truth-telling preferences.

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



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

Relation to the results discussion in the paper—We decided against discussing these results in the paper in the interest of keeping the exposition of the paper more concise and focused.

2.3 Additional Analysis

Since we wanted the preregistration document to provide a comprehensive ex-ante snapshot of our thinking, Preregistration 1 also mentions additional plans for analysis, including conducting additional non-parametric tests of hypotheses 1 – 4 and an analysis of investor belief formation. In the interest of being concise and focused, we have limited the discussion here to presenting the preregistered analyses that we specified as the main tests of the preregistered hypotheses. Many of the other exercises are now redundant due to being superseded by other analyses presented in the paper. All of these results are, however, available on request.

3 Preregistration 2 (12 June 2023)

In Barron & Fries (2023), we preregister the design, sample size, and analysis of SYMMETRIC, COMPETITION, EXPLANATION, and NOEXPLANATION. We decided to add these treatments to gain further insights into the mechanisms governing persuasion in our setup.

3.1 Mapping Terms and Terminology

Table 7 shows how the names that we use to refer to the treatments in the preregistration correspond to the names that we use to describe the treatments in the paper.

Table 7: Names used to refer to treatments in Preregistration 2 and in the paper

Preregistration 2	**Paper**
3Parameters	Explanation
1Parameter	NoExplanation
Symmetric	Symmetric
CompetingNarratives	Competition

3.2 Preregistered Hypotheses

The first hypothesis of the preregistration postulates that, on average, the investors in the Explanation treatment will be more persuaded by the messages they receive (claims with explanations) than the investors in the NoExplanation treatment (claims without explanations).

Preregistration 2: Hypothesis 1 *The advisor's message will influence the investor's final assessment more in Explanation than in NoExplanation.*

We present results related to this hypothesis in Table **??** in the main text. In particular, Column (1) of Table **??** presents the results of the preregistered test, which does not provide support for the hypothesis. We then present additional evidence on heterogeneous treatment effects by examining the influence of the *quality* of the explanation—i.e., asking whether good explanations differ in persuasiveness from bad explanations. We do this by interacting the Explanation treatment dummy with a measure of fit. We preregistered conducting a heterogeneity analysis of this nature that examines the influence of the *quality* (fit) of the explanations—we hypothesized that we would find a negative coefficient estimate for the interaction term (which is what we find and report in the paper). There is, however, one important discrepancy between the analysis that we preregistered and the analysis that we report in the paper. We preregistered that we would study heterogeneous treatment effects by using the EPI (Empirical Plausibility Index) as a measure of heterogeneity in the fit of the explanations. However, we later realized that it is more appropriate to measure the fit of explanations using the APS (Auxiliary Parameter Score) that we define in the paper. The key reason for this is that the EPI measures the fit of the entire narrative (i.e., the claim, θ^A_{post}, and the explanation, c^A and θ^A_{pre}), while the APS measures the fit of the explanation directly (i.e., the fit of c^A and θ^A_{pre} conditional on θ^A_{post}).

To understand how these two different measures of fit could affect the heterogeneity estimation results, it is useful to think about how the interaction parameter between treatment and fit is identified. Recall that we use exactly the same historical company data and advisor messages in Explanation and NoExplanation. Therefore, we can compare pairs of investors

in the two treatments where the only difference is that in NoExplanation, investors only receive the claim, θ^A_{post}, while in Explanation, they receive the claim, θ^A_{post}, plus the explanation. Identification of heterogeneity relies on examining whether the *posterior persuasion gap*, $|\theta^{I,1}_{post} - \theta^A_{post}|$, depends on the fit in the Explanation treatment, controlling for everything else by comparing with a matched investor in NoExplanation who faces an identical choice minus the explanation. To identify the influence of the auxiliary (explanation) parameters, ideally we would like to compare instances where two investors see the same historical company data and receive the same θ^A_{post}, so that the only difference between messages is exogenous variation in the auxiliary (explanation) parameters. We are able to do this in our Competition treatment, but not in the explanation treatments, where we rely on organically constructed messages. This means that there can be endogenous relationships between the data, claims, and explanations. Therefore, the EPI does not provide a clean measure of the quality of explanations because it captures variation in the fit of both the explanation *and* the claim, θ^A_{post}. This is important because the explanations are not exogenous to the claim, θ^A_{post}. One might, therefore, expect a correlation between fit and EPI already in NoExplanation, because some θ^A_{post}—values are more coherent with the data than others. The APS is a cleaner measure of auxiliary parameter fit as it measures fit conditional on θ^A_{post}—i.e., it more precisely captures variation in fit due to the explanation parameters, c^A and θ^A_{pre}.[3]

The empirical results are in line with these considerations. Table 8 presents results that allow for a comparison of both measures. Column (1) reproduces the heterogenous treatment effect result presented in Table **??** in the main text. Column (2) presents results from an identical specification that substitutes the APS with the EPI. The interaction term becomes insignificantly different from zero, which is consistent with the idea that a lot of the influence of the variation in the EPI is already present in NoExplanation. The last two columns remove the Round×linked investor fixed effects, which allows us to identify the non-interacted (i.e., NoExplanation) fit coefficients. Columns (3) and (4) suggest that both fit measures are related to the posterior distance in NoExplanation. However, this relation is stronger when we use the EPI instead of the APS. This is in line with our discussion above. It also suggests to us that substituting the EPI with the APS to investigate a potential heterogeneous treatment effect of introducing explanations is appropriate.

[3]It is worth noting that it also makes sense that the APS is correlated with the posterior persuasion gap in NoExplanation. The reason for this is precisely because the the claim, θ^A_{post}, and explanation, c^A, and θ^A_{pre}, are chosen organically by advisors and may therefore be endogenous. For example, if advisors who choose claims that fit the data well also choose explanations that fit well, this could result in a negative correlation between the APS and posterior persuasion gap in NoExplanation. Our identification strategy accounts for this by holding constant the historical data and the claim, θ^A_{post}, by comparing Explanation and NoExplanation and then examines the variation in only the auxiliary (explanation) parameters using the APS.

Table 8: Comparison of heterogenous treatment effect estimates using EPI and APS fit measures

	(1) Posterior distance	(2) Posterior distance	(3) Posterior distance	(4) Posterior distance
Prior distance	0.365***	0.366***	0.482***	0.470***
	(0.0271)	(0.0272)	(0.0253)	(0.0253)
Explanation	3.078*	0.00886	2.812	0.0195
	(1.574)	(0.762)	(1.717)	(0.796)
Explanation × APS	-3.855**		-3.491*	
	(1.811)		(1.997)	
APS			-3.242**	
			(1.258)	
Explanation × EPI		0.120		0.152
		(1.392)		(1.494)
Message EPI				-6.036***
				(0.962)
Round× linked investor FE	Yes	Yes	No	No
Round FE	No	No	Yes	Yes
Observations	3600	3600	3600	3600

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

Hypothesis 2 of the preregistration considers the impact of advisor incentives on persuasion in Symmetric.

Preregistration 2: Hypothesis 2 *In Symmetric, up-advisors persuade investors to increase their assessment and down-advisors persuade investors to decrease their assessment, relative to aligned advisors.*

Relation to the results discussion in the paper—We present results related to this hypothesis in the discussion around Result **??** (Persuasion in pure interpretation and hybrid scenarios). A pre-registered test of this hypothesis is displayed in the right panel of Figure **??** and in Column (2) of Table **??**. These results provide support for the hypothesis.

In the last hypothesis of Preregistration 2, we address persuasion within Competition.

Preregistration 2: Hypothesis 3. *In CompetingNarratives, the investor is more likely to adopt the human advisor's narrative if the robot advisor picks the auxiliary parameters randomly.*

Relation to the results discussion in the paper—We provide a discussion of this hypothesis in the main text when describing Result **??** (Influence of narrative fit). In particular, Column (1) of Table **??** presents a preregistered test, which provides evidence in favor of the hypothesis.

3.3 Additional Analysis

In Preregistration 2, we set out two additional plans for analysis. The first was that we wanted to measure how often, in Round 1 of COMPETITION, the investor chooses the human advisor's narrative over the robot advisor's narrative (which is always equal to the true model). We present the results of this analysis in the Introduction section of the paper. Second, we planned to study whether the human advisor's response to the robot's narrative in COMPETITION is in line with the fit-movement tradeoff. We mention the results of this exercise in the main text in the discussion around Result **??** (Responding to a competing narrative).

In the main text, the structural model that we estimate in Section **??** (Accounting for decision noise) using data from COMPETITION was not preregistered. We decided to investigate the role of noise in our setting after receiving questions related to it when presenting the paper.

References

Banerjee, A., Duflo, E., Finkelstein, A., Katz, L., Olken, B., & Sautmann, A. (2020, April). *In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics* (Tech. Rep. No. w26993). Cambridge, MA: National Bureau of Economic Research. doi: 10.3386/w26993

Barron, K., & Fries, T. (2022, March). Narrative Persuasion. *AEA RCT Registry*.

Barron, K., & Fries, T. (2023, June). Narrative Persuasion 2. *AEA RCT Registry*.

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

Rosner, B., Glynn, R. J., & Lee, M.-L. T. (2006, December). Extension of the Rank Sum Test for Clustered Data: Two-Group Comparisons with Group Membership Defined at the Subunit Level. *Biometrics*, *62*(4), 1251–1259. doi: 10.1111/j.1541-0420.2006.00582.x

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

A Original Preregistration 1 Document

A.1 Introduction

Our experimental design takes inspiration from the ideas discussed in Schwartzstein & Sunderam (2021), however the primary objective of the experiment is not to test the Schwartzstein & Sunderam (2021) model. Rather, we aim to shed light on when and why persuasion using models is likely to occur and what factors can help to protect individuals from being persuaded in this way. We do this by testing the set of hypotheses described below using comparative static comparisons using the exogenous variation generated by our treatment conditions as well as the additional variation created by the experimental design.

Following Schwartzstein & Sunderam (2021) (S&S), we will consider a strategic setting in which there is a persuader / advisor (narrative-sender) and a receiver / investor (narrative-recipient). The receiver has access to data that is informative about the true underlying model. The persuader's objective is to propose a model to the receiver that guides the receiver in interpreting this data. The receiver then takes an action that influences the payoffs of both the persuader and the receiver. Importantly, the persuader's incentives may be either aligned or misaligned with the receiver's – i.e., the persuader might attempt to convince the receiver to take an action that does not serve her own best interests.

In this setting, we will investigate which factors influence the effectiveness of persuasion using models. Specifically, we will ask questions such as the following: *(1) Are receivers worse off when the sender's incentives are misaligned? (2) Does knowing the sender's incentives make receivers skeptical? (3) Does access to private data protect receivers? (Alternatively: Are persuaders less effective when they cannot construct ex post models that fit the receiver's available data?) (4) Are receivers better off if they are encouraged to make sense of the evidence themselves before they receive the sender's message? (5) Does the empirical plausibility of the sender's proposed model affect receiver's trust in the model?*

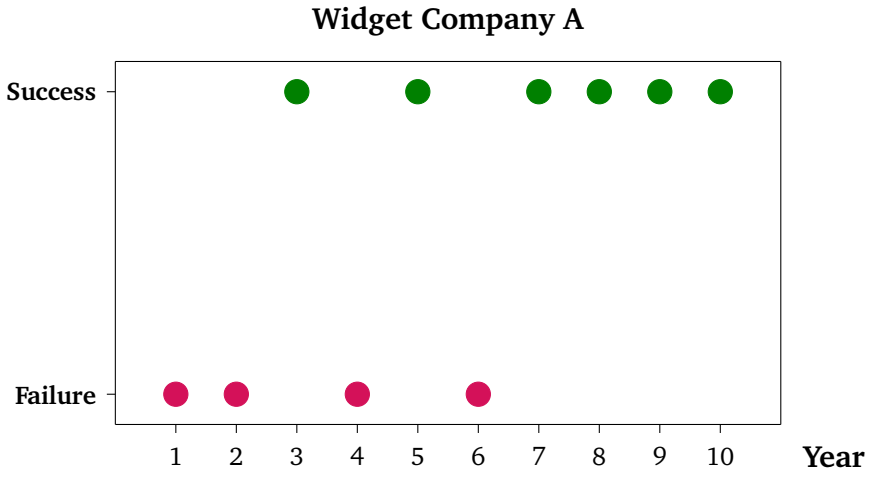
A.2 Experimental Design

In our experiment, we consider a two-player game where one player takes the role of *sender* and the other takes on the role of *receiver*. We frame our experiment using an investment game, such that the receiver is an *investor* whose objective is to assess the likelihood that a fictitious company will be successful (as opposed to unsuccessful) in the coming year. The experiment labels the coming year as "Year 11". The sender is an *advisor* to the investor, and will provide advice about the fictitious company.

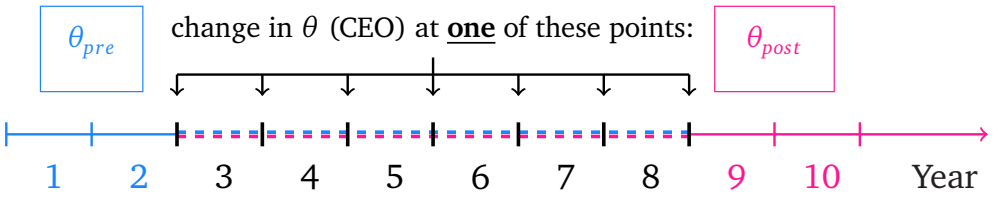
In each round of the experiment, the investor's objective is, therefore, to correctly assess the underlying state of the world (i.e., the likelihood of the company being successful in Year 11). To do this, the investor can draw on the information she observes about the history of success of the company.

