Discussion Paper Series – CRC TR 224

Discussion Paper No. 757
Project A 01

Learning From False Stories

Robin Musolff[1]
Christopher Roth[2]
Florian Zimmermann[3]

June 2026

[1]University of Cologne, Email: robin.musolff@gmail.com.
[2]University of Cologne, CESifo, ECONtribute, NHH Norwegian School of Economics, CEPR and MPI for Behavioral Economics, Email: roth@wiso.uni-koeln.de.
[3]University of Bonn, CESifo and MPI for Behavioral Economics, Email: florian.zimmermann@uni-bonn.de.

Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Learning From False Stories

Robin Musolff Christopher Roth Florian Zimmermann*

May 28, 2026

Abstract

False information often shapes beliefs even after retraction or correction. Using incentivized online experiments, we document a qualitative residue in learning from false information: False quantitative signals generate little to no belief updating, whereas false stories lead to substantial residual belief impact. This effect is robust across a range of design variants. Replacing evaluative stories with more neutral variants eliminates the residue, indicating that the valence of the qualitative information plays a central role. We provide direct evidence that false stories increase mental simulation, measured via self-reports and valence extracted from speech recordings. Respondents are also more confident in beliefs formed after false stories than after false quantitative information, despite being further from the rational benchmark. Additional experiments suggest that the effect also appears among respondents who report not being influenced by the story, consistent with stories partly shaping beliefs below conscious awareness.

JEL codes: D83, D91, C90

Keywords: Stories, Learning, Mental Simulation, Fake News.

*Musolff: University of Cologne, robin.musolff@gmail.com. Roth: University of Cologne, CESifo, ECONtribute, NHH Norwegian School of Economics, CEPR and MPI for Behavioral Economics, roth@wiso.uni-koeln.de. Zimmermann: University of Bonn, CESifo and MPI for Behavioral Economics, florian.zimmermann@uni-bonn.de. Funding from the Deutsche Forschungsgemeinschaft (DFG) through CRC TR 224 (Project A01) and under Germany's Excellence Strategy EXC 2126/2-390838866 is acknowledged. This project received funding from the European Research Council under Horizon 2020 (ERC Starting Grant 101160770—VIRAL), from the Hans-Kelsen Preis of the University of Cologne, and from the Research Council of Norway (FAIR project No 262675). We thank Pedro Bordalo, Ben Enke, Luca Henkel, Nicola Gennaioli, Frederik Schwerter, Michael Thaler and participants at the workshop on frontiers in behavioral research at Bocconi for extremely constructive comments. We particularly thank Thomas Graeber for extremely valuable input into the early stages of this project. Milena Jessen, Philipp Schirmer and Till Stange provided excellent research assistance.

1 Introduction

False information often circulates widely before it is corrected. News articles are retracted after publication, viral social media posts are debunked hours after spreading, and confident accounts of companies, candidates, or events turn out to be wrong. By the time a correction arrives, the original claim has typically reached a large audience and shaped their beliefs. Yet, such corrections may often be insufficient to erase the imprint that the original claim has already left on beliefs. Quantifying this residue, and understanding what governs its magnitude, is central to evaluating the costs of misinformation and to the design of effective corrective interventions.

A natural conjecture is that the form in which false information arrives matters for how much of it survives correction. False claims commonly take one of two forms: stories—qualitative descriptions of specific events—and quantitative information—in our setting, a single binary recommendation from an analyst. The two formats differ along dimensions that may shape belief impact. Discredited quantitative information conveys a number that might be easy to set aside. A discredited story, by contrast, is laden with vivid qualitative detail that may shape what comes to mind when people later think about the topic, engaging mental simulation in ways that persist after the underlying claim has been retracted. Consider the case of AI-generated deepfakes: synthetic video can now place fabricated words in the mouths of real people, packaging fabricated stories in the most vivid medium available.

In this paper, we take up the question of how the form of false information shapes its residual influence on beliefs through a series of controlled experiments. We study how false stories and false quantitative information differentially affect beliefs, and investigate the mechanisms driving this effect. We identify a qualitative residue: exposure to false quantitative signals produces negligible belief movement, whereas exposure to false stories causes substantial belief impact. This effect is robust across experimental design variants. Guided by a simple framework, we show that the residue disappears when evaluative stories are replaced with neutral ones, and that false stories increase mental simulation, evidenced by self-reports and speech analysis. False stories further engender greater confidence in the resulting beliefs, even though those beliefs deviate more substantially from rational Bayesian benchmarks. Finally, the effect persists among respondents who explicitly deny having been influenced, suggesting that learning from stories can operate below the threshold of conscious awareness.

Design. We design controlled experiments to isolate the qualitative residue of false information, defined as the extent to which false stories shift beliefs more than false quantitative signals after debunking. Doing so poses several design challenges. First, because stories are qualitative, it is difficult to specify a Bayesian benchmark against which residual belief movement can be measured. Second, the falsification of the signal must be easy to understand and salient at the point of belief elicitation, so that any residue cannot be attributed to subjects simply missing the

correction. Third, stories and quantitative information must be made sufficiently comparable that differences in updating can be attributed to specific story features.

Our design aims to address all challenges. In the experiment, respondents evaluate hypothetical companies starting from a uniform prior over the number of positive evaluations. Each respondent observes a single analyst recommendation, positive or negative, that is equally likely to be relevant or irrelevant. We then vary the type of information within subjects. In the quantitative-information-only condition, respondents see the recommendation alone for a given company. In the story condition, the same recommendation is accompanied by a qualitative elaboration in the same direction. By fixing the quantitative component, this comparison allows us to isolate the effect of the qualitative features of the story. Between subjects, we additionally vary whether the information is conveyed through text or video. Respondents report incentivized beliefs twice: once under genuine uncertainty about the signal's relevance, and again after explicitly and saliently learning its relevance or irrelevance with certainty.[1] Our key outcome is beliefs after respondents learn that the signal was irrelevant. At that point, the entire signal is nullified: the analyst recommendation alone and story alike carry zero information about company quality. Hence the rational benchmark is exactly zero belief impact: any residual belief movement in either condition is a departure from Bayesian updating. The difference between conditions identifies what qualitative details contained in the story add to belief impact, even when the entire signal has been rendered irrelevant.

Main Results. We document three main findings. First, false quantitative information barely move beliefs. While in our baseline experiment quantitative signals have a belief impact of around 1 percentage point, there is close to zero impact across our robustness experiments. Second, false stories shift beliefs substantially more compared to false signals that are purely quantitative. After respondents learn that their signal was irrelevant, residual belief movement in the story condition is 4.2 percentage points, meaning that the qualitative residue is around 3 percentage points. While magnitudes might appear modest at first, it is important to note that the benchmark is zero: any residual movement after a definitively irrelevant signal constitutes a stark departure from rational updating. Third, stories conveyed by video generate somewhat larger effects than text—approximately 5.2 versus 3.3 percentage points for stories, and 1.7 versus 0.9 for quantitative signals. Even though this corresponds to a close to 46% larger qualitative residue (3.5 pp vs 2.4 pp), this difference is not statistically significant at conventional levels.

Robustness. The qualitative residue for false information is highly robust. First, we probe whether enhanced salience of the (ir)relevance information eliminates the residue. While this information was already displayed prominently in the baseline experiment and control questions

[1] We establish irrelevance by saliently informing respondents that information is about another company and hence uninformative for their guess. This mirrors a frequently employed strategy of misinformation in reality, namely so-called out-of-context false information (Tonglet et al., 2025).

verified respondents' understanding, we conducted two additional experiments that further raise its salience: one in which the (ir)relevance information appears on the same screen as the belief elicitation rather than on the preceding screen, and one in which respondents must explicitly confirm they have read it before proceeding. In both cases, the qualitative residue remains highly statistically significant and of similar magnitude.

Second, we examine how the timing of corrections to misinformation affects our results. In practice, false information may be debunked prior to the formation of beliefs. Furthermore, in an era of widespread fact-checking, individuals sometimes encounter information that has already been identified as false at the point of first exposure. We demonstrate that the qualitative residue in learning from false information remains robust to two timing variations along those lines. In one variation, the interim belief elicitation is removed from our baseline design. Hence, false information is fully debunked prior to the belief elicitation. In another variation, there is no initial uncertainty about (ir)relevance to begin with. Hence, respondents are fully informed about the signal's irrelevance *before* observing it and stating their posterior beliefs.

Mental Simulation. We outline a simple framework, based on Bordalo et al. (2025a), that can generate the qualitative residue for false information. When a story arrives, its vividness and contextual richness trigger mental simulation: the agent retrieves firms from her memory database using similarity-based recall, which tilts her prior in the direction of the story's valence before any explicit updating begins. She then updates rationally on the analyst signal from this contaminated prior. Upon debunking, she can undo the rational update, but not the prior shift. Purely quantitative signals, by contrast, trigger less mental simulation, and hence lead to less (or no) prior contamination. The framework makes several additional predictions that we then test in mechanism experiments.

An important prediction of our simple framework is that the valence of stories is central to mental simulation: high-valence stories tilt retrieval toward one side of the belief distribution, whereas more neutral content produces more balanced retrieval and a smaller bias. In an additional experiment, we document that replacing evaluative stories with more neutral variants eliminates the qualitative residue: belief movements after false neutral stories are not statistically distinguishable from those after false quantitative signals.

A further experiment provides direct evidence for the key mechanism of our framework, that stories engage mental simulation more strongly than purely quantitative signals. After false stories are debunked, respondents report more vivid mental imagery of the company, compared to debunked quantitative signals. Most strikingly, when asked to describe the company in their own words in a speech recording, the emotional valence of their spontaneous speech tracks the direction of the nullified story roughly twice as strongly as after false quantitative information. Importantly, because the classifier operates on the audio waveform, it captures prosodic features such as intonation, rhythm, energy, rather than lexical content. In other words, respondents

may use neutral words while their delivery betrays the affective imprint of the (nullified) signal. This revealed behavior mitigates concerns about social desirability bias or experimenter demand effects.

Awareness. An open question from our framework is whether people cannot correct for the biased prior once stories are debunked because they simply do not know how to do the correction (for instance because they forgot or are uncertain about the original prior), or because mental simulation operates below awareness. We provide suggestive evidence that both mechanisms might be at play. In separate experiments, we ask respondents at the end of the experiments whether they think the nullified story impacted their final beliefs nonetheless. We interpret this as a proxy for awareness and show that the qualitative residue is largely driven by respondents who report being influenced by the false story yet cannot fully correct for it: awareness of the story's pull does not confer the ability to undo it. At the same time, roughly a third of the average effect appears among respondents who report not being influenced, consistent with stories partly operating below conscious awareness.

Calibration. Belief distortions matter more for aggregate outcomes when decision-makers have high confidence in the beliefs they hold: biases that coincide with high confidence have disproportionate effects on market outcomes (Enke et al., 2023). We elicit respondents' confidence in their posterior beliefs and find that confidence is significantly higher after false stories than after false quantitative signals, even though story-based beliefs lie further from the rational benchmark. The miscalibration runs deeper still: confidence rises more steeply with the magnitude of the belief error under stories than under quantitative information. False stories thus not only distort beliefs more, but also leave respondents most confident precisely when their beliefs are furthest from the truth.

Related literature. We contribute to a literature on misinformation (Allcott and Gentzkow, 2017; Bursztyn et al., 2023; Chopra et al., 2022; Lazer et al., 2018; Pennycook and Rand, 2020, 2026; van der Linden, 2023), retractions, and learning from false or fully uninformative signals (Gneezy and Serra-Garcia, 2025; Kieren and Weber, 2025; Serra-Garcia and Gneezy, 2021). Gonçalves et al. (2025) find that retractions result in diminished belief updating in the context of statistical information, reflecting that updating from retractions is more complex. Thaler (2024) provides evidence of motivated interpretation of fully uninformative quantitative signals in the context of political information.[2] None of these papers, however, compares the belief impact

[2]Literature in psychology documents related phenomena: the continued-influence effect, whereby discredited information shapes reasoning even when people remember the correction (Ecker et al., 2011; Johnson and Seifert, 1994; Lewandowsky et al., 2012; Ross et al., 1975); narrative transportation, which sustains story-consistent beliefs even when readers know the story is fictional (Green and Brock, 2000); processing fluency, whereby ease of comprehension is conflated with truth (Reber et al., 2004).

of false stories versus false quantiative information, analyzes the role of mental simulation or decomposes the residual influence of false stories into conscious and subconscious components.