However, prior to the investor reporting their assessment of the company's likelihood of success, the advisor sends a message to the investor. The advisor always knows the true model generating the data. In addition, in most treatment conditions, the advisor also observes the data that the receiver has access to. The advisor may use this message to try to persuade the investor to hold a belief that is biased in a certain direction by distorting the investor's interpretation of the data containing the history of past outcomes.

The Data Generating Process: The history of past outcomes consists of the past ten periods (years) of the company's performance. This data shows whether the company was "successful" or "unsuccessful" in each of the past ten years (i.e., from Year 1 to Year 10). The following provides an illustrative example of how one particular history could be represented.



In each year, the probability of the company being successful is determined by an underlying fundamental, θ. This fundamental changes exactly once during the ten years. More specifically, it is common knowledge that θ_{pre} is drawn from $U[0,1]$ prior to Year 1, and then is redrawn once at some point after Year 2 and before Year 9, denoted by $\theta_{post} \sim U[0,1]$. Therefore, the probability of success in each of the ten years is governed by $(\theta_{pre}, \theta_{post})$. In the experiment, we frame the change in the fundamental state as a change in the CEO of the company. The value of θ_{pre} then summarizes the probability of success for the period before the CEO changed (the pre period) and θ_{post} denotes the probability of success for the period after the CEO changed (the post period). The following figure illustrates the structure of the historical data.



Consequently, the last two periods in the historical dataset are commonly known to be (i)

governed by a different probability of success to the first two periods, and (ii) informative about the success probability of the company in Year 11.

To formalise the setup, let $c \in \{2,3,4,5,6,7,8\}$ denote the period before the the structural change (i.e., if $c = 2$, then the structural change occurred at the end of year 2, or equivalently, at the beginning of year 3). We specify a data generating process where c is uniformly distributed, which we also disclose to participants.[4] The variable c summarises the true model: it specifies that the last $10-c$ years of data are relevant for whether the company is successful under the new CEO. Therefore, θ_{pre} denotes the realised probability of success up to and including year c and θ_{post} denotes the realised probability of success after year c.

The Advisor's Additional Information: The advisor is fully informed about the underlying data generating model—i.e. the advisor knows the true values of the three fundamental parameters: $(c^T, \theta^T_{pre}, \theta^T_{post})$. The investor knows that the advisor has this additional information.

The Advisor's Message: The advisor sends three pieces of information to the investor: (i) an estimate $c^S \in \{2,3,4,5,6,7,8\}$ of the year when there was a structural change, and (ii) estimates $\theta^S_{pre} \in [0,1]$ and $\theta^S_{post} \in [0,1]$ of the success probability prior to and after the structural change, respectively.

The Investor's Decision: The investor observes the advisor's report $(c^S, \theta^S_{pre}, \theta^S_{post})$ and then submits her own estimate of θ^R_{post}.

The Investor's Incentives: The investor is incentivised to estimate θ_{post} as close as possible to θ^T_{post}. We will use the binarized scoring rule (Hossain & Okui, 2013) to ensure that the investor's payment will be maximized (in expectation) if she reports her true belief about θ_{post}.

The Advisor's Incentives: Participants in the experiment who are assigned the role of advisor will be randomly assigned into one of three incentive conditions. In all three conditions, the advisor's payment will be a function of their matched investor's θ_{post}-report. Under the three conditions, the advisor's payment is either: (a) increasing in the investor's estimate of θ_{post}, (b) decreasing in the investor's estimate of θ_{post}, or (c) increasing in the accuracy of the investor's estimate of θ_{post}. Each advisor keeps the same incentives for the duration of the experiment.

This will be incentivized using an strategic version of the binarized scoring rule (BSR), where there are two key differences from the standard BSR. First, the belief report that is relevant for determining the probability of receiving the bonus payment is made by *another individual*, not oneself. Second, in incentive conditions (a) and (b), the θ^S_{post} reported by the

[4] In other words, participants know that c is randomly drawn from a uniform distribution over $\{2,3,4,5,6,7,8\}$. (In the experiment, we frame this as being at the beginning of years 3 to 9, rather than at the end of years 2 to 8.) Participants also know that both θ_{pre} and θ_{post} are independently drawn from uniform distributions, $U[0,1]$. This is done independently for each of the ten companies (i.e., for each of the ten rounds of the experiment).

investor is compared to extreme θ_{post} values, $\theta_{post} = 1$ or $\theta_{post} = 0$ respectively, rather than being compared to the true θ_{post} to determine the advisor's payment. In incentive condition (c), the advisor's payoff is calculated in the same way as the investor's payoff (i.e., their incentives are perfectly aligned).

Strategic Information about Incentives: Investors are told about the different types of advisors that they may face. Specifically, they are told about the distribution of advisors with each of the three incentive types, namely that the probability of being matched with each advisor type in each round is one-third. In treatment SKEPTICISM, investors will additionally be informed about the incentives of their specific matched advisor in each round (more details below).

Advisors know the incentives of investors. In all treatment conditions, advisors are also always told that investors may or may not know their matched advisor's incentives.

General Comments about the Design: The basic idea of this design is that the advisor (in contrast to the investor) knows the underlying DGP $(c^T, \theta^T_{pre}, \theta^T_{post})$, which provides an opportunity for gains from communication between both players, since the advisor is more informed but the advisor's payoff depends on the investor's action. Depending on advisor's incentives, the advisor might sometimes try to deceive the investor into reporting an overly optimistic or pessimistic belief about θ_{post}. Specifically, the advisor can use the other dimensions of the report, (c^S, θ^S_{pre}), as supporting evidence for trying to shift this belief about θ_{post}.

We have chosen to deviate from S&S in that we usually do not elicit the investor's prior beliefs about the model (i.e., before persuasion); that is, her prior beliefs about $(c, \theta_{pre}, \theta_{post})$.

The reason for this is two-fold. First, we wish to study situations in which senders (advisors) present data to receivers (investors) at the same time as they communicate their theory explaining the data, as opposed to the receiver first constructing their own personal theory of the data. This conjunction of receiving the data and a potential theory at the same time reflects many real-world situations. Second, we wish to explicitly study whether being encouraged to form a personal theory of the data *prior* to receiving a potential theory from an advisor has a protective function in helping to insulate receivers (investors) from persuasion.

A.2.1 Treatment Conditions:

To address our research questions, we will consider four core between-subject treatment conditions.

BASELINE: Our BASELINE treatment follows the structure described above. The other three treatments involve small deviations from the BASELINE condition in which we exogenously vary one specific feature of the decision environment.

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

Skepticism: To investigate whether knowing their specific matched advisor's incentives makes investors skeptical, investors are made aware of the advisor's incentives. Because we are interested in investor behavior, we hold the advisor's information set constant by telling advisors also in this treatment that investors may or may not know their incentives.[5]

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

This treatment will allow us to evaluate whether being encouraged to try to make sense of the data oneself first serves a protective function against persuasion using models.[6]

A.2.2 Procedures:

The experiment will be conducted via the platform, Prolific. Participants will take part in the experiment in groups of 6. Within each group, 3 participants are randomly assigned to the role of the sender (advisor) and 3 are assigned to the role of the receiver (investor). Each advisor is randomly assigned to one of the three incentive conditions (i.e., there will be one advisor from each of the three incentive conditions within each group of 6). Both advisors and investors keep their role for the duration of the experiment; advisors additionally stay within their incentive condition throughout the experiment.

The experiment consists of ten rounds. In each round, each investor is randomly matched with an advisor within their group of six (i.e., the three investors are randomly matched with the three advisors).

[5]We control for investors' higher-order beliefs by informing them that advisors do not know that investors know what their matched advisor's incentives are.

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

Within each of the ten rounds, the true underlying data generating process will be held constant across all matched investor-advisor pairs. Specifically, the triple of fundamentals, $(c^T, \theta^T_{pre}, \theta^T_{post})$, is held constant within a specific round across all subjects.[7] However, conditional on these fundamentals, the observed historical data of success and failure of the company is drawn independently for each matched pair of participants. This provides us with exogenous variation in the data observed by subjects, conditional on a particular set of fundamentals governing the success of a the company in that round.

Participants are paid for one randomly chosen round of the experiment and do not receive any feedback until the end of the experiment. The absence of feedback implies that investor behavior cannot affect advisors. Advisors only influence investors directly through the messages they send. We, therefore, can implement the experiment in a simpler way where we first collect all advisor choices for the ten rounds, and thereafter collect investor choices.

A.3 Hypotheses and Analysis

A.3.1 Definitions and Measures

Our main hypotheses and analysis relate to the following objects that we collect in each round, for each matched sender-receiver pair:

(i) The sender's message, $(c^S, \theta^S_{pre}, \theta^S_{post})$.

(ii) The receiver's assessment, θ^R_{post}.

(iii) The realized historical dataset of successes and failures, $h = (\omega_1, \ldots, \omega_{10})$.[8]

In addition, we collect the fundamental parameters of the true data generating process, $(c^T, \theta^T_{pre}, \theta^T_{post})$, which vary across rounds, but are held constant across participants within a given round.

In order to organise the discussion of our hypotheses below, it will be useful to define some derivative measures that we can construct from this information. We organise these measures into three categories: (i) measures that compare a participant's message or assessment to the truth, (ii) measures that compare a participant's message or assessment to the observed historical data, and (iii) measures that provide an indication of the degree to which the sender is able to persuade the receiver (i.e., to shift their assessment).

Measures Relative to the Truth: The two measures of primary interest in this class are the distance between sender's message about θ_{post} and the true value, and the distance between the receiver's assessment of θ_{post} and the true value:[9]

[7] We will therefore randomly draw ten realizations of $(c^T, \theta^T_{pre}, \theta^T_{post})$; one for each round of the experiment. These will apply to all participants in all sessions of the experiment.

[8] Where $\omega_t \in \{0, 1\}$ with $P(\omega_t = 1) = \theta_{pre}$ if $t \leq c$ and $P(\omega_t = 1) = \theta_{post}$ if $t > c$.

[9] In addition, we construct an indicator variable that takes a value of one if the sender lies in their message, and zero if they tell the truth: $\mathbb{I}(\theta^S_{pre} \neq \theta_{pre} \vee c^S \neq c \vee \theta^S_{post} \neq \theta_{post})$

(i) Distance between the sender's message and the truth: $D^T(\theta^S_{post}) := |\theta^S_{post} - \theta_{post}|$

(ii) Distance between the receiver's assessment and the truth: $D^T(\theta^R_{post}) := |\theta^R_{post} - \theta_{post}|$

Measures Relative to the Historical Data: For each round and each matched pair of participants, the historical success data comprises ten realizations of the underlying data generating process in that round. It is therefore informative to compare participants' messages and assessments to the information that they observe.

To do this, for each observed dataset, we determine the data-optimal model, namely the model that is most likely to have generated the data, $(c^{DO}, \theta^{DO}_{pre}, \theta^{DO}_{post})$. Following S&S, we take the maximum likelihood estimate of $(c, \theta_{pre}, \theta_{post})$ for a given dataset h as the model that most likely generated the data. Given this data-optimal model, we can compare the message and assessment of the sender, $(c^S, \theta^S_{pre}, \theta^S_{post})$, to the optimum. We will do this by constructing an empirical plausibility index (EPI) which takes on values between 0 and 1 and is equal to 1 if the sender's message is equal to the data-optimal model. If the sender's message is equal to the model that is least likely given the data, the EPI will take on a value of 0. Values of the EPI that are strictly between 0 and 1 reflect cases of intermediate plausibility. We use the EPI as a measure of the distance in plausibility between the sender's message and the data-optimal model:

$$\text{EPI}(c^S, \theta^S_{pre}, \theta^S_{post}|h) := \frac{\mathcal{L}(c^S, \theta^S_{pre}, \theta^S_{post}|h)}{\mathcal{L}(c^{DO}, \theta^{DO}_{pre}, \theta^{DO}_{post}|h)}, \tag{1}$$

where $\mathcal{L}(\cdot|h)$ is the likelihood function conditional on the historical data h. In Appendix A.5.1, we provide further details on the construction of the EPI and discuss its relation to other benchmarks.

Measures of Persuasion: In our analysis, it will be of interest to have measures of the degree to which senders are able to persuade receivers. The measures discussed above already contribute to this by showing how far receivers' assessments are shifted away from the truth, or from the data. However, we also want to construct measures that indicate the degree to which receivers follow the message of the sender.[10] To do this, we construct the following measures:

(i) Distance between the sender's message and the data-optimal model:

$D^{DO}(\theta^S_{post}) := |\theta^S_{post} - \theta^{DO}_{post}|$

(ii) Distance between the sender's message and the receiver's assessment:

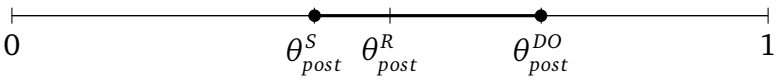
$D^S(\theta^R_{post}) := |\theta^R_{post} - \theta^S_{post}|$

[10]For a single receiver in a single round, one can think of this as an indication of the receiver's trust in the sender's report. When considering the average across all rounds for a single receiver, one can think of this as a measure of the receiver's gullibility.