This paper also contributes to a growing literature on how stories shape beliefs relative to statistical information.[3] Thaler et al. (2025) show that communicators anticipate this asymmetry: experts are substantially more likely to use language rather than numbers when their goal is to persuade rather than inform.[4] Graeber et al. (2024) study selective recall of stories and statistics and provide evidence on the importance of similarity in driving such recall. Our paper, instead, studies contemporaneous *resistance to debunking* of false stories versus purely quantitative information when there are no memory constraints. Our additional contribution is to provide direct behavioral evidence, leveraging paralinguistic features, that mental simulation is engaged differentially across stories and quantitative signals.

We also relate to several literatures on the cognitive foundations of belief formation. First, we relate to literature studying failures of Bayesian updating (Augenblick et al., 2024; Ba et al., 2024; Conlon et al., 2026; Enke, 2020; Enke and Zimmermann, 2019; Esponda et al., 2024, 2023; Hartzmark et al., 2021). Second, our work connects to work examining cue-based belief formation (Bordalo et al., 2025b,b,c, 2026; Conlon and Kwon, 2026; Gennaioli and Shleifer, 2010; Graeber et al., 2026b; Henkel et al., 2026) and mental simulation (Bordalo et al., 2025a). Conlon and Kwon (2026) develop and experimentally validate a model in which explicitly uninformative cues distort beliefs by bringing similar states to mind, with the magnitude of belief shifts predictable from independently elicited similarity judgments. Our work also builds on a large literature in psychology on mental simulation (Schacter et al., 2007, 2017; Szpunar, 2010; Taylor et al., 1998).

The rest of the paper is structured as follows. In Section 2, we describe the baseline experimental design. Section 3 presents our key results on the qualitative residue for false news as well as robustness evidence. Section 4 presents our simple framework of mental simulation. In Section 5 we present our mechanism evidence, before concluding in Section 6.

[3]For a review of the literature on stories versus statistics, see Graeber et al. (2026a). Our focus on stories is distinct from a growing literature on narratives (Shiller, 2017). Stories, as we study them, are data which participants use to form beliefs: concrete accounts of specific events, such as a particular analyst's assessment of a particular company. Narratives, by contrast, organize how people interpret data (Andre et al., 2022, 2026; Barron and Fries, 2024, 2025; Eliaz and Spiegler, 2020, 2024; Eliaz et al., 2025; Fan and Fries, 2026; Schwartzstein and Sunderam, 2021, 2024; Spiegler, 2016).

[4]We also build on evidence on stories from psychology (Nisbett and Borgida, 1975; Nisbett and Ross, 1980; Tulving, 1972).

2 Experimental Design and Sample

2.1 Design

Our baseline design is guided by four objectives. First, beliefs must be elicited in an incentive-compatible way, so that responses reflect genuine assessments. Second, we need a clean rational benchmark for residual belief movement after debunking, even though the qualitative nature of stories makes them difficult to incorporate into a standard Bayesian framework. Third, the debunking itself must be easy for subjects to understand and salient at the point of belief elicitation, so that any residue cannot be attributed to subjects missing the correction. Fourth, the two formats must be sufficiently comparable, with both arriving in a way that feels natural, so that any difference in their effect can be attributed to specific qualitative features.

Task structure. For each of four hypothetical companies, respondents learn that the company has received a number of analyst evaluations, each either positive or negative. We truthfully inform respondents that the fraction of positive evaluations is drawn from a uniform distribution, independently across companies. This likely induces a fuzzy prior among respondents. Respondents then observe a signal, a single analyst's recommendation, either positive or negative, and report an interim belief about the fraction of positive evaluations. At that point, the relevance of the signal remains uncertain. In particular, respondents are told that with 50 percent probability, the signal is about the focal company (and hence relevant), and with 50 percent it is about another unrelated company (and hence irrelevant). The latter mirrors many real-world cases of false information, where events (e.g., a crime committed by a refugee or a war crime against civilians) are falsely but purposefully attributed to a particular time and place, when in fact they occurred elsewhere, at a different time.[5] After stating their interim belief, respondents learn whether the signal was relevant or irrelevant for the focal company and report a posterior belief. This yields eight incentivized belief reports per respondent.

Overview of treatments. Within this common task structure, the four companies differ along two within-subject dimensions. *Information type*: for each company, the signal is either the analyst's quantitative recommendation alone ("statistic only") or the same recommendation accompanied by a qualitative story ("statistic + story").[6]The story is matched to the direction of the recommendation, so the two conditions are aligned on directional content and differ only in format. Each respondent sees two companies in each format. *Signal relevance.* As described above, after the interim belief the respondent learns whether the signal was relevant.

[5]These type of fake news are often refereed to as out of context misinformation (Aneja et al., 2023; Brennen et al., 2020; Tonglet et al., 2025).

[6]We use the term *statistic* as shorthand for this quantitative single-observation signal throughout, even though statistics are conventionally understood as summaries of many data points.

An irrelevant signal nullifies the recommendation, rendering the statistic-only and statistic-plus-story conditions informationally equivalent to receiving no signal at all, since neither carries any information about company quality. Treatment assignment is blocked, so that each respondent encounters exactly one of the four format $\times$ relevance cells. Signal valence is independently randomized. Figure 1 illustrates the design. Between subjects, we additionally vary whether the information is conveyed through text or video.

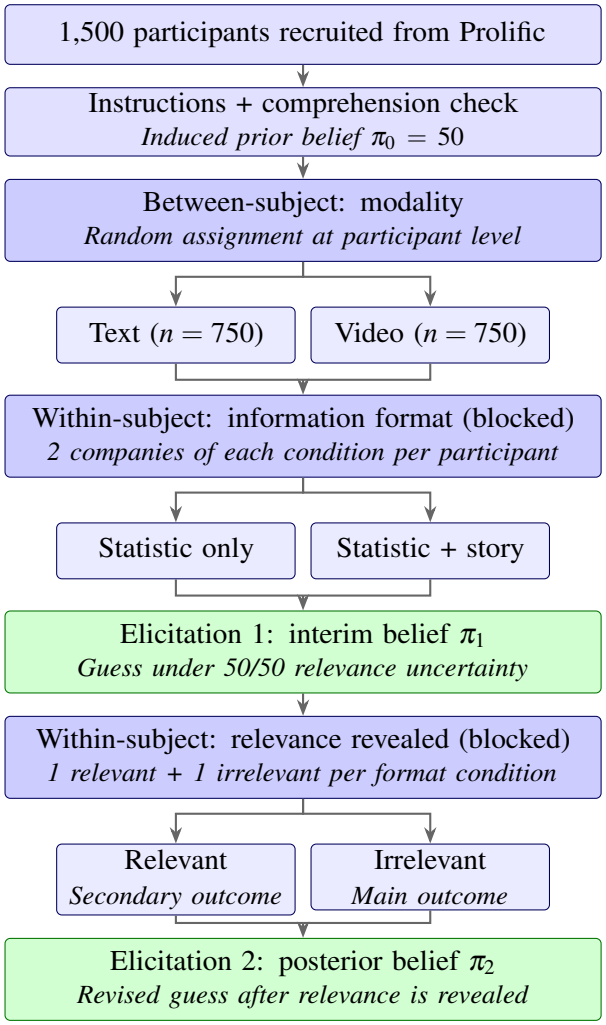


Figure 1: Experiment Design

Story vs. statistic treatment. The statistic treatment provides information about the analyst's recommendation. For example, in the positive case, participants read:

> *A randomly drawn analyst thinks the company is good.'*

The story augments the analyst's recommendation with a brief qualitative elaboration. The story is matched to the direction of the analyst's recommendation. For the positive case, respondents receive the following information:

A randomly drawn analyst thinks the company is good and said the following about the company: 'My evaluation of this company is positive. What sets this company apart is the exceptional caliber of its leadership team. The executives bring a rare combination of vision and operational discipline, consistently making decisions that prioritize long-term value over short-term gains. Their track record of navigating challenging environments with composure and clarity inspires confidence across the organization. A strong culture flows from the top, attracting and retaining talented people at every level. When leadership is this capable and aligned, the entire enterprise benefits. This is a management team that earns trust, and in my assessment, that trust is well placed.'

Modality treatment. Between subjects, we vary whether the story and statistic are conveyed through text or a short video featuring an AI-generated animated analyst reading the same text (with subtitles), enriching the sensory channel while holding the transcript fixed. Figure A5 shows an example screenshot for the video treatment. Each company features a different AI-generated animated analyst in order to minimize potential informational spillovers across companies.[7]

Belief elicitation. Beliefs are elicited twice per company. The *interim belief* is reported after observing the signal but before learning about its relevance. The *posterior belief* is reported after learning the signal's relevance with certainty. Beliefs are incentivized using a binarized scoring rule (Hossain and Okui, 2013).[8] For each respondent, one of the eight beliefs is randomly selected to determine a $30 bonus paid to one in ten participants, with the bonus probability decreasing in the squared distance between the stated belief and the truth.

Bayesian benchmark. The irrelevance manipulation is central to our identification strategy. For an informative signal, computing a Bayesian benchmark requires assumptions about the signal's diagnosticity and, for stories in particular, about how qualitative content enters a Bayesian update at all, neither of which is straightforward. When the signal is irrelevant, these difficulties vanish: the Bayesian posterior collapses to the prior regardless of the signal's format, the analyst's quality, or any assumptions about how stories map into informational content. The benchmark for residual belief movement is therefore exactly zero, and identical across our two main conditions.[9]

[7] All stories and links to the videos can be found in Appendix B.2.

[8] The probability of receiving the bonus payment is given by $\Pr(\text{bonus}) = 1 - \frac{1}{10{,}000}(\text{guess} - \text{truth})^2$, where the guess and truth are expressed in percentage points. This scoring rule is incentive compatible even under risk aversion. Danz et al. (2022) show that binarized scoring rules can introduce bias toward 50 in stated beliefs. This is not a concern for our identification, which relies on the within-subject comparison of posteriors across formats; any such bias would apply symmetrically to the statistic-only and story conditions and thus cancels out in the differential. Our central conclusions are also corroborated by speech-prosody evidence (Section 5.2), which is immune to concerns about scoring rules.

[9] For a recent survey of models that depart from Bayes' rule, see Ortoleva (2024).

Outcome measure. Against this benchmark, our primary outcome is *posterior belief movement* after irrelevant information: the direction-adjusted distance between the posterior belief and the prior of 50%, signed so that positive values correspond to movement in the direction of the signal. Any non-zero value is a departure from Bayesian updating, and any difference between stories and statistics identifies what we call the *qualitative residue*, the residual belief movement attributable to the qualitative content of the story.

Comprehension and attention checks. Before the main task, respondents complete an attention screen and a battery of six comprehension questions covering: the number of companies presented, the independence of company draws, the real-stakes nature of the task, the appropriate use of relevant information, the appropriate disregard of irrelevant information, and the resolution of relevance uncertainty. Respondents who fail any comprehension question after two attempts cannot proceed to the main task. We also implement an audio attention check (numbers played through speakers to be transcribed) to minimize the risk posed by bots. The full set of instructions is reproduced in Appendix B.

2.2 Sample

Sample, recruitment, and exclusion criteria. We collect 1,500 completed responses for the baseline experiment, with 750 per modality condition (video versus text). Respondents are recruited via Prolific, a survey platform widely used in social-science research (Peer et al., 2022). The data collection took place between March 31, 2026 and April 1, 2026; the median completion time was approximately 13 minutes; respondents received a completion payment of $2.75 (equivalent to an hourly rate of $12.27). As pre-specified, we exclude respondents who (i) fail an attention check, or (ii) fail any comprehension question after two attempts. Online Appendix A.5 reports demographics and attrition rates across all studies; attrition does not differ significantly across conditions and is therefore not a threat to internal validity, since our key treatment variation is within-subject.

Pre-registration. The data collection was pre-registered: `https://aspredicted.org/tp9xe3.pdf`. The pre-registration specifies the sample size, primary outcome, main specification, exclusion criteria, and the within-subject treatment comparison. We discuss any deviations from the pre-registration in Online Appendix A.9; all main-text analyses follow the pre-registration unless explicitly stated otherwise. An overview of all data collections including links to the pre-registrations can be found in Appendix Table A1.

3 Results

This section presents our main results. We document a sharp asymmetry across formats: against a Bayesian benchmark of zero residual updating, debunked statistics shift posterior beliefs only modestly, while debunked stories shift them substantially more, yielding what we term a *qualitative residue*. The residue replicates across a battery of preregistered experiments that vary the salience and timing of the relevance information.

3.1 Empirical Strategy

For respondent i evaluating company j, define posterior belief movement as

$$Y_{ij} = d_{ij} \cdot (\hat{\pi}_{ij}^{\text{post}} - 50), \tag{1}$$

where $\hat{\pi}_{ij}^{\text{post}} \in [0, 100]$ is respondent i's reported posterior belief (in percentage points) for company j, and $d_{ij} \in \{-1, +1\}$ is a signing variable defined as

$$d_{ij} = \begin{cases} +1 & \text{if the signal in scenario } j \text{ for respondent } i \text{ was positive,} \\ -1 & \text{if the signal was negative.} \end{cases}$$

This direction adjustment ensures that Y_{ij} measures movement *in the direction of the (nullified) signal*, regardless of valence. Under Bayesian updating with a fully nullified signal, $E[Y_{ij}] = 0$ in both conditions; any positive value of Y_{ij} indicates residual influence of the signal on beliefs.

Our main specification restricts attention to posterior beliefs when the signal was irrelevant.[10] Concretely, we estimate

$$Y_{ij} = \alpha + \beta \, Story_{ij} + u_{ij}, \tag{2}$$

where $Story_{ij} \in \{0, 1\}$ indicates whether respondent i received a story for company j. Standard errors are clustered at the respondent level to account for the within-subject blocked design (Abadie et al., 2023). The coefficient α identifies the residual belief impact of false statistics against the rational benchmark of zero; the coefficient β identifies the qualitative residue, that is, the additional belief impact contributed by qualitative elaboration. The intercept α is unrestricted: $\alpha = 0$ corresponds to fully Bayesian updating in the statistic-only condition; $\alpha > 0$ indicates that even bare statistics carry some residual influence after debunking.

To assess modality effects, we estimate the saturated within-between specification

$$Y_{ij} = \alpha + \beta_1 \, Story_{ij} + \beta_2 \, Video_i + \beta_3 \, (Story_{ij} \times Video_i) + u_{ij}, \tag{3}$$

[10] We present results for interim beliefs under uncertain relevance, and posterior beliefs after relevant signals, in Appendix A.4. These complementary analyses verify that immediate updating is well-behaved across formats and that respondents distinguish appropriately between relevant and irrelevant signals.

where $Video_i \in \{0,1\}$ indicates assignment to the video modality (between subjects). The coefficient β_3 tests whether the qualitative residue is amplified by video delivery.

3.2 The Qualitative Residue

Main effects. Table 1 and Figure 2 present the main result. We see that false statistics barely moves beliefs. While in our baseline experiment statistics have a belief impact of around 1 percentage point ($SE = 0.42$), there is close to zero impact across our robustness experiments (see Section 3.4). False stories, on the other hand, shift beliefs substantially more. Residual belief movement in the story condition is 4.2 percentage points ($SE = 0.47$), meaning that the qualitative residue is around 3 percentage points. Table 1 reports the corresponding regression estimates: the constant identifies the residual belief impact of false statistics ($\hat{\alpha} \approx 1.3$ pp, $p = 0.003$), and the story coefficient identifies the qualitative residue ($\hat{\beta} \approx 3.0$ pp, $SE = 0.51$, $p < 0.001$).

While the magnitudes might appear modest, two features make them informative. First, recall that the rational benchmark is zero belief movement. Hence, any residual movement constitutes a departure from Bayesian updating, and the false-story condition produces a departure three times larger. Second, the setting deliberately minimizes the story's advantage: The story is short, generic, and attached to a hypothetical company.

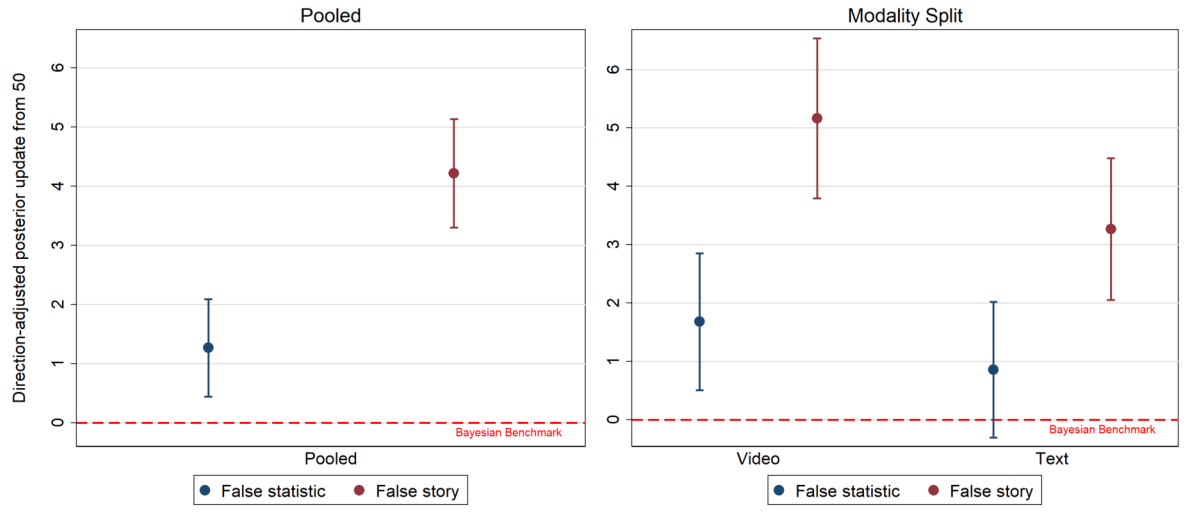


Figure 2: POSTERIOR BELIEFS AFTER IRRELEVANT SIGNALS. Left panel: direction-adjusted posterior belief movement from the prior of 50, pooling across modalities. Right panel: disaggregated by story modality. Blue: false statistic; red: false story. The dashed line marks the Bayesian benchmark of zero.

Modality differences. The right panel of Figure 2 disaggregates belief impact by modality. False-story respondents in the video condition show posterior movement of approximately 5.2

percentage points, compared with 3.3 in text. False-statistic respondents show approximately 1.7 percentage points in video versus 0.9 in text. The qualitative residue is 46 percent larger in video (approximately 3.5 pp) than in text (approximately 2.4 pp), but the difference between modalities is not statistically significant (interaction coefficient $\hat{\beta}_3 \approx 1.1$ pp, $p = 0.298$; equation 3, Table 1, column (4)).

Result 1. False stories produce residual posterior belief movement roughly three times larger than false statistics. The qualitative residue is statistically significant against a benchmark of zero. Stories conveyed through video generate somewhat larger belief shifts than text-based stories.

Table 1: Main analysis: effect of false story on posterior beliefs

Dependent variable	Posterior belief movement			
	Pooled (1)	Video (2)	Text (3)	Interaction (4)
Story	2.949***	3.484***	2.413***	2.413***
	(0.515)	(0.763)	(0.691)	(0.691)
Video				0.825
				(0.842)
Video × Story				1.071
				(1.029)
Baseline	1.269***	1.681***	0.856	0.856
	(0.421)	(0.598)	(0.593)	(0.593)
N	3,000	1,500	1,500	3,000

Notes: This table uses data from the Baseline Experiment. OLS regressions. Dependent variable is the direction-adjusted posterior belief from 50; sample restricted to irrelevant conditions (false statistic and false story). Baseline is the false statistic condition. Columns (2) and (3) restrict to video and text modality, respectively; column (4) adds a video indicator and its interaction with false story. Standard errors clustered at the participant level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

3.3 Heterogeneity

We probe two sources of heterogeneity in the qualitative residue: demographics and order of presentation. These analyses were not pre-registered.

Demographics. Appendix Table A8 examines heterogeneity by gender, age, and education. The qualitative residue is positive and statistically significant in every demographic subsample we examine. A joint test of equality of the qualitative residue across the six demographic subsamples fails to reject the null of a common effect ($p = 0.436$), consistent with a broad-based phenomenon.

Order of presentation. The within-subject blocked design randomizes the order in which respondents encounter the four format-relevance cells. We can hence look at results separately for each round. Notice that, by focusing on round 1, we can actually exploit the between-participant variation of false stories versus false statistics. Appendix Table A9 reports the qualitative residue separately for each round (scenario position). The qualitative residue is positive and significant in all four positions; a joint test of equality across positions does not reject equality ($p = 0.542$). Similarly, Appendix Table A10 splits the sample by whether the false story scenario was encountered before or after the false statistic scenario. The qualitative residue is virtually identical across the two groups (2.8 vs. 3.1 percentage points), and the difference is far from significant ($p = 0.818$), ruling out carryover effects between the two irrelevant conditions.

3.4 Robustness

In this design, we assess the robustness of our main findings to variation in the salience of the relevance information and the timing of the debunkig.

Salience of relevance information. In the main design, respondents learn whether the signal is relevant or irrelevant on the screen immediately prior to the posterior belief elicitation. The information about relevance was already displayed very prominently in the baseline experiment and control questions verified respondents' understanding.[11] We conducted two additional experiments that further increase salience, each preregistered with a target of 750 completed responses.

The first displays the relevance information directly on the decision screen. The design uses video-only presentation and is otherwise identical to the main experiment. The left panel of Figure 3 presents the results. The qualitative residue remains statistically significant, with a magnitude broadly consistent with the main experiment ($\hat{\beta} = 2.2$ pp, $p = 0.002$).

The second adds a confirmation step: before stating their posterior belief, respondents must confirm that they have read the relevance information.[12] This design also uses video-only presentation. The right panel of Figure 3 presents the results. The qualitative residue is again robust ($\hat{\beta} = 4.4$ pp, $p < 0.001$), ruling out the possibility that respondents overlooked or forgot the relevance revelation.

In sum, we show that the qualitative residue is robust to these design variations. It survives both enhanced salience of the relevance information and a forced confirmation step. This further

[11] Of the 1,500 baseline respondents, 1,271 (84.7%) answered all comprehension questions correctly on the first attempt. Appendix Table A2 replicates Table 1 for this subsample; results are virtually unchanged. Appendix Table A3 further shows that even among participants who fully corrected to the prior in the false statistic condition, the false story still produces a significant residual posterior movement, retaining roughly 70% of the full-sample qualitative residue.

[12] 99.3% of participants confirmed reading the relevance information for all scenarios.

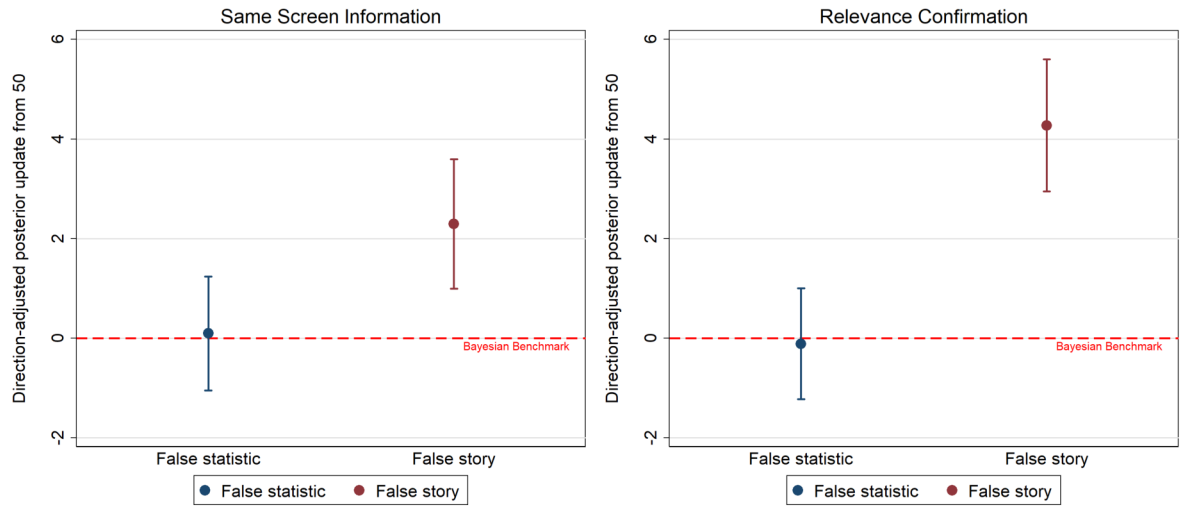


Figure 3: POSTERIOR BELIEFS AFTER IRRELEVANT SIGNALS. Left panel: relevance information displayed on the same screen as the posterior belief elicitation. Right panel: respondents must confirm reading the relevance information before stating their belief. Direction-adjusted posterior belief movement from the prior of 50. Dots are sample means; bars are 95% confidence intervals. The dashed line marks the Bayesian benchmark of zero.

clarifies that respondents understand that the entire signal has been nullified and yet remain influenced by the story.