(iii) The ratio of the distance that the receiver moves away from the data-optimal point to the distance that the sender *tries* to move the receiver from the data-optimal point:
$T := \frac{\theta^R_{post} - \theta^{DO}_{post}}{\theta^S_{post} - \theta^{DO}_{post}}$

A trust measure of $T = 1$ means the receiver is highly trusting of the sender; a trust value of $T = 0$ means that the receiver is maximally skeptical of the sender. Moreover, $T < 0$ and $T > 1$ suggest excessive skepticism and trust, respectively.[11] The following figure illustrates this measure:

Figure 2: Illustration of our measure of trust.



A.3.2 Hypotheses

Our main hypotheses are stated below. They concentrate on comparing receiver behaviour using two dimensions of exogenous variation: (i) different treatment conditions, and (ii) different sender types (i.e., senders with aligned or misaligned incentives).

When interpreting the hypotheses, and the associated empirical analysis plans, it is important to keep in mind that in three of our treatments (Baseline, Skepticism, and Sequential), we hold the instructions of the senders completely constant. Since senders also receive no feedback between rounds, this implies that sender behavior in these treatments should be approximately balanced on average, which allows a clean comparison of the receiver behavior in response to sender messages across these treatment conditions. In our fourth treatment, PrivateData, both the senders' and the receivers' instructions change in comparison to Baseline, since both learn that the sender will not observe the historical dataset prior to sending a message to the receiver. This implies that a treatment comparison between PrivateData and another treatment (e.g., Baseline) should be interpreted as a change in the equilibrium play of senders and receivers.

With regards to sender types, in the hypothesis section, we will often distinguish between receivers who face a sender with *aligned* versus *misaligned* incentives. A sender has aligned incentives if their payment is maximized when the receiver adopts the true θ_{post} of the data generating model. A sender who is incentivized to induce the receiver to report an estimate of θ_{post} that is shifted towards either 0 or 1 is misaligned. As mentioned above, we introduce exogenous variation in the sender incentives within each of our treatment conditions.

Following the section below in which we describe our main hypothesis, we also discuss a set of secondary hypotheses which focus more on within-treatment variation and sender behaviour.

[11] We would normally expect to see $T \in [0, 1]$ for each observation (i.e., that the receiver's report is between the data-optimal model and the sender's message). Therefore, checking for violations of this may be used as a form of rationality check on receiver behaviour.

Main Hypotheses

Influence of persuasion by senders (in BASELINE): We study the impact of sender incentives on receiver assessments by comparing the distance of the receiver's assessment to the true model, $D^T(\theta^R_{post})$, within the BASELINE treatment. Our first hypothesis is that receiver assessments are further from the truth when they face a sender with misaligned incentives. This provides a test of whether senders are able to persuade receivers to shift their beliefs, despite receivers observing objective data.

Hypothesis 1. *In BASELINE, the distance between the receiver's assessment and the truth is larger when sender incentives are misaligned than when sender incentives are aligned.*

We will test this hypothesis using the following regression model:

$$D^T(\theta^R_{post}) = \beta_0 + \beta_1 \times \mathbb{I}(\text{Misaligned sender}) + \rho_r + \varepsilon$$

and estimating it via OLS. In the equation above, $\mathbb{I}$(Misaligned sender) is an indicator function which takes a value of 1 if sender incentives are misaligned, ρ_r are round fixed effects and ε is an error term.[12] We will account for repeated observations and potential within matching group spillovers by clustering errors at the matching group level.[13] Using the estimates from this equation, we will test whether $\beta_1 > 0$. In addition, we will also present results from a Wilcoxon rank-sum test that tests whether the distributions of $D^T(\theta^R_{post})$ differ by alignment of the sender. When reporting these tests we will again account for repeated measurement and within matching group spillovers by reporting a test statistic for receiver outcomes which adjusts for clustered errors at the matching group level (see, e.g., Rosner, Glynn, & Lee, 2006).

Comparative statics using between-treatment variation

The following three hypotheses all involve exploiting the variation provided by our treatment conditions. We measure how persuasion changes in the various treatments relative to BASELINE using OLS regressions of the following kind:

$$D^T(\theta^R_{post}) = \beta_0 + \beta_1 \times \mathbb{I}(\text{Treatment}) + \rho_r + \varepsilon. \tag{2}$$

As our main persuasion measure to test our hypotheses, we take the distance between the truth and the receiver's assessment, $D^T(\theta^R_{post})$. To augment these results, we will also report the results of similar regressions which use the distance between the receiver's assessment and

[12] Since the true model is held constant within each round of the experiment, the ρ_r parameters absorb both round and true model fixed effects.

[13] It is worth noting that since senders receive no feedback at all during the experiment, the within matching group spillovers are more limited in scope than usual in experiments where subjects interact in groups. In our experiment, interaction between players only operates in one direction: from senders to receivers via the messages. Receivers also do not receive any feedback on the outcomes of their decisions prior to the end of the experiment.

the sender's message, $D^S(\theta^R_{post})$, as an alternative outcome measure.[14] The regressions will also typically include round fixed effects (ρ_r). In addition, we will also report nonparametric Wilcoxon rank-sum tests results for each hypothesis comparing the distribution of the outcome variable between two treatments. As before, both for the regression and the nonparametric test, standard errors will be clustered at the matching group level.

Influence of receiver skepticism: We study the impact of receiver skepticism by comparing the distance between the receiver's assessment and the truth when matched to a *misaligned* sender between Baseline and Skepticism. Since the sender's instructions are held constant between Baseline and Skepticism, the treatment comparison holds sender behaviour fixed and only potentially changes receiver behaviour. Our hypothesis is that, when moving from Baseline to Skepticism, this distance between the receivers assessment and the truth will decrease.[15]

Hypothesis 2. *When matched with a sender with misaligned incentives, the distance between the receiver's assessment and the truth is smaller in Skepticism than in Baseline.*

We test this hypothesis by estimating the regression specified in equation (2) for the Baseline and Skepticism treatments and testing whether $\beta_1 < 0$.

Influence of the receiver forming their own prior model: Here, we study the impact of encouraging the receiver to form their own personal interpretation of the data before they receive the assessment from the advisor. We do this by comparing the distance between the receiver's assessment and the truth when matched with a misaligned sender between the Baseline and Sequential treatments. Our hypothesis is that, when matched to a misaligned sender, receivers' assessments are closer to the truth in Sequential than in Baseline.

Hypothesis 3. *When matched with a sender with misaligned incentives, the distance between the receiver's assessment and the truth is smaller in Sequential than in Baseline.*

We test this hypothesis by estimating the regression specified in equation (2) for the Baseline and Sequential treatments and testing whether $\beta_1 < 0$.

Influence of receiver private data: We study the protective role of the receiver having access to private data by comparing the distance between the receiver's assessment and the truth

[14]It is important to note that the treatment comparisons involving the distance between the receiver's report and the sender's message, $D^S(\theta^R_{post})$, have a simple interpretation when comparing the three treatments in which the sender's information set is held identical (i.e., Baseline, Skepticism, and Sequential). However, treatment comparisons of this object involving the PrivateData treatment are more complicated to interpret, since both sender and receiver behavior may change. It is for this reason that we focus on $D^T(\theta^R_{post})$ as our primary object of interest. This has a clear interpretation across all four treatments.

[15]Our hypothesis here focuses only on misaligned senders. However, when considering aligned senders, it is possible that communication between an aligned sender and receiver improves when moving from Baseline to Skepticism as the receiver knows in Skepticism exactly when their matched sender is aligned. As a corollary to Hypothesis 2, we will also test for this possibility by comparing the distance between the receiver's assessment and the truth when matched to an aligned sender between Baseline and Skepticism.

when matched to a *misaligned* sender between BASELINE and PRIVATEDATA. We hypothesize that, when matched to a misaligned sender, receivers' assessments are closer to the truth in PRIVATEDATA than in BASELINE. Another interpretation of this hypothesis is that it is a test of whether senders who are able to construct an *ex post* narrative or model that is tailored to the exact historical data series that receivers observe are able to be more persuasive.

Hypothesis 4. *When matched with a sender with misaligned incentives, the distance between the receiver's assessment and the truth is smaller in PRIVATEDATA than in BASELINE.*

We test this hypothesis by estimating the regression specified in equation (2) for the BASELINE and PRIVATEDATA treatments and testing whether $\beta_1 < 0$.

Secondary Hypotheses Our secondary hypothesis are organized to test for certain empirical regularities using within-treatment variation. They mostly follow the theoretical predictions of the framework adapted from S&S to fit our experimental design. Appendix A.5.3 sets up the adapted framework, presents predictions, and discusses how they can be tested using data from the experiment. A reader interested in more detailed theoretical justifications of the following hypotheses can refer to this Appendix.

Since we use within-treatment variation for our secondary hypotheses, when studying receiver behaviour, we will predominantly focus on the BASELINE treatment in order to hold constant other contextual factors.[16] We will also focus on the subset of rounds where receivers are matched with a *misaligned* sender to study persuasion. For this reason, we will also collect a larger sample size in our BASELINE treatment.[17] When studying sender behavior, we are able to exploit the fact that the senders face an identical choice problem in the BASELINE, SKEPTICISM, and SEQUENTIAL treatments (i.e., senders in these three treatments receive identical instructions—they only differ in the receivers they are matched with, but are not aware of these differences and do not receive feedback from these receivers during the experiment). Therefore, we pool the senders from these three treatments for our within-treatment comparisons for senders.

[16] As a robustness check, we will also report the results for all receivers in the Appendices of the paper, including fixed effects to control for treatment differences, as well as fixed effects that account for potential interactions between the treatment and the alignment of the senders' incentives.

[17] A second reason for collecting a larger sample for our BASELINE treatment is that we use the BASELINE treatment as a comparison group in most of our main hypotheses, which makes it efficient to collect a larger sample for this treatment in comparison to the other treatments.

Secondary Hypotheses Regarding Receiver Behavior:

The influence of the empirical plausibility of narratives on receiver trust: This hypothesis addresses the question: are receivers more willing to follow a message that fits the data well?

We study the impact of the receiver receiving an empirically plausible message (i.e., a message that fits the observed historical data well) by relating the distance between the sender's message and the receiver's assessment, $D^S(\theta^R_{post})$, to the empirical fit of the sender's message, as measured by the Empirical Plausibility Index (EPI). We hypothesize that the better the sender's message fits the observed data, the smaller the distance between the sender's message and the receiver's assessment. Essentially, this says that receivers will be more willing to follow a sender's message if it fits the data they observe well.

Hypothesis 5a. *The distance between the sender's message and the receiver's assessment decreases in the EPI.*

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

$$D^S(\theta^R_{post}) = \beta_0 + \beta_1 \text{EPI}(c^S, \theta^S_{pre}, \theta^S_{post}|h) + \alpha + \rho_r + \varepsilon$$

and testing whether $\beta_1 < 0$. In the equation above, α denotes the estimated effect of being matched to an aligned sender. The round indicator variable, ρ_r, captures experimental round fixed effects. We will cluster standard errors at the matching group level.

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

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

local optima.[18] To adjust for possible changes in the sender's message quality across different histories, we condition the hypothesis on the value of the EPI evaluated at the sender's model.

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

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

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

and testing whether $\beta_2 > 0$, where the variable "EPI has multiple peaks" is a binary variable that takes a value of one when the EPI has more than one local maximum. Fixed effects and standard errors are calculated in the same way as in the specification for Hypothesis 5a.

Secondary Hypotheses Regarding Sender Behavior:

To conduct our within-treatment hypothesis tests pertaining to senders, we will pool sender data from all treatments where senders face an identical decision problems (i.e., the Baseline, Skepticism, and Sequential treatments).

The influence of incentives on sender behaviour: In a first comparison, we ask how senders react to different incentives. We will do this by comparing the distance between the sender's message and the truth, $D^T(\theta^S_{post})$, between aligned and misaligned senders. Our hypothesis is that the messages of misaligned senders are further from the truth.

Hypothesis 6a. *The distance between the sender's message and the truth of the post report, $D^T(\theta^S_{post})$, is larger for misaligned senders than for aligned senders.*

Constructing a convincing narrative: A related test of sender strategies concerns the narrative part of the sender's problem: A sender might adjust their choices of c and θ_{pre} to make their report of θ_{post} more convincing.[19] As we show in the Appendix, an upward incentive-biased sender should deviate from reporting the data-optimal year of change c^{DO} only if a different year increases the number of successes or decreases the number of failures in the post period. Conversely, a downward incentive-biased sender should deviate only if a different year decreases the number of successes or increases the number of failures in the post period. We

[18] Another way to think about this is that, if the log likelihood function of the model for a given history is relatively flat in θ_{post}, the sender is less swayed by the receiver's message, even if the communicated model has a high EPI because alternative models exist that also have a high EPI. We proxy flatness of the log likelihood function by distinguishing between flat (multiple peaked) and non-flat (single peaked) functions. Figure **??** in Appendix A.5.3 plots this function for an example history.

[19] This is despite the fact that the sender's incentives depend only on the receiver's θ_{post} report, implying that distortions of c and θ_{pre} serve a pure story-telling role.

hypothesize that this behaviour leads to a systematic bias in the choice of θ_{pre} away from the true model, which results in upward incentive-biased senders reporting a smaller value than the truth and downward incentive-biased senders reporting a larger value than the truth. In other words, the bias in the choice of θ_{pre} operates in the opposite direction to the choice of θ_{post} for misaligned senders.