Variations in timing of debunking. We investigate whether the timing of misinformation corrections influences our findings. In real-world settings, false claims are sometimes debunked before individuals have formed beliefs about them. Moreover, in today's fact-checking landscape, people frequently come across information that has already been flagged as false at the moment they first encounter it. We conducted two additional experiments that vary the timing of debunking along these lines.

First, we conducted an experiment that removes the interim elicitation entirely: respondents observe the signal under uncertainty regarding its relevance, and then learn its relevance before stating a single posterior belief. The design uses video-only presentation and is otherwise identical to the main experiment. We collect 1,000 completed responses.

The left panel of Figure 4 presents the results. The qualitative residue replicates with magnitude comparable to the main experiment: false stories produce residual posterior movement of approximately 4 percentage points, compared with approximately 1 for false statistics ($\hat{\beta} = 2.9$ pp, $p < 0.001$). The residue is not an artifact of the two-stage measurement.

Second, we conducted an experiment ($N = 1{,}000$, video-only) where participants learn about the relevance at the beginning of each round, before observing the additional information. Hence, they already know that the information is irrelevant before seeing it. The right panel of Figure 4 shows that the qualitative residue remains unchanged ($\hat{\beta} = 3.1$ pp, $p < 0.001$), implying that it

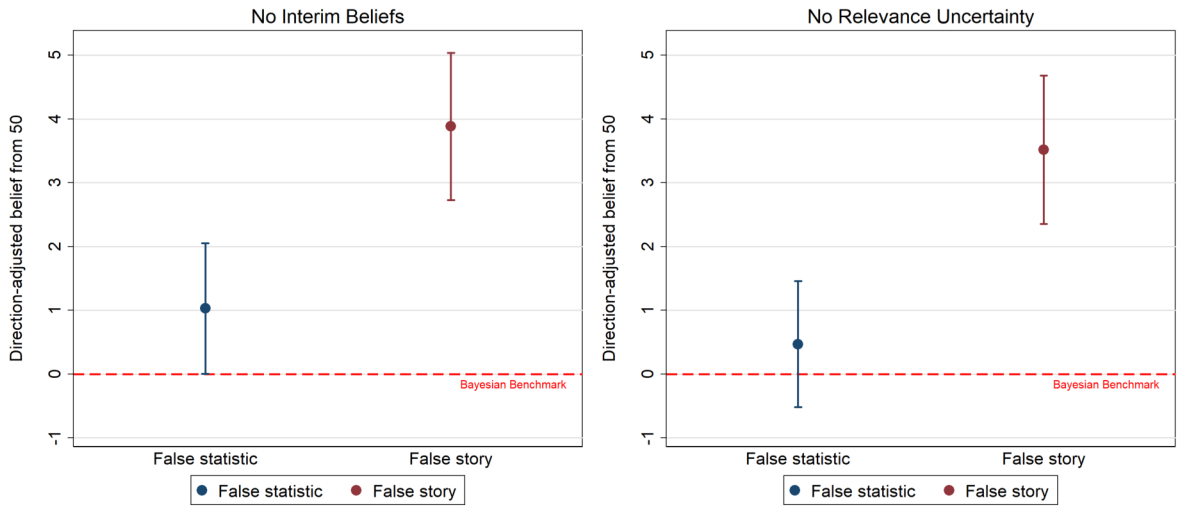


Figure 4: POSTERIOR BELIEFS AFTER IRRELEVANT SIGNALS. Left panel: design without interim beliefs, i.e. the relevance of the signal was uncertain when participants observed the signal, but resolved before they stated their posterior guess. Right panel: design without uncertainty about the relevance of the signal, i.e. participants learned about relevance or irrelevance prior to observing the signal. Direction-adjusted posterior belief movement from the prior of 50. Dots are sample means; bars are 95% confidence intervals. The dashed line marks the Bayesian benchmark of zero.

is not crucial for our results that uncertainty about relevance is only gradually revealed.

Table 2 summarizes the qualitative residue estimates across all experimental designs, documenting its robustness across a wide range of variations in salience, timing, and presentation format.

Result 2. The qualitative residue is robust to enhanced salience of the relevance information, removal of the interim belief elicitation, and removal of initial uncertainty about signal relevance. In all cases, false stories produce significantly larger residual belief movement than false statistics.

Table 2: QUALITATIVE RESIDUE ACROSS EXPERIMENTS

			Posterior belief movement after irrelevant signals (pp)			
Experiment	N	Modality	False stat $\hat{\alpha}$	False story $\hat{\alpha}+\hat{\beta}$	Residue $\hat{\beta}$	p-value
Main experiment						
Baseline (pooled)	1,500	Text & video	1.27	4.22	2.95	<0.001
Video subsample	750	Video	1.68	5.17	3.48	<0.001
Text subsample	750	Text	0.86	3.27	2.41	<0.001
Round 1 only	780	Text & video	0.48	4.07	3.59	0.003
Robustness experiments						
Same-screen information	750	Video	0.09	2.29	2.20	0.002
Relevance confirmation	750	Video	-0.11	4.27	4.38	<0.001
No interim beliefs	1,000	Video	1.03	3.89	2.86	<0.001
No relevance uncertainty	1,000	Video	0.47	3.52	3.05	<0.001
Mechanism experiments						
Neutral stories	750	Text	0.80	1.68	0.87	0.267
Mental simulation	1,000	Video	1.26	4.63	3.37	<0.001

Notes: Each row reports point estimates from $Y_{ij} = \alpha + \beta\, Story_{ij} + u_{ij}$, restricted to respondents who learned the signal was irrelevant. $\hat{\alpha}$ captures the residual belief movement of false statistics against the rational benchmark of zero; $\hat{\beta}$ identifies the qualitative residue. The Baseline (pooled) row reports the pooled estimate across modalities; the indented sub-rows decompose this into the within-subject text and video samples (corresponding to columns (1)–(3) of Table 1). Standard errors clustered at the participant level. The Bayesian benchmark is zero in all rows.

4 A Simple Framework of Mental Simulation

This basic framework closely follows our experimental design. The framework is an amended and simplified version of Bordalo et al. (2025a). It is meant to organize our main results and as a means of deriving testable predictions regarding the underlying mechanisms of the qualitative residue.

4.1 Information Structure

Firm Quality and the Prior. Each company is characterized by its quality $q \in [0,1]$, defined as the *fraction of analysts who evaluate it positively*. The agent has a uniform prior over firm quality:

$$q \sim \text{Uniform}[0,1].$$

We work with the agent's belief about q, summarized by its expectation. All beliefs are reported on a $[0,1]$ scale; the induced prior belief is hence $\pi_0 = \frac{1}{2}$.

Analyst Signal and Signal Relevance. The agent observes the recommendation of one randomly drawn analyst. The analyst's recommendation is $r_q \in \{+1, -1\}$ (positive or negative).

The signal's relevance $\rho \in \{1, 0\}$ is unknown to the agent at the time she forms her interim belief. With equal probability the recommendation is relevant ($\rho = 1$, the analyst genuinely evaluated this firm) or irrelevant ($\rho = 0$, the recommendation pertains to a different firm and carries zero information about q). Formally:

$$\Pr(\rho = 1) = \Pr(\rho = 0) = \tfrac{1}{2}.$$

Relevance is revealed to the agent *after* she reports her interim belief.

Rational Benchmark. The agent reports two incentivized beliefs:

- **Interim belief** π_1: after observing the signal but before learning ρ.
- **Posterior belief** π_2: after learning ρ with certainty.

Our key outcome is **posterior belief movement** when $\rho = 0$, defined as the simple difference between posterior belief and prior:

$$\Delta\pi \equiv \pi_2 - \pi_0.$$

When the agent learns $\rho = 0$, the signal is fully nullified. The Bayesian posterior collapses back to the prior regardless of the format in which r was delivered:

$$\pi_2^{\text{rational}} = \pi_0 = \tfrac{1}{2}, \qquad \Delta\pi^{\text{rational}} = 0.$$

Information Format. The information r is presented in one of two formats:

1. **Statistic only.** The agent sees the analyst's recommendation r_q alone.
2. **Statistic + story.** The agent sees the analyst's recommendation r_q plus a story describing the reasoning underlying the recommendation.

Information is captured as $r = (r_q, V_r^{\text{context}})$ and contains the quantitative signal as well as qualitative features of information. V_r^{context} is a vector of length M that captures whether certain qualitative features are part of the information (e.g., valence of the signal, quality of the CEO, market competition, innovativeness) and if present, whether they are positive or negative. The vector begins with the type of object, in this case whether it is about a firm, which is stored as (Yes, No). The vector next contains the valence of the information. For this and all subsequent features, the value of this feature is stored as (Pos, Neg). If a feature is not mentioned or is mentioned in a neutral fashion, this is stored as NA. A statistic

lacks such qualitative features (beyond the valence), hence it is captured as, for instance, $r = (r_q, Yes, Pos, NA, ..., NA)$, whereas a story at least contains some qualitative features such that, for instance, $r = (r_q, Yes, Pos, Pos, Neg, Pos, ..., NA)$. We assume that a positive story contains substantially more positive than negative qualitative features, and vice versa for negative stories. This is satisfied by construction in our baseline experimental stimuli (see Appendix B.2).

4.2 Mental Simulation

Upon receipt of information, the agent engages in mental simulation: the information prompts retrieval of firms from her memory database using similarity-based recall. The similarity between the firm that comes to mind and the information determines the degree and vividness of the mental simulation. Retrieval determines which firms are top of mind and as a consequence shifts her operative prior over q. Once this new prior has been formed, the agent is fully Bayesian in the way she incorporates quantitative information.

The Mental Database and Similarity-Based Retrieval. The agent maintains a mental database $\mathscr{D}$ of firms, each also characterized by a quality index q as well as a vector V_d^{context}. V is also of length M and contains the same features in the same order as V_r^{context}. We assume the database is symmetric, so the agent has both good and bad firms in her database: the average quality of a randomly drawn firm is $\frac{1}{2}$. We further assume that good firms, beyond the valence, have strictly more positive than negative features and vice versa for bad firms.

Once a signal is received, this triggers mental simulation and firms from the mental database come to mind. Mental simulation is guided by similarity-based recall where the intensity or vividness of the mental simulation is governed by similarity. The recall probability $p(d)$ for a given firm in the database is given by the relative similarity of its contextual features to the cue:

$$p(d) \;=\; \frac{S\big(V_r^{\text{context}},\, V_d^{\text{context}}\big)}{\sum\limits_{d' \in \mathscr{D}} S\big(V_r^{\text{context}},\, V_{d'}^{\text{context}}\big)}, \tag{4}$$

where $S(\cdot,\cdot)$ is the similarity function. We define $S(\cdot,\cdot)$ as follows,

$$S: \prod_{i=1}^{M} V_i \times \prod_{i=1}^{M} V_i \to [0,1],$$

and require that it is symmetric, increasing in the number of features that share the same value (positive or negative)[13], equals 1 if and only if the two vectors are identical and equals 0 if and only if no feature is shared by the two vectors.

[13]Only features that share the same value increase similarity. Features that are absent or neutral *NA* do not increase similarity.

Prior Contamination Through Mental Simulation. We assume that similarity-based recall determines which firm from the database comes to mind through mental simulation. The quality of this firm then forms the new "contaminated" prior.[14]

Thus, the agent's contaminated prior over firm quality after retrieval is shifted from $\pi_0 = \frac{1}{2}$ to $\tilde{\pi}_0$.[15] Due to similarity-based recall, upon receipt of a story, $E(\tilde{\pi}_0)$ is shifted in the direction of the story. In other words, stories with positive valence induce an upward shift of the prior, stories with negative valence induce a downward shift. Since statistics do not contain contextual features except for the object and valence, the shift in prior will be less pronounced. This directly follows from our assumption that good firms in the database have strictly more positive than negative features.

This shift in prior occurs through the simulation process, *before* the agent explicitly processes the signal r_q.

Assumption 1: The retrieval-induced prior shift overwrites the original prior. The agent experiences it as a genuine adjustment to her background model of firm quality. She therefore cannot reverse it upon learning $\rho = 0$.

Assumption 1 could capture a process of prior adjustment where the old prior is not stored in working memory and hence cannot be restored. Alternatively, we could think of mental simulation as a process that operates below awareness or that is not explicitly linked to the information from the perspective of the agent and can hence not be corrected for. Indeed, a literature in cognitive psychology argues that mental simulation largely operates below awareness (Barsalou, 1999; Clark, 2016).

Bayesian Updating and Debunking. Starting from the contaminated prior $\tilde{\pi}_0$, the agent updates on the analyst signal r exactly as in the statistician mode, but now using her shifted prior.

Upon learning $\rho = 0$, the agent recognizes that her Bayesian update from $\tilde{\pi}_0$ was based on a now-irrelevant signal. She reverses this component. However, she cannot reverse the prior contamination. Hence $\pi_2 = \tilde{\pi}_0$.

4.3 The Qualitative Residue

We can state the following Proposition:

Proposition 1 (Qualitative Residue)**.** *Among respondents who learn that the signal was irrelevant*

[14]The framework models the story's effect as operating through prior contamination rather than stronger updating on the analyst signal. The *No Relevance Uncertainty* experiment is consistent with this interpretation: the qualitative residue survives even when respondents learn of the signal's irrelevance before observing it, leaving little scope for signal-based updating.

[15]Note that the contaminated prior $\tilde{\pi}_0$ is different from the elicited interim belief π_1: the interim belief is reported after the agent has updated on the analyst signal from $\tilde{\pi}_0$, whereas $\tilde{\pi}_0$ arises before the signal is processed.

($\rho = 0$), posterior belief movement satisfies:

$$|E(\pi_2(story)) - \pi_0| > |E(\pi_2(stat)) - \pi_0|.$$

Corollary 1 (Neutral Valence Stories)**.** False stories with a neutral valence should lead to a smaller belief impact compared to stories with strong valence. This is because such stories will share fewer similarities with firms in the mental database.

5 Mechanisms

The main experiment establishes a qualitative residue: false stories shift beliefs more than false statistics. Guided by our framework, we now characterize this residue with different pre-registered mechanism experiments.

5.1 Story Valence

The stories in the main experiment are evaluative: they provide a persuasive account of why the company is good or bad. In other words, the stories have a clear valence. According to our framework, more neutral stories should lead to a smaller belief movement of false stories. Hence, we conducted a separate experiment where we replaced the valenced stories with more neutral variants that describe the company in less evaluative terms, holding the quantity of qualitative information fixed while removing persuasive content. The design is otherwise identical to the main experiment restricted to text-only presentation.[16] For example, participants read:

> *A randomly drawn analyst thinks the company is good and said the following about the company: 'My evaluation of this company is positive. The company is in the software industry. Its employees are based in offices in several locations. The firm's products relate to data storage, internal communications, and digital record-keeping. The company charges its clients on a recurring basis. A portion of the company's budget is spent on developing new products. Its clients are organizations of varying sizes in both the public and private sector. The company's operations are organized into several internal divisions. The firm has been in operation for a number of years.'*

[16]The neutral stories experiment uses text-only presentation because our video stimuli feature animated analyst avatars, imagery, and music calibrated to the evaluative content. Producing neutral video stories would have either required redesigning these elements alongside the text, changing more than just the content, or left neutral text paired with evaluative audiovisual elements, introducing a mismatch. Text-only presentation allows us to vary only the story content while holding everything else fixed. The story-statistic gap is statistically significant in the text condition of the main experiment ($\hat{\beta} = 2.41$ pp, $p < 0.001$), confirming that text delivery is sufficiently powerful to detect the effect when evaluative content is present.

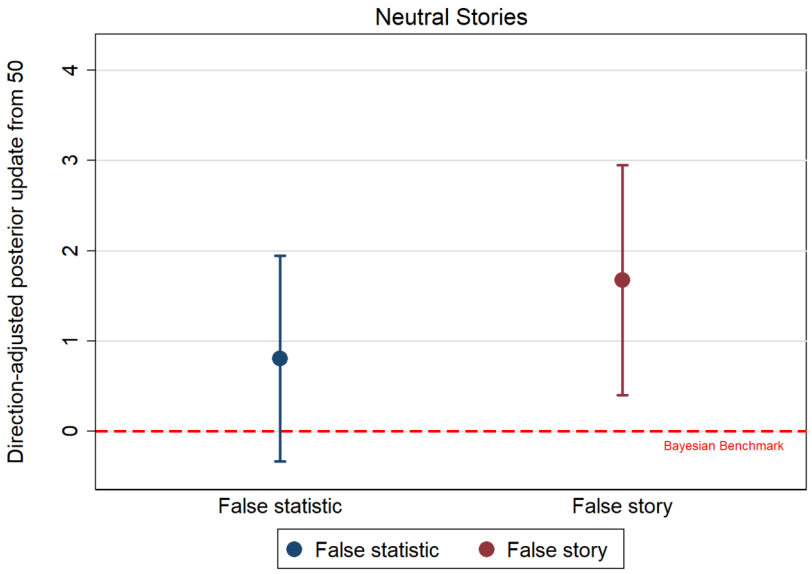


Figure 5: POSTERIOR BELIEFS AFTER IRRELEVANT SIGNALS (NEUTRAL STORIES). Direction-adjusted posterior belief movement from the prior of 50. Dots are sample means; bars are 95% confidence intervals. The dashed line marks the Bayesian benchmark of zero.

Figure 5 presents the results. With neutral stories, in line with our framework, the qualitative residue is no longer statistically detectable: belief movements in the false statistic and false story conditions are not significantly different from one another ($\hat{\beta} = 0.87$ pp, $p = 0.267$; 95% CI: [-0.67, 2.41]). Stripped of evaluative content, qualitative format loses its differential pull on beliefs. The residue documented in the main experiment operates through the persuasive content of the story, not through qualitative elaboration per se. Mental simulation amplifies evaluative narratives; absent valence, there is little left for it to amplify.

Result 3. Without evaluative content, the qualitative residue is no longer statistically significant. The residue is driven by the persuasive content of stories rather than by qualitative format alone.

5.2 Mental Simulation

What does the qualitative elaboration do that the bare statistic does not? Our framework proposes that the contextual richness of stories triggers more intense mental simulation. We conduct an additional experiment that elicits direct measures of mental simulation alongside the belief task, providing both self-reported and revealed evidence that stories indeed engage simulation to a greater extent than statistics.

Design. The experiment replicates the main design with video-only presentation and adds several measures elicited immediately after the posterior belief for the two irrelevant-signal scenarios. First, respondents provide an *open-ended audio response* describing the company,

speaking for at least 20 seconds.[17] The exact prompt instructs respondents: "Please share your overall impression of the company. What comes to mind when you think about this company? You might think about what kind of company it seems like, how you feel about it, or anything else that comes to mind. Please speak for at least 20 seconds, and briefly play back your recording to ensure it has been recorded correctly before proceeding." Crucially, the prompt makes no reference to the analyst, the recommendation, or its relevance, respondents are asked only for their general impression of the company. Second, respondents rate the *vividness of their mental imagery* of the company on a 1–7 scale ("How vividly can you imagine the company?"). We collect 1,000 completed responses.

Speech-emotion classifier. We extract the affective valence of each recording using a speech emotion recognition model that takes raw audio waveforms as input and predicts continuous valence scores by processing the acoustic signal through a fine-tuned WavLM Large transformer (Goncalves et al., 2024). The model is trained on the MSP-Podcast corpus, the largest publicly available speech-emotion dataset, and outputs a continuous valence score capturing the affective tone of vocal delivery. Because the classifier operates on the audio waveform rather than transcribed text, it captures prosodic features—intonation, rhythm, energy—rather than lexical content, which is the desired behavioral signature: respondents may use neutral words while their delivery betrays the affective imprint of the (nullified) signal. Research in psychology shows that such behavioral signatures are less subject to conscious control (DePaulo et al., 2003; Ekman and Friesen, 1969; Scherer, 2003).

Pre-processing and quality control. We exclude recordings that are silent, shorter than 10 seconds, or flagged as low-quality by automatic checks (e.g., overwhelming background noise). 96.4% of recordings pass these checks.[18] The mean recording duration is approximately 31 seconds. The distribution of valence scores across recordings is well-spread, with no evidence of ceiling or floor effects. As a face validity check, Odyssey valence scores are significantly higher for positive-signal recordings than for negative-signal recordings ($p < 0.001$), confirming that the classifier captures meaningful variation in affective tone in our data.

Results. Figure 6 summarizes the three main outcomes. Panel (a) replicates the qualitative residue: false stories produce posterior movement of approximately 4.6 percentage points compared with 1.2 for false statistics ($\hat{\beta} = 3.4$ pp, $p < 0.001$). This reproduces the core finding. Panel (b) shows that the self-reported measure moves in the same direction as beliefs: respondents rate their mental imagery as significantly more vivid after false stories than after false statistics

[17] Speech data have been used in prior work by economists (Galasso et al., 2024; Graeber et al., 2025, 2026b; Jabarian and Henkel, 2025). For a discussion of broader methodological issues when analyzing such data, see Haaland et al. (2025).

[18] Given our within participant variation this is not a threat to internal validity.

(3.7 versus 3.0 on a 7-point scale; difference = 0.77, $p < 0.001$).

Panel (c) turns to the revealed measure of mental simulation obtained from the speech recordings. We plot mean speech-emotion valence separately for positive and negative signals in each format condition. The valence gap between positive and negative signals is substantially larger in the story than in the statistic condition (gap-of-gaps = 0.036, $p < 0.001$): after a false story, respondents' unscripted descriptions exhibit stronger affective alignment with the direction of the nullified signal than after a false statistic.[19] Crucially, this occurs in a task that never mentions the analyst, the recommendation, or its relevance—respondents are asked only for their overall impression of the company. The spontaneous affective coloring of their language nonetheless tracks the direction of a signal they know to be informationally void. This pattern is consistent with Manzoni et al. (2026), who show that emotionally charged information can shape policy views beyond its informational content and affect how people respond to statistical evidence.

Result 4. False stories elevate mental simulation along all three measures. Respondents report more vivid imagery after false stories than after false statistics, and the valence gap between positive and negative signals extracted from unscripted speech is roughly twice as large in the story condition.

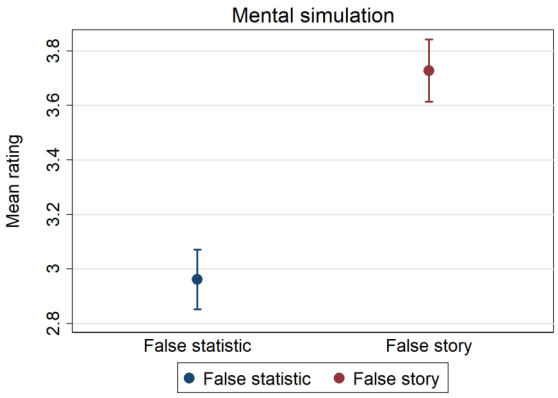
The speech-valence finding is particularly informative because it does not rely on self-report and is therefore less subject to experimenter demand effects or social desirability bias (Bursztyn et al., 2025; Haaland et al., 2023; De Quidt et al., 2018). Even a respondent who consciously discards the analyst's recommendation continues to produce signal-consistent affective content when asked to describe the company in their own words. This is consistent with stories functioning as cues that bring signal-consistent mental content to mind through similarity-based recall (Conlon and Kwon, 2026; Enke et al., 2024), and with the broader claim that qualitative elaboration engages perceptual and simulation channels that bare statistics do not (Barsalou, 1999; Bordalo et al., 2025a).