Hypothesis 6b. *The distance between the sender's message and the truth of the pre report, $D^T(\theta^S_{pre})$, is larger for misaligned senders than for aligned senders.*

To test the previous two hypotheses, we specify and estimate regressions of the form

$$D^T(\theta^S) = \beta_0 + \beta_1 \mathbb{I}(\text{Misaligned sender}) + \rho_r + \varepsilon, \tag{3}$$

that either use $D^T(\theta^S_{post})$ (Hypothesis 6a) or $D^T(\theta^S_{pre})$ (Hypothesis 6b) as an outcome variable. We will test whether $\beta_1 > 0$. We will take account of repeated measurement by clustering standard errors at the sender level.[20]

Balancing persuasiveness against the truth (Aligned senders): Aligned senders face a tradeoff between sending a truthful message and sending a message that *more plausibly induces the truth*. Whether this tension induces the sender to bias their report of θ_{post} away from the data-optimal model may depend on the difference between θ^T_{post} and θ^{DO}_{post}. If this difference is positive (i.e., $\theta^T_{post} > \theta^{DO}_{post}$), an aligned sender has an incentive to bias their report upwards moving it closer to the truth, while they have an incentive to bias it downward towards the truth if the difference is negative (i.e., $\theta^T_{post} < \theta^{DO}_{post}$).[21] This leads us to the following hypothesis which asks whether aligned senders follow such a strategy that involves reporting a weighted average of the truth, θ^T_{post}, and the data-optimal parameter, θ^{DO}_{post}:

Hypothesis 7a. *The distance between the data-optimal model and the aligned sender's report, $D^{DO}(\theta^S_{post})$, increases in the distance between the truth and the data optimal report $|\theta^T_{post} - \theta^{DO}_{post}|$.*

We test this hypothesis by estimating the following model for senders from the pooled Baseline, Skepticism, and Sequential treatments:

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

and testing whether $\beta_2 > 0$.

[20] When reporting regressions on sender outcomes, we will take a less conservative clustering approach than when reporting on receiver outcomes, since senders do not receive feedback from other participants in their matching group.

[21] Note that this hypothesis could be formulated in two different, but equivalent, ways—either by considering that reports can be biased away from the data-optimal model or that they can be biased away from the true data generating model. Both are captured by the following intuition: We expect aligned senders' reports to reflect a compromise that biases their reports away from the data-optimal model and towards the true model (i.e., we expect that the average aligned sender will choose a report that represents some linear combination of the true θ^T_{post} and the data-optimal θ^{DO}_{post}).

Gravitational pull of the truth is weaker for misaligned senders: A final, related hypothesis is that misaligned senders should be less responsive to the true model than aligned senders. Essentially, the misaligned senders have incentives to persuade the receiver to move away from the truth, and they are constrained only by the receivers information set (i.e., the historical data, which yields the data-optimal model) and their own truth-telling preferences. If misaligned senders have no truth-telling preferences, they will completely disregard the truth and it will play no role in influencing their report. In this hypothesis we check whether misaligned senders: (a) are influenced by the truth, and (b) whether the size of this influence (pull towards the truth) is smaller than it is for aligned senders.

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

In the regression model specified above, we test whether: (i) $\beta_2 + \beta_3 > 0$, namely whether misaligned senders are responsive to the truth at all, and (ii) $\beta_3 < 0$, namely whether they are less responsive than aligned senders.

Tentative plans for additional exploratory analysis: In addition to the analysis specified above that is aimed at testing the hypotheses that we have outlined, we plan to also include some more exploratory analyses in the paper. We view it as being potentially useful to provide a description (snapshot) of our tentative plans regarding this exploratory analysis, although we note that this analysis is likely to change in the final version of the paper (in the paper, we will indicate which analyses were pre-registered and which are exploratory). Our tentative plans include the following: We plan to estimate regression models that explain the sender's report as a function of their type, the data optimal model, the period they report on, and the true model. We also plan to measure the percentage of sender messages that are consistent with utility maximization and investigate how messages deviate from the theoretical benchmark. Appendices A.5.2 and A.5.3 contain further details.

A.4 Sample Size

As most of our planned hypothesis tests either: (i) compare the Baseline treatment to one of our other treatment conditions, or (ii) compare participants within the Baseline treatment, we will collect more observations for Baseline than for the other treatments. In particular, we plan to collect data from 360 participants (180 senders and 180 receivers) in Baseline and from 180 participants in each of the remaining treatments. The sample size gives us 80% power to detect a minimum treatment effect of 2.3 when considering the distance between the receiver's assessment and the truth at the 5%-level.[22] We based the power analysis on data we collected in a pilot of the Baseline treatment where we found that the distance between

[22]In the power analysis, we randomly draw observations from the pilot data and simulate regression results that include round fixed-effects and which cluster standard errors at the matching group level.

the receiver's assessment and the truth, our main outcome variable of interest, had a mean of 17.467 and a standard deviation of 14.094.

Therefore, in total, we will collect approximately 900 observations in these four treatment conditions: 360 in BASELINE, 180 in SKEPTICISM, 180 in SEQUENTIAL and 180 in PRIVATEDATA. Within each treatment, half will be senders and half receivers. Amongst senders, one-third will be randomly assigned to each incentive condition.

A.5 Appendix to the Preregistration 1 Document

A.5.1 Construction of the Empirical Plausibility Index

In this section, we show how we determine the model that is most likely to have generated a history of outcomes. Possible models consist of parameter combinations $(c, \theta_{pre}, \theta_{post})$. The data set consists of a vector $h = (\omega_1, \omega_2, ..., \omega_{10})$, where $\omega_t \in \{0, 1\}$. An $\omega_t = 1$ denotes "success" and an $\omega_t = 0$ denotes "failure". For each possible parameter combination and data set, we can calculate the empirical likelihood as follows:

$$\begin{aligned}\mathcal{L}(c, \theta_{pre}, \theta_{post}|h) &= \prod_{t=1}^{c}(\theta^S_{pre})^{\omega_t}(1-\theta^S_{pre})^{1-\omega_t} \times \prod_{t=c+1}^{10}(\theta^S_{post})^{\omega_t}(1-\theta^S_{post})^{1-\omega_t}, \\ &= (\theta^S_{pre})^{\omega_1+...+\omega_c}(1-\theta^S_{pre})^{c-(\omega_1+...+\omega_c)} \times (\theta^S_{post})^{\omega_{c+1}+...+\omega_{10}}(1-\theta^S_{post})^{10-c-(\omega_c+...+\omega_{10})}, \\ &= (\theta^S_{pre})^{k_{pre}}(1-\theta^S_{pre})^{c-k_{pre}} \times (\theta^S_{post})^{k_{post}}(1-\theta^S_{post})^{10-c-k_{post}}.\end{aligned} \tag{4}$$

In the equation above, $k_{pre} \equiv \sum_{t=1}^{c} \omega_t$ denotes the number of successes before the structural break and $k_{post} \equiv \sum_{t=c+1}^{10} \omega_t$ denotes the number of successes after the structural break. We further know that, fixing c, the maximum likelihood estimator of θ_{pre} and θ_{post} is equal to $\theta^{DO}_{pre}(c) = k_{pre}/c$ and $\theta^{DO}_{post}(c) = k_{post}/(10-c)$. Therefore, the optimal year of change c^{DO} for a given data set h is equal to $\arg\max_{c\in\{2,3,...,8\}} \mathcal{L}(c, \theta^{DO}_{pre}(c), \theta^{DO}_{post}(c)|h)$.

We evaluate the empirical plausibility index of sender's messages (EPI) for a given data set by comparing the empirical likelihood of the sender's model to the that of the model that is most likely to have generated the data as follows:

$$\text{EPI}(c^S, \theta^S_{pre}, \theta^S_{post}|h) := \frac{\mathcal{L}(c^S, \theta^S_{pre}, \theta^S_{post}|h)}{\max_c \mathcal{L}(c, \theta^{DO}_{pre}(c), \theta^{DO}_{post}(c)|h)}.$$

Since, for any data set, there always exists a model which induces a minimized likelihood value of zero,[23] the empirical plausibility index is scaled to take on values between zero and one. An empirical plausibility index of one suggests that the sender sent the model that is most likely

[23]For a given c, if either $k_{pre} < c$ or $k_{post} < 10-c$, setting $\theta_{pre} = \theta_{post} = 1$ will result in a likelihood value of zero. If the history only consists of successes, so that $k_{pre} = c$ and $k_{post} = 10-c$, setting $\theta_{pre} = \theta_{post} = 0$ will result in a likelihood value of zero. A model which induces a zero likelihood value thus always exists. Since the likelihood function can never take on negative values, we conclude that its minimum value is zero.

to have generated the data set, while a value of zero suggests that the sender sent the model which is least likely to have generated the data set.

Relation to Schwartzstein & Sunderam (2021) S&S conceptualize an agent who will be persuaded by a model whenever that model provides a better empirical fit of the data than an initial default model held by the agent. The empirical fit is thereby measured by the likelihood conditional on the data and the agent's prior beliefs. Whenever some model induces a higher EPI than another model, an agent in S&S would prefer the first model. To show this equivalence more precisely, we derive the posterior distribution over $(c, \theta_{pre}, \theta_{post})$ that a Bayesian agent with prior belief $\psi(c, \theta_{pre}, \theta_{post})$ would hold after observing h. Denote this posterior distribution by $f(c, \theta_{pre}, \theta_{post}|h, \psi)$. Using Bayes' rule,

$$f(c, \theta_{pre}, \theta_{post}|h, \psi) = \frac{f(h|c, \theta_{pre}, \theta_{post})\psi(c, \theta_{pre}, \theta_{post})}{\sum_{x=2}^{8} \int_{y\in[0,1]} \int_{z\in[0,1]} f(x, y, z|h)\psi(x, y, z)\,\mathrm{d}y\mathrm{d}z}.$$

Now, $\psi(c, \theta_{pre}, \theta_{post})$ is constant for all potential messages, since we specified a data generating process where all parameters are uniformly distributed and independent of one another. Further, the denominator in the equation above is constant over all potential messages. It follows that the joint distribution is directly proportional to $f(h|c, \theta_{pre}, \theta_{post})$, which is equal to the likelihood function in (4). As a consequence, any message which maximizes the likelihood function also maximizes the joint distribution of parameters. Therefore, a message that suggests a model with EPI $= 1$ would always (weakly) persuade an agent regardless of the default model in S&S. More generally, if for any two models $(c', \theta'_{pre}, \theta'_{post})$ and $(c'', \theta''_{pre}, \theta''_{post})$ EPI$(c', \theta'_{pre}, \theta'_{post}) >$ EPI$(c'', \theta''_{pre}, \theta''_{post})$, an agent in S&S would judge the former model more plausible.

Comparison to the beta-binomial updating formula A different popular belief benchmark in the literature is to compare stated beliefs about certain parameters to their objective Bayesian expected value. We will consider the case where an agent forms a degenerate belief about c and subsequently arrives at non-degenerate beliefs for θ_{pre} and θ_{post} using Bayesian updating. Before seeing any data, agents hold a uniform prior over θ_{pre} and θ_{post}. A uniform distribution on $[0, 1]$ can be represented by a beta distribution with parameters $\alpha = 1$ and $\beta = 1$. The mean of a beta distribution is given by

$$\frac{\alpha}{\alpha + \beta}.$$

Upon seeing n realizations of the state (success/failure), out of which k are successes and l are failures, agents update the parameters of the beta distribution to $\tilde{\alpha} = \alpha + k$ and $\tilde{\beta} = \beta + l$. The posterior *mean* belief is thus equal to

$$\frac{\tilde{\alpha}}{\tilde{\alpha} + \tilde{\beta}} = \frac{\alpha + k}{\alpha + \beta + n}.$$

This is also known as the beta-binomial updating formula. However, the posterior *mode* of a beta distribution is given by

$$\min\left\{\frac{\tilde{\alpha}-1}{\tilde{\alpha}+\tilde{\beta}-2},1\right\} \text{ if } \tilde{\alpha}>1,\tilde{\beta}\geq 1.$$

Consider the following example of a Bayesian agent who observes $h' = (1,1,1,1,0,0,1,0,1,0)$ and believes that $c = 4$. If non-degenerate, their posterior belief over θ_{pre} is distributed according to a beta distribution with $\alpha = 5$ and $\beta = 1$. Their *mean* belief of θ_{pre} is thus equal to $5/6$. In contrast, an agent in S&S would find an estimate of θ_{pre} more persuasive that maximizes the likelihood function. This estimate is equal to the empirical frequency of successes in periods 1-4; $4/4$. Similarly, the expected value of θ_{post} according the beta-binomial updating formula is $3/8$, whereas the maximum likelihood estimate of θ_{post} for $c = 4$ is $2/6$. These considerations imply that $\text{EPI}(4, 4/4, 2/6|h') > \text{EPI}(4, 5/6, 3/8|h')$. It is straightforward to verify that the maximum likelihood estimates coincide with the mode of the updated beta distribution. Therefore, it is best to think of our EPI measure as quantifying the plausibility of a model under the assumption that agents evaluate the model's likelihood against the historical data and accept the model whenever the likelihood is sufficiently high. When they accept the model, they form a degenerate belief about $(c, \theta_{pre}, \theta_{post})$, which is equal to the parameters of the accepted model.

A.5.2 Analysis of Senders' Messages

To gain a more fine-grained insight into sender strategies, we will specify and estimate regression models that explain a sender's report of θ_{pre} and θ_{post} as a function of the sender's type, the empirical data, the period they report on, and the true model. We offer two approaches to specifying such models.