5.3 Alternative Explanations

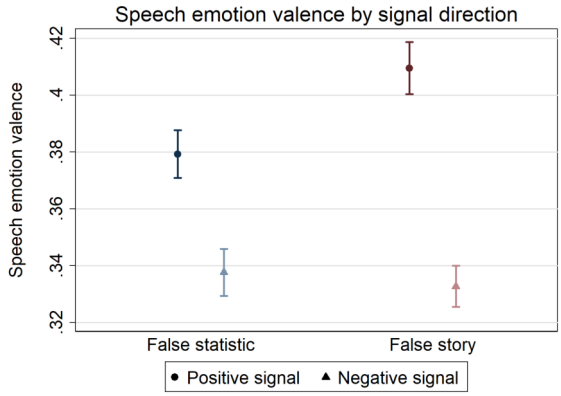
In principle, another candidate mechanism for the qualitative residue is the cognitive complexity of retraction inference (Gonçalves et al., 2025): if stories are more complex than statistics—a story being a statistic with additional content—then retracting a story is a harder inference, which on its own could generate diminished updating relative to the rational benchmark. Several pieces

[19] We replicate this finding using a transcript-based sentiment classifier applied to automatic transcripts of the same recordings. A RoBERTa model fine-tuned for sentiment classification (Camacho-Collados et al., 2022) yields a significant and quantitatively comparable interaction between signal direction and format (gap-of-gaps = 0.343, $p < 0.001$; see Appendix Figure A3). The two measures are moderately correlated ($r = 0.63$), yet both detect the same pattern, consistent with each capturing affective mental imagery through partially overlapping channels (acoustic tone vs. lexical content).

(a) Qualitative residue in the mental simulation experiment.



(b) Self-reported measure: mental imagery vividness



(c) Speech-emotion valence by signal direction

Figure 6: MECHANISM EVIDENCE: MENTAL SIMULATION EXPERIMENT. Panel (a): direction-adjusted posterior belief movement by information format; dots are sample means, bars are 95% confidence intervals. Panel (b): self-reported vividness of mental imagery of the company (1–7 scale) by information format. Panel (c): speech emotion valence extracted from respondents' open-ended audio descriptions of the company, disaggregated by signal direction within each format (circles: positive signal; triangles: negative signal). Speech valence is estimated from raw audio using a fine-tuned WavLM Large classifier (Goncalves et al., 2024). All outcomes are elicited after respondents learn the signal was irrelevant. Blue: false statistic; red: false story.

of evidence speak against this account in favor of mental simulation. First, the *Neutral stories* result (Section 5.1) holds the retraction structurally fixed and removes only evaluative valence. A complexity-based account predicts the qualitative residue should persist—the elaborated signal is at least as complex to process—but the residue disappears, consistent with simulation operating through similarity-based recall on valenced features. Second, the *No Relevance Uncertainty* experiment (Section 3.4) removes the retraction step entirely: subjects learn the signal is irrelevant before observing it, and the residue remains at 3.1 percentage points. The residue does not require the inferential complexity of revising on a retraction; it follows naturally from prior contamination occurring at exposure.[20]

5.4 Awareness and Correction

The mechanism evidence in Section 5.2 indicates that false stories engage mental simulation to a greater extent than false statistics, pulling the affective content of respondents' spontaneous descriptions toward a signal they know to be irrelevant. A question that remains unresolved within our framework concerns the mechanism by which debunked stories continue to distort prior beliefs: specifically, whether this persistence reflects an inability to perform the necessary correction (due, for example, to the original prior no longer being accessible in memory or respondents being uncertain about the prior), or whether it instead reflects the unconscious nature of mental simulation itself.

Design. We speak to this question in the experiments without interim beliefs as well as without relevance uncertainty. We pool observations from both experiments and report individual results in the appendix.[21] Recall that both studies elicit only a single posterior belief per company, after relevance has been resolved, so that a subsequent self-report refers unambiguously to one belief and one information set.[22] Stories are presented in video format. After all belief elicitations, respondents are reminded of the scenario in which they received an irrelevant story—including the content they observed and the fact that it was irrelevant—and asked whether they used this information when forming their belief. We collect 1,000 completed responses for each study, such that the pooled results are based on 2,000 completed responses.

[20]The elevated confidence we document after false stories (Section 5.5) is also directionally inconsistent with retraction-complexity, which would generally predict lower confidence after a more demanding inference. We view this as suggestive rather than dispositive: there are settings in which complexity can instead lead to higher confidence.

[21]While the pre-registrations for the individual experiments include the analyses presented in this section, we did not explicitly pre-register pooling data from both collections.

[22]This design deliberately does not measure an interim belief to avoid confusion about which belief elicitation we refer to in our question.

Specification. We estimate treatment-effect heterogeneity by self-reported use:

$$Y_i = \alpha + \beta\, T_i + \gamma S_i + \delta\,(T_i \times S_i) + u_i, \tag{5}$$

where T_i indicates the story condition and S_i indicates self-reported use of the irrelevant information. The coefficient β identifies the qualitative residue among respondents who deny being influenced; $\beta + \delta$ identifies it among respondents who report having used the information.

Results. Figure 7 compares the belief residuals in the full sample (left panel) to the subsample of participants who report not having used the irrelevant story (right panel). The residuals for both false statistics and stories are reduced among respondents who report not being influenced, but the movement for stories remains significant at 0.98 ($p = 0.026$), which is roughly a quarter of the full sample residual.[23] Appendix Table A4 additionally shows an elevated belief residual among those participants who report having been influenced by the irrelevant story, consistent with a conscious-but-uncorrectable channel: respondents recognize the story's pull but cannot fully override it. Appendix Figures A1 and A2 present the non-pooled results in the *No Interim Beliefs* and *No Relevance Uncertainty* studies, respectively.

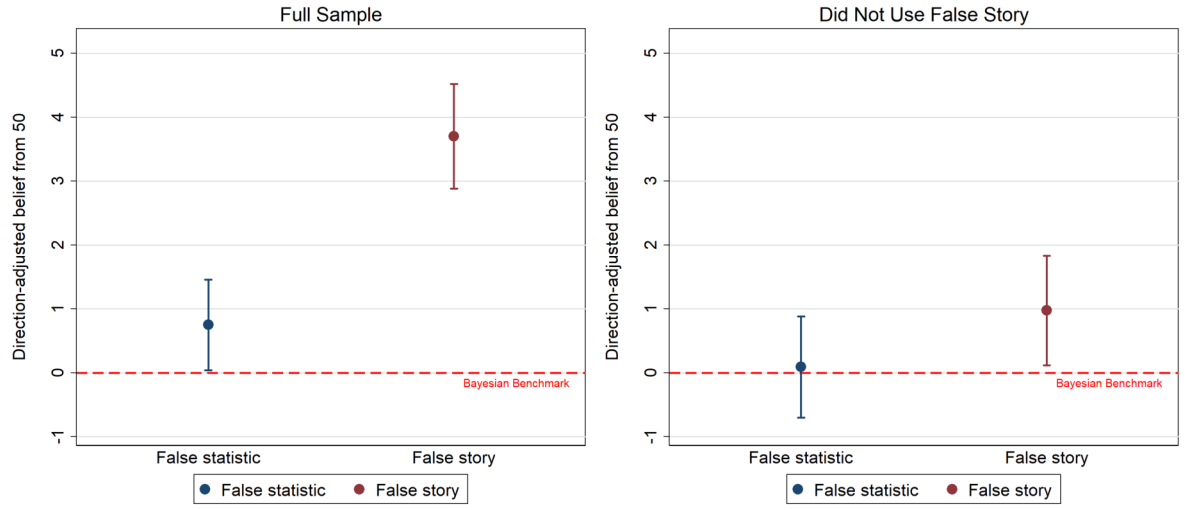


Figure 7: POSTERIOR BELIEFS AFTER IRRELEVANT SIGNALS. This figure pools observations from the *No Interim Beliefs* and *No Relevance Uncertainty* experiments. Left panel: full sample. Right panel: respondents who report not having used the irrelevant false story. Direction-adjusted posterior belief movement from the prior of 50. Dots are sample means; bars are 95% confidence intervals. Blue: false statistic; red: false story. The dashed line marks the Bayesian benchmark of zero.

Note that social desirability does not easily explain the pattern observed here. If respondents were systematically reluctant to admit influence, we would expect underreporting across the

[23] Note that the qualitative residue in this subsample (0.89) is roughly a third of the residue in the full sample (2.95), and remains marginally significant ($p = 0.096$).

board. Instead, self-reports coincide with revealed updating in the irrelevant-signal condition, supporting the validity of the measure.

Result 5. The residue of false stories operates through two channels. A dominant share of the effect is concentrated among respondents who recognize that the false story influenced their belief but cannot fully correct for it. A smaller share appears among respondents who report not being influenced, consistent with an additional automatic channel that shapes beliefs below conscious awareness.

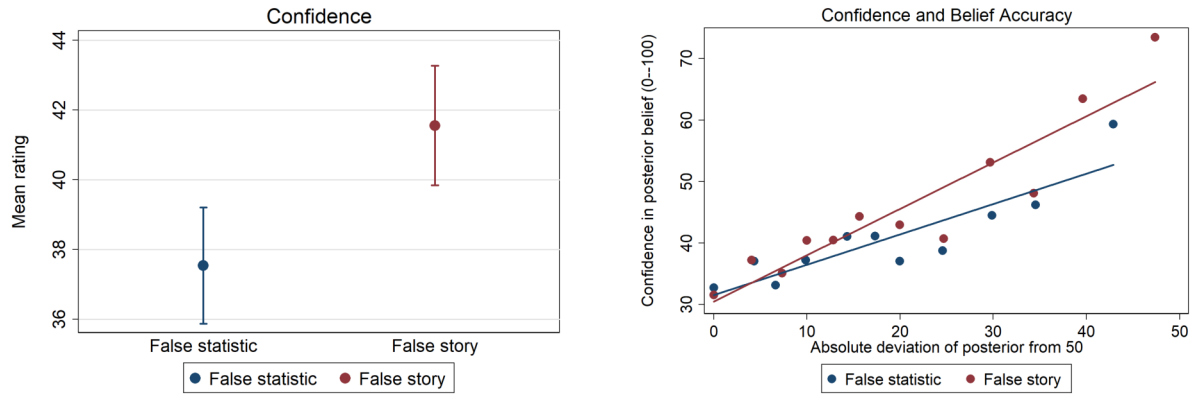
5.5 Calibration

Finally, we examine people's confidence in the optimality of their own decisions. While the framework introduced in Section 4 is agnostic about metacognition and subjective confidence, Enke et al. (2023) provide evidence that the degree to which market interactions mitigate or amplify behavioral biases depends on the correlation between the bias and the decision-maker's confidence. Specifically, biases that coincide with high decision confidence tend to have a more pronounced impact on aggregate outcomes. Against this backdrop, it becomes important to understand whether exposure to false narratives leads to higher or lower confidence relative to exposure to false statistics.

In the experiment described in Section 5.2, respondents were additionally asked to report their *confidence* in the belief they had just stated, on a 0–100 scale. This measure was elicited jointly with the mental simulation questions, immediately following the posterior belief elicitation in the two irrelevant-signal scenarios.

Figure 8a reveals that respondents report significantly higher confidence in the optimality of their beliefs following exposure to false stories than to false statistics (41.6 versus 37.6 on a 0–100 scale; difference = 4.0, $p < 0.001$). Thus, although beliefs formed after exposure to false stories are objectively less accurate, respondents nonetheless express greater confidence in their accuracy.

To quantify the extent of miscalibration between confidence and accuracy, Figure 8b plots mean confidence against the absolute deviation of the posterior belief from 50, our measure of belief error, separately for the false statistic and false story conditions. This analysis was not pre-registered and should be interpreted as exploratory. For perfectly calibrated respondents, we would expect higher confidence when the belief is close to the rational benchmark and lower confidence when it is far away, i.e., a negative slope of confidence on the belief error. Instead, we find positive slopes in both conditions, indicating miscalibration in both formats. The slope is significantly steeper for false stories (r = 0.36) than for false statistics (r = 0.20; difference $p < 0.001$). Respondents who receive a story are not merely more confident on average, their confidence is disproportionately elevated precisely when their beliefs are furthest from the rational benchmark.



(a) Mean confidence by condition

(b) Confidence vs. belief error

Figure 8: CONFIDENCE AND BELIEF ACCURACY BY INFORMATION FORMAT. Both panels are based on the *Mental Simulation* experiment, restricted to irrelevant-signal conditions. Blue: false statistic; red: false story. *Panel (a):* Mean self-reported confidence (0–100) by condition with 95% confidence intervals. *Panel (b):* Binned scatterplot of self-reported confidence against absolute posterior deviation from 50, $|\hat{\pi}^{\text{post}} - 50|$. The correlation is $r = 0.20$ ($p < 0.001$) for false statistics and $r = 0.36$ ($p < 0.001$) for false stories; the difference is significant ($z = 3.87$, $p < 0.001$).