Parametric approach We take the *pre* and *post* report of a sender as the outcome variable (θ_t^S) and specify the regression model

$$\begin{aligned}\theta_t^S =& \beta_0 + \beta_1 \mathbb{I}(\text{t} = \text{post}) + \mathbb{I}(\text{type} = \text{upward}) \times [\beta_2 + \beta_3 \mathbb{I}(\text{t} = \text{post})] \\ &+ \mathbb{I}(\text{type} = \text{downward}) \times \left[\beta_4 + \beta_5 \mathbb{I}(\text{t} = \text{post})\right] \\ &+ \delta_1 \theta_t^T + \delta_2 \theta_t^{DO} + \rho_r + \varepsilon.\end{aligned}$$

We call this the "parametric approach" since we explicitly include θ_t^T and θ_t^{DO} as benchmark controls in the regression. Therefore, estimated effects of the sender type and on the reporting period are relative to these benchmarks. In the later presented semiparametric approach, we instead measure differences in reporting relative to the empirically observed average report. Let us highlight the interpretation of a number of coefficients and their expected signs:

- β_2 and β_3 capture deviations in the reporting behavior of an upward biased sender relative to the behavior of an aligned sender, separately for the *pre* and for the *post* period. We expect $\beta_2 < 0$ and $\beta_3 > 0$ (see hypotheses 6a and 6b).

- β_4 and β_5 capture deviations in the reporting behavior of an downward biased sender relative to the behavior of an aligned sender, separately for the *pre* and for the *post* period. We expect $\beta_4 > 0$ and $\beta_5 < 0$ (see hypotheses 6a and 6b).

To examine aligned senders, notice that they essentially have the same incentives to bias their model away from the data-optimal model as the upward biased sender if $\theta_t^T - \theta_t^{DO} > 0$ and an incentive as the downward biased sender if $\theta_t^T - \theta_t^{DO} < 0$. We introduce the terms

$$\begin{aligned}
&(\theta_t^T - \theta_t^{DO})\big[\beta_6 + \beta_7 \mathbb{I}(\text{t = post})\big] \\
+\mathbb{I}(\text{type = upward})&(\theta_t^T - \theta_t^{DO}) \times \big[\beta_8 + \beta_9 \mathbb{I}(\text{t = post})\big] \\
+\mathbb{I}(\text{type =downward})&(\theta_t^T - \theta_t^{DO}) \times [\beta_{10} + \beta_{11} \mathbb{I}(\text{t = post})]
\end{aligned}$$

in the regression above. Here we accordingly expect that $\beta_6 < 0$ and $\beta_7 > 0$.

Semiparametric approach The experiment provides a high degree of variation in the histories that sender-receiver pairs observe. With the semiparametric approach, we will use this feature of the experimental design to maximize what we can learn from the data. The method we will use consists of "mirroring" histories, as described in the following. One can construct a mirror image of any history of past outcomes h that reverses the timing of success and failure. For example, the history $h = (1,0,1,1,1,0,0,0,0,1)$ has a mirror image history $h' = (1,0,0,0,0,1,1,1,0,1)$. More formally, h' is a mirror image of h if $\omega_t = \omega'_{10-(t-1)}$ for all $t \in \{1,\ldots,10\}$, $\omega_t \in h$ and $\omega'_t \in h'$. Observe that h' is a mirror of h if and only if h is a mirror of h'. We will refer to any two histories (h,h') where h' is a mirror of h as a "mirror pair".[24]

This part of our analysis consists of identifying mirror pairs for which the set of all senders collectively report at least two models (one for each history of the pair) in the experimental data. We then compare the θ^S_{pre} from one history of the pair to the θ^S_{post} from the other history of the pair. This comparison allows us to cleanly identify the directions into which senders bias their reports. A sender who always reports the true data generating model should on average report the same θ_{pre} for history h as θ_{post} for history h' and the same θ_{pre} for history h' as θ_{post} for history h. On the contrary, a sender who exaggerates the post period success probability should report a θ_{post} for history h that is larger than the θ_{pre} for history h', etc. Table 9 presents example of experimental data and the comparisons we will make. To facilitate the analysis we will typically transform the data from a wide format as displayed in table 9 to a long format as displayed in table 10. The long format has only one column for the sender report θ^S which can either denote a report for the *pre* or *post* period probability of success. It doubles the size of the experimental data, as we have two reports (one pre and one post) for each sender and round. We will specify the same models as in the parametric approach but, instead of controlling for $\delta_2 \theta_t^{DO}$ we will include mirror pair dummies $\pi_{p,comp}$. The mirror pair fixed effect indicator, $\pi_{p,comp}$, differs by two variables, the pair id p and a binary indicator $comp \in \{0,1\}$ which varies within mirror pairs. The reason is that not every parameter between two histories

[24]Note that the definition implies that (h,h) is a mirror pair if h is symmetric.

Table 9: Example of experimental data and planned comparisons

θ^S_{pre}	θ^S_{post}	history	sender type	mirror pair id
0.3a	b0.6	h	upward biased	12
0.5c	d0.45	h'	upward biased	12
0.7e	f0.3	h"	aligned	31
0.3g	h 0.7	h"'	aligned	31

Table 10: The long version of Table 9

θ^S	period	history	sender type	mirror pair id	$\mathbb{I}$(reference history)	$\mathbb{I}$(comparable to reference *pre* report)
0.3	pre	h	upward biased	12	1	1
0.6	post	h	upward biased	12	1	0
0.5	pre	h'	upward biased	12	0	0
0.45	post	h'	upward biased	12	0	1
0.7	pre	h"	aligned	31	1	1
0.3	post	h"	aligned	31	1	0
0.3	pre	h"'	aligned	31	0	0
0.7	post	h"'	aligned	31	0	1

is comparable. Instead, we can only compare the *pre* report of a history to the *post* report of the mirror history. For that reason we define for each mirror pair a *reference history* that we use to construct the indicator in the fixed effect to absorb differences in average reporting between the two possible θ reports for each history. For example, in table 10, h is the reference history for mirror pair 12. The pre report of the reference history is comparable to the post report of the nonreference history. Therefore, the indicator $\mathbb{I}$(comparable to reference *pre* report) is 1 in rows 1 and 4 of the table and 0 in rows 2 and 3. Similar comparisons apply to mirror pair 31, where h'' is the reference history.

In comparison to the parametric approach, in the nonparametric approach we do not assume that senders know the data optimal model and choose their message accordingly. Instead, we only assume that, absent any incentives to bias the report away from the data-optimal message and concerns for truth-telling, senders will send a message after seeing history h that mirrors their message after seeing message h' if (h, h') are a mirror pair.

A.5.3 Theoretical Framework

This section sketches a framework that guides our secondary hypotheses. The framework largely follows S&S but is in some ways adjusted to our setting.

Consider a sender whose goal is to persuade a receiver of a certain model. The sender and receiver observe a history of outcomes h. The history records for each of ten years t the success ($\omega_t = 1$) or failure ($\omega_t = 0$) of a company, which is generated by a true data generating model m^T. A model consists of a year of change $c \in \{2, 3, \ldots, 8\}$, a pre-change success probability $\theta_{pre} \in [0, 1]$ and a post-change success probability $\theta_{post} \in [0, 1]$. The

company's outcome in each year up to the year of change is drawn from a binomial distribution with success probability θ_{pre}. In years $t > c$, the company's outcome is drawn from a binomial distribution with success probability θ_{post}. We will use "pre period" to describe the range of years up to the year of change and "post period" to describe the range of years after the year of change. The timing of the game is as follows:

(i) Nature draws three parameters $(c^T, \theta^T_{pre}, \theta^T_{post})$ that form the true data generating model. Each of the parameters is drawn from a uniform distribution and is uncorrelated with the other parameters.

(ii) The true data generating model generates a history h.

(iii) The receiver observes h and draws a default model m^D from a distribution function $M(c, \theta_{pre}, \theta_{post}|h)$.

(iv) The sender observes h and sends a model $m^S = (c^S, \theta^S_{pre}, \theta^S_{post})$ to the receiver.

(v) The receiver decides whether to adopt the sender's model. In case the receiver accepts, they make a report θ^S_{post}. Otherwise, they report the value θ^D_{post} of the default model.

(vi) Sender payoffs realize.

Following S&S, we consider the receiver to be a nonstrategic agent who decides as if their objective is to adopt the most compelling model. Models can be evaluated by their fit, which we take to be equal to the value of the log likelihood function evaluated at the model parameters.[25] For a history that is generated as described above, the log likelihood function is

$$ll(c, \theta_{pre}, \theta_{post}) = k_{pre}(c)log(\theta_{pre}) + f_{pre}(c)log(1-\theta_{pre}) + k_{post}(c)log(\theta_{post}) + f_{post}(c)log(1-\theta_{post}).$$[26]

In the equation, $k_{pre}(c) = \sum_{t=1}^{c} \omega_t$ denotes the number of successes and $f_{pre}(c) = c - k_{pre}(c)$ denotes the number of failures in the pre-period. The values $k_{post}(c) = \sum_{t=c+1}^{10} \omega_t$ and $f_{post}(c) = 10 - c - k_{post}(c)$ similarly denote the number of successes and failures in the post period. For a given year of change c, there is always one pair $(\theta_{pre}, \theta_{post})$ that maximizes the log likelihood function. We denote these likelihood maximizers by $\hat{\theta}_{pre}(c)$ and $\hat{\theta}_{post}(c)$. Closed-form solution exist. In period p, the likelihood maximizer given c is equal to the number of successes divided by the total length of the period; $\hat{\theta}_p(c) = k_p(c)/(k_p(c) + f_p(c))$. The following discussion assumes that the log likelihood function has a unique optimum.[27] We call the model that maximizes the log likelihood function the data-optimal model and denote it by $m^{DO} = (c^{DO}, \hat{\theta}^{DO}_{pre}, \hat{\theta}^{DO}_{post})$. Most of the time, the data-optimal model will be different from the true data-generating model.

Receiver types The receiver's type depends on the drawn default model.[28] The distribution

[25] As discussed in section A.5.1, this is how a Bayesian agent would choose among models in our setting.

[26] Here and in the following, we usually do not condition functions on a particular history h to save notation.

[27] This holds for almost all possible histories. Degenerate histories like $(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$, for which any year can be part of a data optimal model, are the exceptions.

[28] This is a major difference between our framework and S&S, who assume only one type of receiver.

of default models implies a distribution of log likelihood function values with c.d.f. $G(\ell)$ and p.d.f. $g(\ell)$. For simplicity, we assume that g has full support over all possible values of the likelihood function, i.e., $g(\ell) > 0$ for all $\ell \in (-\infty, ll(m^{DO})]$. The default model is private information to the receiver, though the sender knows that its log-likelihood value is distributed according to $G(\ell)$.[29]

Sender types The sender can either be aligned, upward biased or downward biased. The receiver's report will determine the sender's payoff in different ways, depending on the sender's type. In particular, the receiver's report θ^R_{post} maps into the sender's payoff according to a scoring rule

$$1-(\varphi-\theta^R_{post})^2.$$

This rule assigns the sender the maximum score whenever the receiver reports sender's target φ. If the sender is aligned φ is equal to θ^T_{post}, if the sender is upward biased φ is equal to 1, and if the sender is downward biased φ is equal to 0. Since the receiver adopts the sender's model if it provides at least the same fit as the default model, the sender's expected utility from sending a model m^S is

$$\begin{aligned} u(c^S,\theta^S_{pre},\theta^S_{post};h,\varphi) = {} & \mathbb{P}(ll(c^S,\theta^S_{pre},\theta^S_{post})\geq\ell)[1-(\varphi-\theta^S_{post})^2] \\ & + \mathbb{P}(ll(c^S,\theta^S_{pre},\theta^S_{post})<\ell)\mathbb{E}[1-(\varphi-\theta^D_{post})^2|ll(c^S,\theta^S_{pre},\theta^S_{post})<\ell]. \end{aligned}$$

In the equation above, $\mathbb{E}[1-(\varphi-\theta^D_{post})^2|ll(c^S,\theta^S_{pre},\theta^S_{post})<\ell]$ is the sender's expected payoff when the receiver does not adopt the sender's model. We make the simplifying assumption that the sender believes this expectation term to be equal to a value $x \in (0,1)$, which is independent of the sender's message.[30] Plugging in $G(\ell)$ and the sender expectation, the sender's expected utility function is equal to

$$u(c^S,\theta^S_{pre},\theta^S_{post};h,\varphi) = G(ll(c^S,\theta^S_{pre},\theta^S_{post}))[1-(\varphi-\theta^S_{post})^2]+(1-G(ll(c^S,\theta^S_{pre},\theta^S_{post})))x.$$

The maximization problem can then be written as

$$\max_{c,\theta_{pre},\theta_{post}} G(ll(c,\theta_{pre},\theta_{post}))(1-x-(\varphi-\theta_{post})^2).$$

Analysis We analyse sender behaviour. In the first part of the analysis we focus on a misaligned, i.e., upward or downward biased sender. We extend the results to the aligned sender at the end of the section. Throughout the analysis, we will often benchmark sender strategies by comparing them to the strategy that communicates the data-optimal model m^{DO}. Since

[29] As will become clear later, this assumption could be relaxed by quite a bit without qualitatively changing the results. What is important is that the sender knows that (i) the receiver holds a default model that might not be equal to the data-optimal model and that (ii) the receiver adopts the sender's model whenever the sender's model has a log likelihood value at least equal to the receiver's default model.

[30] This is a simplifying assumption as, in principle, knowing that the receiver does not adopt a certain model might be informative about the value of θ^D_{post}. While we are aware of this possibility, we regard it as second-order. The assumption above awards us with tractability and allows us to focus on the direct effects of the sender's report.

$G(ll(c^{DO}, \theta^{DO}_{pre}, \theta^{DO}_{post})) = 1$, the receiver always adopts the data-optimal model upon reception. In the analysis below, we somewhat informally assume that x, the sender's payoff when the receiver does not adopt the sender's model, is close to zero. We do not believe that this is a meaningful restriction: For any x, all the results below would go through under the qualifier that the sender only chooses among models which induce a scoring rule payoff $1-(\varphi-\theta^{R}_{post})^2 \geq x$.

The sender faces a conflict between an *accuracy motive* which induces them to communicate a model with a high fit that is likely adopted by the receiver, and a *direction motive* to convince the receiver to report a particular value of θ_{post}. We start with a result that naturally follows from the accuracy-direction tradeoff. The sender only communicates a non data-optimal θ_{post} if it increases the direction motive.

Observation 1. *Consider the choice of the optimal θ^{S}_{post}:*

(i) An upward biased sender chooses a $\theta^{S}_{post} \geq \theta^{DO}_{post}$.

(ii) A downward biased sender chooses $\theta^{S}_{post} \leq \theta^{DO}_{post}$.

Proof. Consider case (i). The data-optimal model dominates the choice of any model $(c', \theta'_{pre}, \theta'_{post})$ with $\theta'_{post} < \theta^{DO}_{post}$ because any such alternative model decreases accuracy and direction motives. Any model $(c'', \theta''_{pre}, \theta''_{post})$ with $\theta''_{post} > \theta^{DO}_{post}$ instead decreases the accuracy motive but (weakly) increases the direction motive. The claim follows. A symmetric argument can be made for case (ii). □

The direction motive only applies to θ_{post} but not to θ_{pre}. It follows that any sender, regardless of type, will always communicate the likelihood maximizer of θ_{pre} conditional on the year of change.

Observation 2. *For any type of sender who chooses any year of change c^S, θ^{S}_{pre} is equal to $\hat{\theta}_{pre}(c^S)$.*

Proof. The value of θ_{pre} affects the expected utility function only through the effect it has on $ll(\cdot)$ (the accuracy motive). It follows that choosing the value θ_{post} which maximizes the log likelihood is optimal. □

We now turn to the choice of the optimal cutoff c^S. Consider a sender who considers to communicate the data-optimal model with year of change c^{DO}. It turns out that the sender only considers alternative years of change if they can better rationalize a θ_{post} in line with the direction motive.

Observation 3a. *Consider how an upward biased sender chooses the optimal c^S:*

(i) For years $c' > c^{DO}$ the sender prefers a model with year c' over a model with year c^{DO} if and only if:

- *$f_{post}(c') < f_{post}(c^{DO})$ and*
- *$\theta^{S}_{post} > \tilde{\theta}_2(c')$, where $\tilde{\theta}_2(c')$ is a critical value on $(\hat{\theta}_{post}(c^{DO}), 1)$.*

(ii) For years $c' < c^{DO}$ the sender prefers a model with year c' over a model with year c^{DO} if and only if:

- $k_{post}(c') > k_{post}(c^{DO})$, $\hat{\theta}_{post}(c') > \hat{\theta}_{post}(c^{DO})$, *and*
- $\theta^S_{post} \in (\tilde{\theta}^L_2(c'), \tilde{\theta}^H_2(c'))$, *where* $\tilde{\theta}^L_2(c') > \theta^{DO}_{post}$ *and* $\tilde{\theta}^L_2(c') \leq \tilde{\theta}^H_2(c') \leq 1$ *are two critical values.*

Proof. Let us denote the empirical successes and failures implied by the data-optimal model by k^{DO}_j and f^{DO}_j (for $j \in \{pre, post\}$). We compare the data-optimal model to a model $m' = (c', \theta'_{pre}, \theta'_{post})$ with $c' \neq c^{DO}$ and implied empirical successes and failures k'_j and f'_j. Since m^{DO} maximizes the log likelihood function, it follows that $ll(c^{DO}, \theta^{DO}_{pre}, \theta^{DO}_{post}) > ll(c', \theta'_{pre}, \theta'_{post})$ for any m'. Therefore, any model with cutoff c' can only lead to an increase in the sender's expected utility if it increases the direction motive. For an upward biased sender, m' induces a higher direction motive if $\theta'_{post} > \theta^{DO}_{post}$. We show conditions under which the sender prefers to communicate a model $(c', \theta'_{pre}, \theta'_{post})$ over $(c^{DO}, \theta^{DO}_{pre}, \theta'_{post})$. As both models have the same direction motive, the sender prefers the first to the second model only if the log likelihood difference $ll(c', \theta'_{pre}, \theta'_{post}) - ll(c^{DO}, \theta^{DO}_{pre}, \theta'_{post}) > 0$ is positive. This difference has a number of important properties. First, consider the value of the difference when evaluated at θ^{DO}_{post}. Since m^{DO} maximizes the log likelihood function, it follows that $ll(c', \theta'_{pre}, \theta^{DO}_{post}) - ll(c^{DO}, \theta^{DO}_{pre}, \theta^{DO}_{post}) < 0$. Second, the sign of the derivative of the difference with respect to θ'_{post} when evaluated at θ^{DO}_{post} depends on likelihood maximizer of θ_{post} under the alternative model, $\hat{\theta}_{post}(c')$, as follows:

$$\frac{\partial ll(c', \theta'_{pre}, \theta^{DO}_{post})}{\partial \theta_{post}} - \frac{\partial ll(c^{DO}, \theta^{DO}_{pre}, \theta^{DO}_{post})}{\partial \theta_{post}} \begin{cases} \leq 0 & \text{if } \hat{\theta}_{post}(c') \leq \theta^{DO}_{post} \\ > 0 & \text{if } \hat{\theta}_{post}(c') > \theta^{DO}_{post}. \end{cases}$$

In both cases, the derivative of the log likelihood evaluated at the data-optimal model is zero, since it is evaluated at the optimum. The sign of the difference is then fully determined by the sign of the log likelihood derivative evaluated at the alternative model. It is negative if $\hat{\theta}_{post}(c') < \theta^{DO}_{post}$ (the log likelihood is past its peak) and positive if $\hat{\theta}_{post}(c') > \theta^{DO}_{post}$ (the peak is still to come). Another important property of the log likelihood functions is that they cross at most once for values of $\theta_{post} \in [0, 1]$. We show this by taking the derivative of the log likelihood difference with respect to θ_{post};

$$\frac{\partial ll(c', \theta'_{pre}, \theta_{post})}{\partial \theta_{post}} - \frac{\partial ll(c^{DO}, \theta^{DO}_{pre}, \theta_{post})}{\partial \theta_{post}} = \frac{k'_{post} - k^{DO}_{post}}{\theta_{post}} + \frac{f^{DO}_{post} - f'_{post}}{1 - \theta_{post}}.$$

Note that, if they are nonzero, the two terms on the right hand side always have the opposite sign because either $k^{DO}_{post} \geq k'_{post}$ and $f^{DO}_{post} \geq k'_{post}$ or $k^{DO}_{post} \leq k'_{post}$ and $f^{DO}_{post} \leq k'_{post}$ (in both cases, at least one inequality is strict). Setting the derivative equal to zero and rearranging, we find

that

$$\frac{\theta^0_{post}}{1-\theta^0_{post}} = \frac{k^{DO}_{post} - k'_{post}}{f^{DO}_{post} - f'_{post}},$$

which implies a unique θ^0_{post} as a solution. This value is equal to

$$\theta^0_{post} = \frac{k^{DO}_{post} - k'_{post}}{k^{DO}_{post} - k'_{post} + f^{DO}_{post} - f'_{post}}.$$

The log likelihood difference can thus either increase, decrease, first increase and then decrease or first decrease and then increase for $\theta_{post} \in [0,1]$. We now distinguish between a number of cases that determine the shape of the log likelihood difference.

Case 1: $k'_{post} = k^{DO}_{post}$. The critical value θ^0_{post} is equal to zero. It directly follows that the likelihood difference is monotone. If $\hat{\theta}_{post}(c') > \theta^{DO}_{post}$ ($f'_{post} < f^{DO}_{post}$) it increases, if $\hat{\theta}_{post}(c') < \theta^{DO}_{post}$ ($f'_{post} > f^{DO}_{post}$) it decreases.

Case 2: $k'_{post} > k^{DO}_{post}$ *and* $f'_{post} > f^{DO}_{post}$. The derivative of the log likelihood difference changes its sign at θ^0_2. We ask whether $\theta^0_2 \gtrless \theta^{DO}_{post}$. Plugging in values, this is equivalent to showing whether

$$\frac{k^{DO}_{post} - k'_{post}}{k^{DO}_{post} - k'_{post} + f^{DO}_{post} - f'_{post}} \gtrless \frac{k^{DO}_2}{k^{DO}_2 + f^{DO}_2}.$$

After rearranging, we find that

$$\theta^0_2 > \theta^{DO}_{post} \text{ if } \hat{\theta}_{post}(c') > \theta^{DO}_{post} \text{ and } \theta^0_2 \leq \theta^{DO}_{post} \text{ if } \hat{\theta}_{post}(c') \leq \theta^{DO}_{post}.$$

Since we know the sign of the derivative at θ^{DO}_{post}, this pins down the whole shape of the derivative; it first increases and then decreases.

The additional cases $f'_{post} = f^{DO}_{post}$ and $k'_{post} < k^{DO}_{post}$ and $f'_{post} < f^{DO}_{post}$ follow in a similar way. We summarize the results in the table below.

Table 11: Shape of the log likelihood difference for different parameter combinations

	$k'_{post} = k^{DO}_{post}$	$k'_{post} > k^{DO}_{post}$ and $f'_{post} > f^{DO}_{post}$	$f'_{post} = f^{DO}_{post}$	$k'_{post} < k^{DO}_{post}$ and $f'_{post} < f^{DO}_{post}$
$\hat{\theta}_{post}(c') > \theta^{DO}_{post}$	Increasing	First increasing, then decreasing Peak at $\theta^0_{post} > \theta^{DO}_{post}$	Increasing	First decreasing, then increasing Minimum at $\theta^0_{post} < \theta^{DO}_{post}$
$\hat{\theta}_{post}(c') \leq \theta^{DO}_{post}$	Decreasing	First increasing, then decreasing Peak at $\theta^0_{post} \leq \theta^{DO}_{post}$	Decreasing	First decreasing, then increasing Minimum at $\theta^0_{post} \geq \theta^{DO}_{post}$

As a final property of the log likelihood difference, when taking the limit of $\theta_{post} \to 1$ we find that

$$\lim_{\theta'_{post} \to 1} \Big[ll(c', \theta'_{pre}, \theta'_{post}) - ll(c^{DO}, \theta^{DO}_{pre}, \theta'_{post}) \Big] = \lim_{\theta'_{post} \to 1} log(1-\theta'_{post})(f'_{post} - f^{DO}_{post}) + \kappa, \qquad (5)$$

where κ is a number independent of θ_{post}. Since $\lim_{\theta_{post} \to 1} log(1-\theta_{post}) = -\infty$, the difference is positive in the limit if $f^{DO}_{post} > f'_{post}$ and negative if $f^{DO}_{post} < f'_{post}$.

This discussion has a number of implications for sender strategies. Consider the first row of table 11 where $\hat{\theta}_{post}(c') > \theta^{DO}_{post}$.

- If $k'_{post} = k^{DO}_{post}$ it must be that $f'_{post} < f^{DO}_{post}$. Since the difference is positive in the limit as θ'_{post} becomes large, there is one value $\tilde{\theta}_2 \in (\theta^{DO}_{post}, 1)$ so that a model that couples $\theta^S_{post} > \tilde{\theta}_{post}$ with c' has a larger likelihood than a model with c^{DO}.
- If $k'_{post} > k^{DO}_{post}$ and $f'_{post} > f^{DO}_{post}$ there might be a range of values between $(\theta^{DO}_{post}, 1)$ for which a model that couples a θ^S_{post} in that range with c' has a larger likelihood than a model with c^{DO}.
- If $f'_{post} = f^{DO}_{post}$ the difference is increasing. From equation (5), a value $\tilde{\theta}_{post} \in (\theta^{DO}_{post}, 1)$ under which a model with c' and $\theta^S_{post} > \tilde{\theta}_2$ has a larger likelihood only exists if $\kappa > 0$.
- If $k'_{post} < k^{DO}_{post}$ and $f'_{post} < f^{DO}_{post}$ there is one value $\tilde{\theta}_{post} \in (\theta^{DO}_{post}, 1)$ so that a model that couples $\theta^S_{post} > \tilde{\theta}_{post}$ with c' has a larger likelihood than a model with c^{DO}.

Consider the second row of table 11 where $\hat{\theta}_{post}(c') \leq \theta^{DO}_{post}$.