6 Discussion and Conclusion

We document a qualitative residue of false stories: attaching a qualitative elaboration to a quantitative signal shifts incentivized beliefs in the signal's direction even when the entire signal , recommendation plus story, has been debunked. The rational benchmark is zero, yet residual belief movement in the story condition is roughly three times larger than for purely quantitative signals. The effect survives enhanced salience of the relevance revelation, removal of the interim belief elicitation, and removal of initial uncertainty about signal relevance, and disappears for non-evaluative stories. The mechanism appears to be mental simulation: false stories shift the affective valence of respondents' spontaneous speech in the direction of a signal they know to be informationally void—revealed behavior that does not rely on self-report. The distortion extends to metacognition: respondents are more confident in beliefs formed after false stories than after false quantitative information, even though those beliefs are further from the rational benchmark.

False stories in public discourse. The qualitative residue we document offers one micro-foundation for the persistence of false stories in public discourse, even after they have been authoritatively debunked. Three classes of phenomena are illustrative.

First, allegations of voter fraud in the 2020 U.S. presidential election were largely sustained by qualitative stories—claims about specific precincts, specific voting machines, specific suitcases of ballots—rather than by aggregate statistical evidence, which pointed in the opposite direction. Independent reviews, court rulings, and recounts repeatedly debunked the specific claims, yet

substantial fractions of the electorate continued to believe the election had been compromised (Eggers et al., 2021). Our findings suggest that statistical rebuttal may be a structurally weak corrective against story-based falsehoods: the channel through which the original story shifted beliefs, contamination of the prior through mental simulation, is largely insensitive to subsequent quantitative correction.

Second, stories about welfare fraud have shaped policy debates far beyond what statistical incidence would warrant. Beginning with his 1976 campaign, Ronald Reagan repeatedly told a vivid story about a Chicago "welfare queen" with multiple aliases, addresses, and Social Security cards who collected hundreds of thousands of dollars in fraudulent benefits. Investigative reporting later established that the underlying anecdote was substantially exaggerated and that systematic welfare fraud at this scale was vanishingly rare (Levin, 2019). Yet the story shaped public opinion and policy for decades—consistent with our finding that the qualitative residue of a false story is large, persists after correction, and is amplified by vivid presentation.

Third, the rise of AI-generated deepfakes intensifies this pattern. Synthetic video packages false stories in the most vivid narrative format available, and such content can proliferate faster than it can be debunked. Our finding that video elaboration amplifies the qualitative residue, together with the evidence that awareness of a story's falsity does not undo its influence, suggests that even timely correction may be insufficient to neutralize such content.

Implications for policy communication. A natural implication of our mechanism is that effective corrections may need to take the same form as the falsehoods they target. If the residue of false stories runs through mental simulation rather than through the literal informational content of the signal, then statistical rebuttals—however accurate—operate on a channel that the original falsehood largely bypassed. Counter-stories that engage simulation symmetrically, rather than dispassionate fact-checks, may therefore be a more promising corrective, consistent with evidence on prebunking and inoculation (Roozenbeek et al., 2022; van der Linden et al., 2017).

The same logic helps rationalize a pattern on the supply side of communication: strategic communicators have a clear incentive to package claims as stories rather than purely quantitative information (Thaler et al., 2025), both because narrative format is more persuasive on impact and, as our findings show, because the residual influence of a story survives retraction in a way that the influence of a comparable quantitative signal does not. Seen from this angle, the qualitative residue is not only a friction in belief updating but a structural force shaping which kinds of claims circulate in equilibrium. For policy makers and communicators of statistical information, our results suggest that pairing quantitative evidence with concrete, evocative qualitative content may be necessary not only to capture attention but also to leave a residue that survives counter-stories.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge, "When should you adjust standard errors for clustering?," *Quarterly Journal of Economics*, 2023, *138* (1), 1–35.

Allcott, Hunt and Matthew Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, 2017, *31* (2), 211–236.

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart, "Subjective Models of the Macroeconomy: Evidence from experts and Representative Samples," *Review of Economic Studies*, 2022.

_ , Ingar Haaland, Christopher Roth, Mirko Wiederholt, and Johannes Wohlfart, "Narratives about the Macroeconomy," *The Review of Economic Studies*, 2026.

Aneja, Shivangi, Chris Bregler, and Matthias Nießner, "COSMOS: Catching Out-of-Context Image Misuse Using Self-Supervised Learning," in "Proceedings of the AAAI Conference on Artificial Intelligence," Vol. 37 2023, pp. 14084–14092.

Augenblick, Ned, Eben Lazarus, and Michael Thaler, "Overinference from weak signals and underinference from strong signals," *arXiv preprint arXiv:2109.09871*, 2024.

Ba, Cuimin, J Aislinn Bohren, and Alex Imas, "Over-and underreaction to information," *Available at SSRN 4274617*, 2024.

Barron, Kai and Tilman Fries, "Narrative Persuasion: A Brief Introduction," in "Encyclopedia of Experimental Social Science," Berlin and Munich: Encyclopedia of Experimental Social Science, May 2024.

_ and _ , "Narrative persuasion," Technical Report, WZB Discussion Paper 2025.

Barsalou, Lawrence W., "Perceptual Symbol Systems," *Behavioral and Brain Sciences*, 1999, *22* (4), 577–609.

Bordalo, Pedro, Giovanni Burro, Katherine B. Coffman, Nicola Gennaioli, and Andrei Shleifer, "Imagining the Future: Memory, Simulation, and Beliefs," *The Review of Economic Studies*, 2025, *92* (3), 1532–1563.

_ , _ , Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Imagining the Future: Memory, Simulation, and Beliefs," *The Review of Economic Studies*, 2025, *92* (3), 1532–1563.

_ , _ , Nicola Gennaioli, Gad Nacamulli, and Andrei Shleifer, "Ads as Cues," Working Paper 34387, National Bureau of Economic Research 2025.

_ , **Nicola Gennaioli, Florencio Lopez de Silanes, Simon G. Schröder, Andrei Shleifer, and Maarten van Rooij**, "The Psychology of Macroeconomic Expectations," Presented at the NBER Behavioral Finance Working Group Meeting, Spring 2026 2026. April 18, 2026.

Brennen, J. Scott, Felix M. Simon, Philip N. Howard, and Rasmus Kleis Nielsen, "Types, Sources, and Claims of COVID-19 Misinformation," RISJ Factsheet, Reuters Institute for the Study of Journalism, University of Oxford 2020. Available at `https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation`.

Bursztyn, Leonardo, Aakaash Rao, Christopher Roth, and David Yanagizawa-Drott, "Opinions as Facts," *The Review of Economic Studies*, 2023, *90* (4), 1832–1864.

_ , **Ingar K. Haaland, Nicolas Roever, and Christopher Roth**, "The Social Desirability Atlas," Working Paper 33920, National Bureau of Economic Research 2025.

Camacho-Collados, Jose, Kiamehr Rezaee, Tara Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, and Eugenio Martínez-Cámara, "TweetNLP: Cutting-Edge Natural Language Processing for Social Media," in "Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations" Association for Computational Linguistics Abu Dhabi, UAE December 2022, pp. 38–45.

Chopra, Felix, Ingar Haaland, and Christopher Roth, "Do People Demand Fact-Checked News? Evidence from U.S. Democrats," *Journal of Public Economics*, 2022, *205*, 104549.

Clark, Andy, *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, New York: Oxford University Press, 2016.

Conlon, John J. and Spencer Y. Kwon, "Beliefs from Cues," 2026. Working Paper.

Conlon, John J, Malavika Mani, Gautam Rao, Matthew W Ridley, and Frank Schilbach, "Not Learning from Others," *Journal of Political Economy*, 2026.

Danz, David, Lise Vesterlund, and Alistair J. Wilson, "Belief elicitation and behavioral incentive compatibility," *American Economic Review*, 2022, *112* (9), 2851–2883.

DePaulo, Bella M., James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper, "Cues to Deception," *Psychological Bulletin*, 2003, *129* (1), 74–118.

Ecker, Ullrich K. H., Stephan Lewandowsky, Briony Swire, and Darren Chang, "Correcting False Information in Memory: Manipulating the Strength of Misinformation Encoding and Its Retraction," *Psychonomic Bulletin & Review*, 2011, *18* (3), 570–578.

Eggers, Andrew C., Haritz Garro, and Justin Grimmer, "No evidence for systematic voter fraud: A guide to statistical claims about the 2020 election," *Proceedings of the National Academy of Sciences*, 2021, *118* (45), e2103619118.

Ekman, Paul and Wallace V. Friesen, "Nonverbal Leakage and Clues to Deception," *Psychiatry*, 1969, *32* (1), 88–106.

Eliaz, Kfir and Ran Spiegler, "A model of competing narratives," *American Economic Review*, 2020, *110* (12), 3786–3816.

_ and _ , "News Media as Suppliers of Narratives (and Information)," 2024. Working paper, arXiv:2403.09155.

_ , Simone Galperti, and Ran Spiegler, "False Narratives and Political Mobilization," *Journal of the European Economic Association*, 2025, *23* (3), 983–1027.

Enke, Benjamin, "What you see is all there is," *The Quarterly Journal of Economics*, 2020, *135* (3), 1363–1398.

_ and Florian Zimmermann, "Correlation neglect in belief formation," *The Review of Economic Studies*, 2019, *86* (1), 313–332.

_ and Thomas Graeber, "Cognitive uncertainty," *The Quarterly Journal of Economics*, 2023, *138* (4), 2021–2067.

_ , Frederik Schwerter, and Florian Zimmermann, "Associative Memory, Beliefs and Market Interactions," *Journal of Financial Economics*, 2024, *157*.

_ , Thomas Graeber, and Ryan Oprea, "Confidence, Self-selection and Bias in the Aggregate," *American Economic Review*, 2023.

Esponda, Ignacio, Emanuel Vespa, and Sevgi Yuksel, "Mental Models and Learning: The Case of Base-Rate Neglect," *American Economic Review*, 2024, *114* (3), 752–782.

_ , Ryan Oprea, and Sevgi Yuksel, "Seeing What Is Representative," *The Quarterly Journal of Economics*, 2023, *138* (4), 2607–2657.

Fan, Tony Q. and Tilman Fries, "Narratives, Belief Movements, and Economic Fluctuations," 2026. Working paper.

Galasso, Vincenzo, Tommaso Nannicini, and Debora Nozza, "We Need to Talk: Audio Surveys and Information Extraction," CESifo Working Paper 11530, CESifo 2024.

Gennaioli, Nicola and Andrei Shleifer, "What comes to mind," *The Quarterly journal of economics*, 2010, *125* (4), 1399–1433.

Gneezy, Uri and Marta Serra-Garcia, "Why Don't People Lie More? Truth Is (Wrongly) Believed To Be More Persuasive," *The Economic Journal*, 2025. Forthcoming.

Goncalves, Lucas, Ali N. Salman, Abinay R. Naini, Laureano Moro Velazquez, Thomas Thebaud, Leibny Paola Garcia, Najim Dehak, Berrak Sisman, and Carlos Busso, "Odyssey 2024 - Speech Emotion Recognition Challenge: Dataset, Baseline Framework, and Results," in "The Speaker and Language Recognition Workshop (Odyssey 2024)" 2024, pp. 247–254.

Gonçalves, Duarte, Jonathan Libgober, and Jack Willis, "Retractions: Updating from Complex Information," *Review of Economic Studies*, 2025.

Graeber, Thomas, Christopher Roth, and Constantin Schesch, "Explanations," 2025. Working Paper.

_ , _ , and Florian Zimmermann, "Stories, Statistics, and Memory," *The Quarterly Journal of Economics*, 2024, *139* (4), 2181–2225.

_ , _ , and _ , "Stories and Statistics in Belief Formation," *Working Paper*, 2026.

_ , Shakked Noy, and Christopher Roth, "The Transmission of Reliable and Unreliable Information," 2026. Working Paper.

Green, Melanie C. and Timothy C. Brock, "The Role of Transportation in the Persuasiveness of Public Narratives," *Journal of Personality and Social Psychology*, 2000, *79* (5), 701–721.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart, "Designing information provision experiments," *Journal of economic literature*, 2023, *61* (1), 3–40.