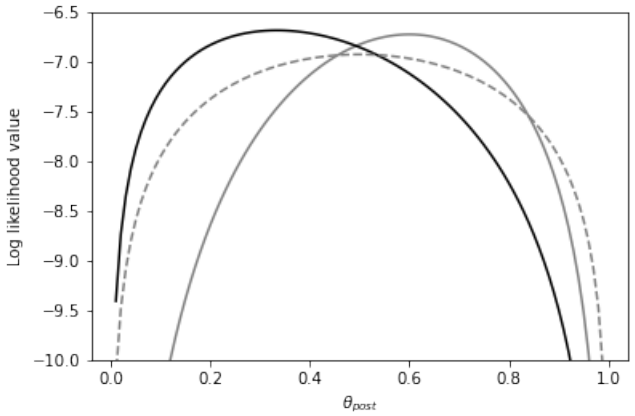
- If $k'_{post} = k^{DO}_{post}$ or $f'_{post} = f^{DO}_{post}$ an upward biased sender would never choose the model c' since its likelihood is lower than that of the data-optimal model for all values $\theta^S_{post} \geq \theta^{DO}_{post}$.
- If $k'_{post} > k^{DO}_{post}$ and $f'_{post} > f^{DO}_{post}$ then the difference starts decreasing before θ^{DO}_{post}. Since it is negative at θ^{DO}_{post}, there is no $\theta^S_{post} \geq \theta^{DO}_{post}$ where the alternative model has a higher likelihood.
- If $k'_{post} < k^{DO}_{post}$ and $f'_{post} < f^{DO}_{post}$ there is one value $\tilde{\theta}_2 \in (\theta^{DO}_{post}, 1)$ so that a model that couples $\theta^S_{post} > \tilde{\theta}_{post}$ with c' has a larger likelihood than a model with c^{DO}.

Finally, note that $c' < c^{DO}$ if and only if $k'_{post} \geq k^{DO}_{post}$ and $f'_{post} \geq f^{DO}_{post}$ and that $c' > c^{DO}$ if and only if $k'_{post} \leq k^{DO}_{post}$ and $f'_{post} \leq f^{DO}_{post}$. The above considerations imply the claims in the observation. □

This result puts restrictions on the years of change an upward biased sender is willing to communicate. In words, the observation says that a sender will only choose a later year if the later year implies fewer failures in the post period. Conversely, the sender will only choose an earlier year if the earlier year implies more successes in the post period. Perhaps surprisingly, the sender is slightly more constrained in choosing an earlier than a later year. The reason for this asymmetry seems to be the following: As θ_{post} becomes very large, the log likelihood function puts a strong penalty on any failure in the second period so that this term dominates the function value (intuitively, with a high θ_{post} failures are difficult to explain). Therefore, a later year under which fewer failures happen in the post period becomes more attractive. On the converse, an earlier year does not lower the number of failures, which makes it unattractive for high values of θ_{post}.

Since, for a fixed θ_{post}, the direction motive is held constant for any c, the sender prefers the year of change which maximizes the log likelihood function. The figure below plots log likelihood functions for different values of c and for an example history $h = (0, 1, 1, 0, 0, 1, 1, 0, 1, 0)$. The black line displays the log likelihood function with year of change 7. The figure shows that, for intermediate values of θ_{post}, this cutoff is dominated by a model with $c = 5$ (the gray line) which adds two additional successes to the post period. For very high values of θ_{post} both models are dominated by a model with a later year of change of $c = 8$, whose log likelihood is displayed by the dashed line. This later year of change minimizes the number of failures in the second period.

Figure 3: Log likelihood functions for different c



Note: The graph plots values of three log likelihood functions for different values of θ_{post} and for history $h = (0, 1, 1, 0, 0, 1, 1, 0, 1, 0)$. The black line plots the log likelihood of model $(7, \hat{\theta}_{pre}(7), \theta_{post})$, the grey line of model $(5, \hat{\theta}_{pre}(5), \theta_{post})$, and the dashed line of model $(8, \hat{\theta}_{pre}(8), \theta_{post})$.

We obtain a similar result for the downward biased sender.

Observation 3b. *Consider how a downward biased sender chooses the optimal c^S:*

(i) For years $c' > c^{DO}$ the sender prefers a model with year c' over a model with year c^{DO} if and only if:

- $k_{post}(c') < k_{post}(c^{DO})$ *and*
- $\theta^S_{post} > \tilde{\theta}_2(c')$*, where* $\tilde{\theta}_2(c')$ *is a critical value on* $(\hat{\theta}_{post}(c^{DO}), 1)$.

(ii) For years $c' < c^{DO}$ the sender prefers a model with year c' over a model with year c^{DO} if and only if:

- $k_{post}(c') < k_{post}(c^{DO})$*,* $\hat{\theta}_{post}(c') < \hat{\theta}_{post}(c^{DO})$*, and*
- $\theta^S_{post} \in (\tilde{\theta}^L_2(c'), \tilde{\theta}^H_2(c'))$*, where* $\tilde{\theta}^L_2(c') > \theta^{DO}_{post}$ *and* $\tilde{\theta}^L_2(c') \leq \tilde{\theta}^H_2(c') \leq 1$ *are two critical values.*

Having gained insight into the choice of c^S, we close with an observation on the sender's optimal model

Observation 4. *Consider the sender's choice of the optimal model* $(c^S, \theta^S_{pre}, \theta^S_{post})$. *Denote by* c^{max} *the year for which* $\hat{\theta}_{post}(c^{max}) = max\left\{\hat{\theta}_{post}(c)\right\}_{c\in\{2,\dots,8\}}$ *and by* c^{min} *the year for which* $\hat{\theta}_{post}(c^{min}) = min\left\{\hat{\theta}_{post}(c)\right\}_{c\in\{2,\dots,8\}}$.

(i) The upward biased sender chooses a model for which either $\theta^S_{post} > \hat{\theta}_{post}(c^{max})$ *or* $ll(c^S, \hat{\theta}_{pre}(c^S), \theta^S_{post}) \geq ll(c^{max}, \hat{\theta}_{pre}(c^{max}), \hat{\theta}_{post}(c^{max}))$ *holds.*

(ii) The downward biased sender chooses a model for which either $\theta^S_{post} < \hat{\theta}_{post}(c^{min})$ *or* $ll(c^S, \hat{\theta}_{pre}(c^S), \theta^S_{post}) \geq ll(c^{min}, \hat{\theta}_{pre}(c^{min}), \hat{\theta}_{post}(c^{min}))$ *holds.*

Proof. Consider case (i). Suppose by contradiction that none of the conditions hold. Then the sender could increase the accuracy and the direction motive by transmitting model $(c^{max}, \hat{\theta}_{pre}(c^{max}), \hat{\theta}_{post}(c^{max}))$ instead of model $(c^S, \theta^S_{pre}, \theta^S_{post})$, a contradiction. Starting from a model with $\theta^S_{post} > \hat{\theta}_{post}(c^{max})$, the accuracy motive decreases when moving to model $(c^{max}, \hat{\theta}_{pre}(c^{max}), \hat{\theta}_{post}(c^{max}))$. Starting from a model with $ll(c^S, \hat{\theta}_{pre}(c^S), \theta^S_{post}) \geq ll(c^{max}, \hat{\theta}_{pre}(c^{max}), \hat{\theta}_{post}(c^{max}))$, the direction motive decreases. Therefore, at least one but not both conditions must hold. A symmetric argument can be made to show case (ii). □

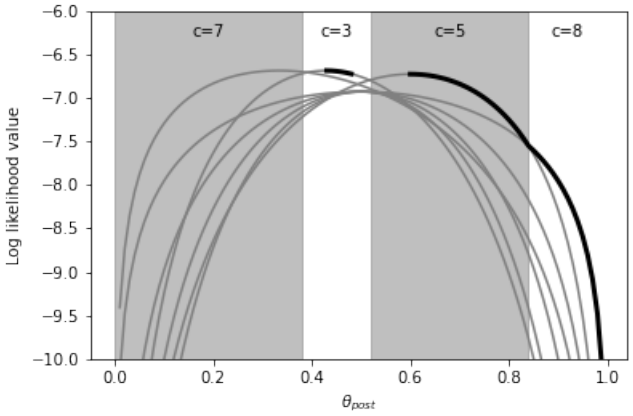
Figure 4 plots log likelihood functions of an example history for all possible years of change. It illustrates the upward biased sender's problem to pick among combinations of c and θ_{post}. The black line displays combinations which are consistent with Observation 4. This line has gaps, as some combinations are dominated by other combinations. For example, the data-optimal model in the example has year $c^{DO} = 3$, which makes any θ_{post} to the left of the peak of its likelihood function suboptimal. The c^{max} in this example is equal to 5, which is why the black line continues without gaps vor values of θ_{post} larger than $\hat{\theta}_{post}(5)$.

Aligned sender The direction motive of the aligned sender depends on the true data generating model. For example, if $\theta^T_{post} < \theta^{DO}_{post}$, the aligned sender has an incentive to communicate a θ^S_{post} smaller than the data-optimal value. Whether the aligned sender biases reports upward or downward depends on whether the difference $\theta^T_{post} - \theta^{DO}_{post}$ is smaller or larger than zero.[31] Therefore, the same qualitative theoretical results as for the upward biased sender also hold for the aligned sender when $\theta^T_{post} > \theta^{DO}_{post}$. When instead $\theta^T_{post} < \theta^{DO}_{post}$, the predictions for the aligned sender follow those of the downward biased sender. We however note that the misaligned senders represent extreme cases. Therefore, the predictions for the aligned sender would be quantitatively smaller.

Observation 5a. *If* $\theta^T_{post} > \theta^{DO}_{post}$, *part* (i) *of Observation 1 and Observation 3a also apply to the aligned sender. If* $\theta^T_{post} < \theta^{DO}_{post}$, *part* (ii) *of Observation 1 and Observation 3b also apply to the aligned sender.*

[31] This discussion largely ignores the case where $\theta^T_{post} = \theta^{DO}_{post}$, which is unlikely to ever be exactly true. We note that in this unlikely case, the sender only has an accuracy motive, i.e., the sender will communicate the data-optimal model.

Figure 4: Combinations of c and θ_{post} consistent with utility maximization, upward biased sender



Note: The graph plots values of log likelihood functions for all possible years of change, different values of θ_{post}, and for history $h = (0, 1, 1, 0, 0, 1, 1, 0, 1, 0)$. The years at the top of the figure highlight years that are optimal for values of θ_{post} within the shaded area. The black line highlights values of θ_{post} that are consistent with utility maximization.

Observation 5b. *Consider the aligned sender's choice of the optimal model* $(c^S, \theta^S_{pre}, \theta^S_{post})$*. Denote by* c^{max} *the year for which* $\hat{\theta}_{post}(c^{max}) = min\left\{|\theta^T_{post} - \hat{\theta}_{post}(c)|\right\}_{c\in\{2,\dots,8\}}$.

(i) The aligned sender chooses a model for which either $|\theta^T_{post} - \hat{\theta}_{post}(c)| > |\theta^T_{post} - \hat{\theta}_{post}(c^{max})|$ *or* $ll(c^S, \hat{\theta}_{pre}(c^S), \theta^S_{post}) \geq ll(c^{max}, \hat{\theta}_{pre}(c^{max}), \hat{\theta}_{post}(c^{max}))$ *holds.*

A.5.4 Implications for the Empirical Analysis

The observations above provide benchmarks for sender behavior. In particular, we can measure the percentage of biased sender messages that are consistent with parts (i) and (ii) of Observation 4 and, using Observation 5b, we can do a similar exercise for aligned senders.

B Original Preregistration 2 Document

B.1 Experimental Design

The basic design of these treatments will follow the BASELINE and INVESTORPRIOR treatments of the experiment reported in the January 2023 CESifo working paper entitled "Narrative Persuasion".[32] These treatments are intended to extend and complement the evidence reported in that working paper.

B.1.1 Investor-only Treatments

Building on the INVESTORPRIOR treatment, the two new investor-only treatments will only elicit behavior from participants in the role of an investor (we will not collect new advisor data for these two treatments). The primary objective of these treatments is to evaluate the role played by justifications when constructing a convincing narrative. Essentially, we turn on or off the presence of a justification (in the form of the justification parameters, θ_{pre} and c) in support of the advisor's assessment of the main variable of interest (θ_{post}). We describe the design of these two treatments, 3PARAMETERS and 1PARAMETER, below.

3Parameters: In this treatment, investors are told that they will face a decision that is very similar to the one that investors faced in a previous experiment (the previous INVESTORPRIOR treatment). They then read the instructions that were shown to the investors in INVESTORPRIOR. They are told that they have been linked to an investor from this previous experiment and that this implies that, in each round of the experiment, they will be shown the same company data that was shown to their linked investor. They will also receive the same advice that the linked investor received in each round of the experiment. This advice comes in the form of the three narrative parameters, $(\theta^A_{pre}, \theta^A_{post}, c^A)$, that advisors chose in the previous INVESTORPRIOR treatment. Each round of the experiment proceeds in the same way as in the INVESTORPRIOR treatment: Investors first see only the company data and make an initial assessment about the underlying process that they think best explains the data (i.e., about $\theta_{pre}, \theta_{post}$ and c). They then receive the advisor's message and make a final assessment about θ_{post}. The procedures we take in this experiment ensure that all information provided to participants is truthful. In particular, we make sure that participants will indeed observe the same sequence of company data and advice as their linked investor.

1Parameter: This treatment closely follows the 3PARAMETERS treatment. The sole difference is that in 1PARAMETER investors only receive the previous advisor's θ^A_{post}-assessment. That is, they are not shown θ^A_{pre} and c^A.

Linking participants to investors from the past experiment We will randomly match every participant in the new experiment to an investor who participated in the INVESTORPRIOR

[32] Available here: https://www.cesifo.org/node/73857

treatment of the past experiment. The new participant will then face the same decisions as their linked investor. That is, in every round of the experiment, participants will see the same company data and receive the same θ^A_{post} parameter as advice as their linked investor. The two treatments will vary whether the two auxiliary parameters (θ^A_{pre} and c^A) are shown to the participant or not.

Specifically, we will link each past investor of the InvestorPrior treatment to:

- Two new investors in the 3Parameters treatment.
- Two new investors in the 1Parameter treatment.

B.1.2 Investor-Advisor Treatments

Symmetric This treatment follows the design of the Baseline treatment closely, with the main difference being that advisors in this treatment do not know the true data generating process (and investors know that advisors do not know). This implies that advisors have no additional information relative to investors—their information sets are identical. The implication of this is that in a fully rational model there should be no information transmission because the advisor has no additional information. A second difference is that, when constructing the message, the computer interface will show advisors a graph being drawn on the historical company data that visualizes their message. Investors will receive the company data with the advisor's message drawn onto it. The objective of this treatment is to examine whether advisors can shape how investors interpret objective information even when advisors have no additional private information. We will be able to compare advisors within this treatment who face different incentives (and investors who are matched with advisors with different incentives). Therefore, the key source of variation that we exploit here is the variation in advisor incentives *within* the Symmetric treatment (as opposed to comparing behavior in Symmetric to behavior in another treatment condition).

In this Symmetric treatment, participants take part in the experiment in a matching group of 3 investors and 3 advisors. Importantly, we will balance the company data such that in every round of the experiment all members of a matching group will see the same data. The reason for this is that it will ensure that we are able to compare advisor-investor pairs where the data is held completely constant, but the incentives of the advisor are varied.

CompetingNarratives After participants complete ten rounds of the Symmetric treatment, as a surprise, they will participate in the CompetingNarratives treatment for five rounds. The key difference from the Symmetric treatment is that, in CompetingNarratives, when advisors are constructing their message they see a competing message constructed by a robot advisor.

The investor then observes the messages of both the robot advisor and the human advisor without knowing which message was constructed by whom (the advisor knows that the investor

will observe both messages). After observing the two messages, the investor then makes an assessment. In contrast to SYMMETRIC, this assessment is a binary choice, i.e., the investors must form an assessment of which of the two messages contains a θ_{post} that is more likely to be true. (Investors are incentivized through a binarized scoring rule to choose the message with the more accurate θ_{post}-value.) Advisors face the same incentives as they did during the ten rounds of the SYMMETRIC treatment, i.e., they are incentivized to try to induce the investor to adopt a message that includes a θ_{post}-value that is either as close as possible to one (up-advisor), as close as possible to zero (down-advisor), or accurate (aligned advisor).

In COMPETINGNARRATIVES, participants are told that the robot advisor is programmed to try to help the investor (i.e., the robot's objective is to help the investor to be accurate) but that there is variability in how skilled the robot advisor is in constructing messages: In practice, the robot advisor's message will be constructed in the following way:

- In Round 1, the robot advisor sends the true data generating process $(\theta_{pre}^T, \theta_{post}^T, c^T)$.
- In rounds 2-5, the robot advisor always sends the true θ_{post}. However, the other two parameters are determined in one of two ways—either they are chosen randomly, or they are chosen to be close to the data-optimal values. Specifically, they are chosen in the following way:
  - (Data-optimal supporting parameters) In half of the rounds, the robot advisor will calculate the data-optimal c and θ_{pre} terms, conditional on the true θ_{post} and the data. We will slightly perturbe θ_{pre} by adding a noise term, $\eta \sim U[-.03, .03]$. The rationale for adding noise is to avoid having the data-optimal values that are exactly equal to the fraction of successes in the pre-period, making the robot advisor easy to detect.
  - (Random supporting parameters) In the other half of the rounds, the robot advisor's c is randomly drawn from $U\{2, 8\}$ and θ_{pre} is drawn from $U[0, 1]$.

  Whenever the θ_{pre} parameters generated by the procedures described above takes on an extreme value of either 0 or 1, we will replace it by a randomly generated, less extreme value. In particular, we will replace a value of 0 with a value that is drawn from $U[0, .09]$ and we will replace a value of 1 with a value that is drawn from $U[.9, .99]$. We do this to avoid having substantial density at 0 an 1.

  All θ values take on a value between 0 and 1 and are rounded to two digits after the decimal place.

Therefore, we will generate exogenous variation in the fit (EPI) of the robot advisor's message in rounds 2 - 5. In terms of the implementation, we want to be able to compare two investor-advisor pairs who observe the same data, where the advisor has the same incentives, but where one advisor is competing with a robot advisor who produces a good justification for θ_{post} in terms of fit (i.e., by sending the *data-optimal* supporting parameters) and the other

advisor produces a less good justification (i.e., by sending *randomly chosen* supporting parameters). Therefore, in each of rounds 2-5, we will always have two matching groups who all see the same data and where the robot advisor sends the same θ_{post}. The difference between the two groups will be that in one group the robot advisor will send the data-optimal auxilliary parameters while in the other group the robot advisor will send randomly chosen auxiliary parameters. Within each of these two groups, the full message sent by the robot advisor is held constant (i.e., the c^R and θ^R_{pre} parameters sent by the robot are varied between the two groups and held constant within).

B.2 Hypotheses and Analysis

B.2.1 Investor-only Treatments

In the treatment comparisons, we will take advantage of the fact that we link two investors in each of the two treatments (3Parameters and 1Parameter) to an investor from the InvestorPrior treatment. This implies that these four investors in the different treatments see exactly the same data and all observe a message from the same advisor. We will, therefore, often control for shared-linked-investor×round fixed effects. These fixed effects hold the observed company history and the θ^A_{post} seen by investors constant. This allows us to isolate the effect of the variation in the auxiliary parameters induced by the treatments on our outcomes of interest. We will denote these fixed effects in the regressions below by λ.

When running regressions, we will cluster the standard errors at the investor level. We denote the individual-specific error term in the regressions by ε.

Main Hypothesis

Hypothesis 8. *The advisor's message will influence the investor's final assessment more in 3Parameters than in 1Parameter.*

We will investigate this hypothesis by estimating the following regression model:

$$\Delta D^I = \beta \times \mathbb{I}(\text{treatment} = \text{3Parameters}) + \lambda + \varepsilon.$$

where the dependent variable, ΔD^I, is defined as the change in the absolute distance between the investor's assessment of θ_{post} and the advisor's θ^A_{post} between $t = 0$ and $t = 1$ (i.e., the change that can be attributed to the advisor's message).[33] We ask whether this change is larger in the 3Parameters treatment in comparison to the 1Parameter treatment by testing

[33] Specifically, we define $\Delta D^I := D^{I,1}(\theta^A_{post}) - D^{I,0}(\theta^A_{post})$ where $D^{I,1}(\theta^A_{post}) := |\theta^{I,1}_{post} - \theta^A_{post}|$ and $D^{I,0}(\theta^A_{post}) := |\theta^{I,0}_{post} - \theta^A_{post}|$. Essentially, $D^{I,0}(\theta^A_{post})$ denotes the absolute distance between the investor's initial assessment at $t = 0$ (before meeting the advisor) and the advisor's assessment of θ_{post}. Then, $D^{I,1}(\theta^A_{post})$ is the absolute distance between the investor's assessment and the advisor's assessment at $t = 1$ after the investor has received a message from the advisor. Therefore, ΔD^I reflects the change in the distance between the investor's assessment and the advisor's assessment due to the investor receiving a message from the advisor.

whether $\beta < 0$.Essentially, we are asking whether persuasion is more pronounced in the 3Parameters treatment in comparison to the 1Parameter treatment by asking whether there is more movement towards the advisor's message in 3Parameters than in 1Parameter.

Heterogeneity:

Since persuasion in the 3Parameters treatment is expected to be most effective when the empirical fit (EPI) of the message is high, we will also test for heterogeneity in the effect of the treatment using the following regression.[34]

$$\Delta D^I = \beta_1 \cdot \mathbb{I}(\text{treatment} = \text{3Parameters}) + \beta_2 \cdot \mathbb{I}(\text{treatment} = \text{3Parameters}) \times \text{EPI} + \lambda + \varepsilon.$$

where EPI is a continuous measure of the empirical fit of the advisor's message—the EPI of the best-fitting message for a given historical dataset is 1, while the EPI of the worst-fitting message is 0; messages with intermediate levels of fit are associated with intermediate values of the EPI. Essentially, the regression allows us to assess whether the empirical fit interacts with the treatment. We will test whether $\beta_3 < 0$, which would indicate that for messages where the empirical fit is high, the auxiliary narrative components are important for persuasion.

B.2.2 Investor-Advisor Treatments

When analyzing data from Symmetric, we will often control for Round×History fixed effects.[35] We will denote these fixed effects by γ.

When analyzing data from CompetingNarratives, we will often control for Round×History×θ^R_{post} fixed effects. In the notation, R superscripts denote advice given by the robot advisor (e.g., m^R is the message sent by the robot advisor and θ^R_{post} is the θ_{post}-component of the robot's message). We will denote these fixed effects by η.

When running regressions, we will cluster the standard errors on the matching group level when analyzing investor outcomes. We will cluster the standard errors at the advisor level when analyzing advisor outcomes. We denote the individual-specific error term by ε.

Main Hypotheses

Hypothesis 9. *In Symmetric, up-advisors persuade investors to increase their assessment and down-advisors persuade investors to decrease their assessment, relative to aligned advisors.*

We will test this hypothesis by estimating the regression equation

$$\theta^{I,1}_{post} = \beta_0 \cdot \mathbb{I}(\text{up-advisor}) + \beta_1 \cdot \mathbb{I}(\text{down-advisor}) + \gamma + \varepsilon.$$

[34]To obtain a measure of message fit, we use the empirical plausibility index (EPI). We defined this EPI measure, and provided a detailed discussion of it, in the earlier pre-registration (AEARCTR-0009103) that described the treatments discussed in the January 2023 CESifo working paper. Essentially, the EPI takes on values between 0 and 1, such that the best-fitting message (for the relevant historical data) takes a value of 1 and the worst-fitting message takes a value of 0. For further details, please refer to pages 8-9 and 19-20 of preregistration: AEARCTR-0009103.

[35]For clarity, by "History" we are referring to the company history—i.e., the observed company data. Within a round, we therefore have a fixed effect for investors who observe the same historical company data.

Here, $\mathbb{I}$(up-advisor) is an indicator variable for whether the investor received advice from the up-advisor (and vice versa for $\mathbb{I}$(down-advisor)). We expect that $\beta_0 > 0$ and $\beta_1 < 0$.

Hypothesis 10. *In CompetingNarratives, the investor is more likely to adopt the human advisor's narrative if the robot advisor picks the auxiliary parameters randomly.*

We will test this hypothesis by estimating the regression equation

$$\mathbb{I}(\text{adopt } m^A) = \beta \cdot \mathbb{I}(\text{robot advisor sends random auxiliary parameters}) + \eta + \varepsilon.$$

In the equation above, $\mathbb{I}$(robot advisor sends random auxiliary parameters) is an indicator variable equal to one if the robot advisor's auxiliary parameters were randomly drawn from uniform distribution. The control group in this case are robot advisors who send data-optimal auxiliary parameters. The dependent variable, $\mathbb{I}$(adopt m^A), is a binary variable that takes a value of 1 when the investor adopts the message sent by the human advisor. We expect that $\beta > 0$. This would mean that the investor adopts the human advisor's messages more often when the robot advisor is selecting the auxiliary narrative justification parameters randomly rather than optimally (i.e., when they provide a less good justification in terms of fit). When running these regressions, we will only include data from rounds 2-5.

B.2.3 Additional Analyses

For Round 1 of CompetingNarratives, we will calculate how frequently the investor adopts the advisor's narrative instead of the true data generating model sent by the robot advisor. This will allow us to assess whether human advisors are able to construct narratives that are more persuasive than the truth.

We will also investigate how human advisors react to different strategies of robot advisors. Here, we expect that the fit of the human advisor's message, EPI(m^A), will be higher in cases where the robot advisor chooses the auxiliary parameters in a data-optimal way instead of randomly. We also expect that up-advisors will send a θ^A_{post} further away from their target of 1 and down-advisors a θ^A_{post} further away from their target of 0 if the robot advisor chooses the auxiliary parameters in a data-optimal way instead of randomly. The rationale for this is that when competing with a poor-fitting message, there is more scope for trying to move the investor's belief further while still sending a narrative the fits better than the competing message.[36]

B.3 Sample Size

We will have a sample of 180 investors in each of the 1Parameter and 3Parameters treatments. In a pilot, we found that ΔD^I, our main outcome variable in this treatment, took on

[36]Essentially, assuming that there is a tradeoff between movement of θ^A_{post} and the fit of the narrative, then when competing with a message that fits very well, one needs to also send a narrative that fits well to "beat" it on fit. This allows less flexibility for trying to shift the investor's θ_{post} further.

a mean value of -11.921 (s.d. 14.868). Based on a power analysis and given the sample size of 180 in each treatment, we will have 83% power to detect a minimum treatment effect of -1.846 (the effect found in the pilot) at the 5-% significance level.

In the SYMMETRIC / COMPETINGNARRATIVES, we will also collect 360 observations—here, this will constitute 180 investors and 180 advisors. In our pilot, we found that $\theta_{post}^{I,1}$ takes on a mean value of 50.671 (s.d. 31.801) and that $\mathbb{I}$(adopt m^A) takes on a mean value of 0.467 (s.d. 501). For the analysis planned for Hyp. 2 and given the sample size, we are 99% powered to to detect a β_0 of at least 4.7 and a β_1 of -8.8 or less (i.e., larger in magnitude, but negative) at the 5-% significance level. The assumed sizes of these coefficients are based on pilot results. For the analysis planned for Hyp. 3, our pilot results suggest a β-estimate of .20. Given the sample size, we have 99% power to detect such an effect at the 5-% significance level.