_ , _ , Stefanie Stantcheva, and Johannes Wohlfart, "Understanding Economic Behavior Using Open-Ended Survey Data," *Journal of Economic Literature*, 2025, *63* (4), 1244–1280.

Hartzmark, Samuel M., Samuel D. Hirshman, and Alex Imas, "Ownership, Learning, and Beliefs," *The Quarterly Journal of Economics*, 2021, *136* (3), 1665–1717.

Henkel, Luca, Christoph Oslislo, and Frederik Schwerter, "Cues, Attention, and Charitable Giving," 2026. Working paper.

Hossain, Tanjim and Ryo Okui, "The binarized scoring rule," *Review of Economic Studies*, 2013, *80* (3), 984–1001.

Jabarian, Brian and Luca Henkel, "Voice AI in Firms: A Natural Field Experiment on Automated Job Interviews," August 2025. SSRN working paper, last revised January 26, 2026.

Johnson, Hollyn M. and Colleen M. Seifert, "Sources of the Continued Influence Effect: When Misinformation in Memory Affects Later Inferences," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1994, *20* (6), 1420–1436.

Kieren, Pascal and Martin Weber, "Expectation Formation Under Uninformative Signals," *Management Science*, 2025, *71* (6), 5123–5141.

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain, "The Science of Fake News," *Science*, 2018, *359* (6380), 1094–1096.

Levin, Josh, *The Queen: The Forgotten Life Behind an American Myth*, New York: Little, Brown and Company, 2019.

Lewandowsky, Stephan, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook, "Misinformation and Its Correction: Continued Influence and Successful Debiasing," *Psychological Science in the Public Interest*, 2012, *13* (3), 106–131.

Manzoni, Elena, Elie Murard, Simone Quercia, and Sara Tonini, "Emotions, Beliefs, and Policy Views," *Journal of the European Economic Association*, 2026, p. jvag016.

Nisbett, Richard E. and Eugene Borgida, "Attribution and the Psychology of Prediction," *Journal of Personality and Social Psychology*, 1975, *32*, 932–943.

_ **and Lee Ross**, *Human Inference: Strategies and Shortcomings of Social Judgment*, Englewood Cliffs, NJ: Prentice-Hall, 1980.

Ortoleva, Pietro, "Alternatives to Bayesian Updating," *Annual Review of Economics*, 2024, *16*, 545–570.

Peer, Eyal, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer, "Data quality of platforms and panels for online behavioral research," *Behavior Research Methods*, 2022, *54* (4), 1643–1662.

Pennycook, Gordon and David G. Rand, "Who Falls for Fake News? The Roles of Bullshit Receptivity, Overclaiming, Familiarity, and Analytic Thinking," *Journal of Personality*, 2020, *88* (2), 185–200.

_ **and** _, "Consensus, Disagreements, and Open Questions in the Psychology of Misinformation," *Annual Review of Psychology*, 2026. Forthcoming.

Quidt, Jonathan De, Johannes Haushofer, and Christopher Roth, "Measuring and bounding experimenter demand," *American Economic Review*, 2018, *108* (11), 3266–3302.

Reber, Rolf, Norbert Schwarz, and Piotr Winkielman, "Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience?," *Personality and Social Psychology Review*, 2004, *8* (4), 364–382.

Roozenbeek, Jon, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky, "Psychological Inoculation Improves Resilience against Misinformation on Social Media," *Science Advances*, 2022, *8* (34), eabo6254.

Ross, Lee, Mark R. Lepper, and Michael Hubbard, "Perseverance in Self-Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm," *Journal of Personality and Social Psychology*, 1975, *32* (5), 880–892.

Schacter, Daniel L., Donna Rose Addis, and Randy L. Buckner, "Remembering the Past to Imagine the Future: The Prospective Brain," *Nature Reviews Neuroscience*, 2007, *8* (9), 657–661.

_ , Roland G. Benoit, and Karl K. Szpunar, "Episodic Future Thinking: Mechanisms and Functions," *Current Opinion in Behavioral Sciences*, 2017, *17*, 41–50.

Scherer, Klaus R., "Vocal Communication of Emotion: A Review of Research Paradigms," *Speech Communication*, 2003, *40* (1–2), 227–256.

Schwartzstein, Joshua and Adi Sunderam, "Using models to persuade," *American Economic Review*, 2021, *111* (1), 276–323.

_ and _ , "Sharing Models to Interpret Data," Technical Report, National Bureau of Economic Research 2024.

Serra-Garcia, Marta and Uri Gneezy, "Mistakes, Overconfidence, and the Effect of Sharing on Detecting Lies," *American Economic Review*, 2021, *111* (10), 3160–3183.

Shiller, Robert J, "Narrative economics," *American Economic Review*, 2017, *107* (4), 967–1004.

Spiegler, Ran, "Bayesian networks and boundedly rational expectations," *The Quarterly Journal of Economics*, 2016, *131* (3), 1243–1290.

Szpunar, Karl K., "Episodic Future Thought: An Emerging Concept," *Perspectives on Psychological Science*, 2010, *5* (2), 142–162.

Taylor, Shelley E., Lien B. Pham, Inna D. Rivkin, and David A. Armor, "Harnessing the Imagination: Mental Simulation, Self-Regulation, and Coping," *American Psychologist*, 1998, *53* (4), 429–439.

Thaler, Michael, "The Fake News Effect: Experimentally Identifying Motivated Reasoning Using Trust in News," *American Economic Journal: Microeconomics*, 2024, *16* (2), 1–38.

_ , **Mattie Toma, and Victor Yaneng Wang**, "Numbers Tell, Words Sell," Working Paper 11600, CESifo 2025.

Tonglet, Jonathan, Gabriel Thiem, and Iryna Gurevych, "COVE: COntext and VEracity Prediction for Out-of-Context Images," in "Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)" Association for Computational Linguistics Albuquerque, New Mexico April 2025, pp. 2029–2049.

Tulving, Endel, "Episodic and Semantic Memory," in Endel Tulving and Wayne Donaldson, eds., *Organization of Memory*, New York: Academic Press, 1972, pp. 381–403.

van der Linden, Sander, *Foolproof: Why Misinformation Infects Our Minds and How to Build Immunity*, New York: W. W. Norton & Company, 2023.

_ , **Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach**, "Inoculating the Public against Misinformation about Climate Change," *Global Challenges*, 2017, *1* (2), 1600008.

For online publication only:

Learning From False Stories

Robin Musolff Christopher Roth Florian Zimmermann

The online appendix collects supplementary materials. Appendix A reports an overview of all data collections, supplementary figures and heterogeneity tables for the baseline experiment, the usage split underlying Section 5.4, secondary outcomes for interim and relevant-signal beliefs, demographics and completion rates across studies, and pre-registration links. Appendix B reproduces the full instructions of the baseline experiment and all analyst stories, together with the prompts used to generate them.

A Additional Results

Appendix Table A1: OVERVIEW OF DATA COLLECTIONS

Experiment	*N*	Modality	Design features	Pre-registration
Main experiment				
Baseline	1,500	Text & video	Evaluative stories; 750 respondents per modality	tp9xe3
Robustness experiments				
Same-screen information	750	Video	Relevance information displayed on the posterior belief screen	y5jm62
Relevance confirmation	750	Video	Respondents confirm reading the relevance information before stating their belief	yn33h5
No relevance uncertainty	1,000	Video	Single posterior belief; no uncertainty about signal relevance; self-reported usage of irrelevant story	c2n5ru
No interim beliefs	1,000	Video	Single posterior belief; self-reported usage of irrelevant story	pb57th
Mechanism experiments				
Neutral stories	750	Text	Non-evaluative, descriptive stories; qualitative content held fixed	5tw8x7
Mental simulation	1,000	Video	Adds open-ended speech recordings, imagery vividness, and confidence measures	u22f56

Notes: All experiments (except *No relevance uncertainty*) share the same two-stage signal design: respondents observe an analyst recommendation under genuine uncertainty about its relevance, and then learn with certainty whether the signal was relevant or irrelevant. The story condition pairs the recommendation with a qualitative elaboration matched to its direction. Each data collection was separately pre-registered on AsPredicted; the final column reports the identifier of each record. Identifiers link to the corresponding pre-registration document, e.g. the Baseline pre-registration is available at `https://aspredicted.org/tp9xe3.pdf`.

A.1 Baseline Experiment: Correct Comprehension Check on First Attempt

Appendix Table A2: Main analysis: first-attempt passers subsample

Dependent variable	Posterior belief movement			
	Pooled	Video	Text	Interaction
	(1)	(2)	(3)	(4)
Story	2.729***	3.142***	2.316***	2.316***
	(0.549)	(0.824)	(0.728)	(0.728)
Video				0.479
				(0.895)
Video × Story				0.826
				(1.099)
Baseline	1.302***	1.542**	1.063*	1.063*
	(0.447)	(0.640)	(0.626)	(0.626)
N	2,542	1,270	1,272	2,542

Notes: This table uses data from the Baseline Experiment. Replicates Table 1 restricting to the 84.7% of participants ($N = 1{,}271$) who answered all comprehension questions correctly on the first attempt. OLS regressions. Dependent variable is the direction-adjusted posterior belief from 50; sample restricted to irrelevant conditions. Baseline is the false statistic condition. Standard errors clustered at the participant level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A.2 Baseline Experiment: Bayesian Answer for False Statistic

Appendix Table A3: Story effect by whether participants returned to the prior in the statistic condition

Dependent variable	Posterior belief movement		
	Stat $= 50$	Stat $\neq 50$	Full sample
	(1)	(2)	(3)
Story	2.078^{***}	3.401^{***}	2.949^{***}
	(0.541)	(0.730)	(0.515)
Baseline	0.000	1.928^{***}	1.269^{***}
	(0.000)	(0.639)	(0.421)
N	1,026	1,974	3,000
p-value: Stat $= 50$ vs. Full sample		0.145	

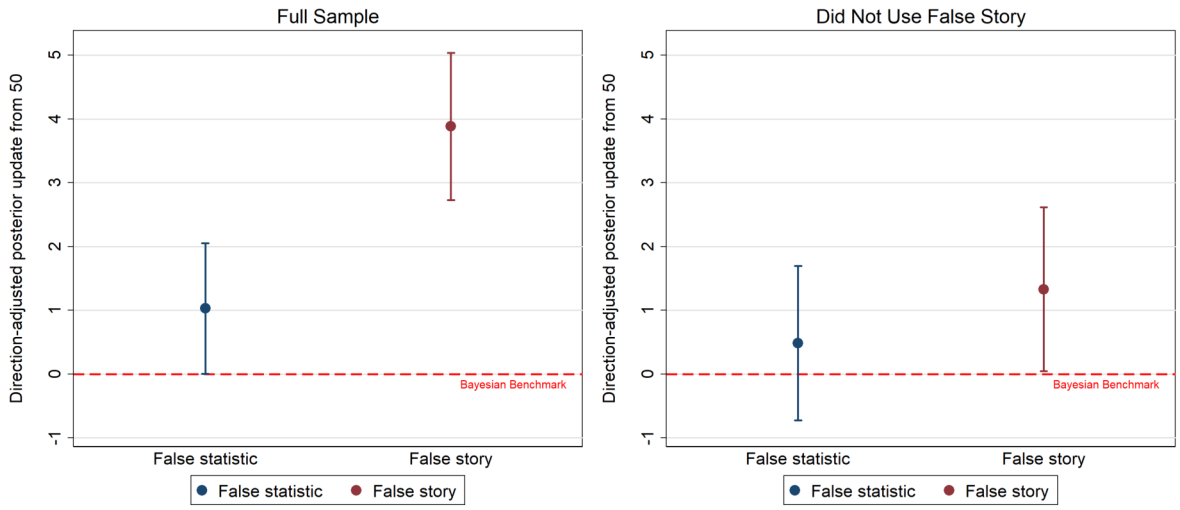
Notes: This table uses data from the Baseline Experiment. OLS regressions of the direction-adjusted posterior belief from 50 on the false-story indicator, restricted to irrelevant conditions. Column (1) restricts to participants who reported exactly 50 in the false statistic condition; column (2) restricts to the remaining participants; column (3) uses the full sample. The constant in column (1) is zero by construction. The *p*-value at the bottom tests whether the qualitative residue in the Stat $= 50$ subsample differs from the full-sample qualitative residue, computed from a full-sample interaction regression that interacts the false-story indicator with an indicator for the Stat $= 50$ subsample. Standard errors clustered at the participant level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A.3 Usage Split

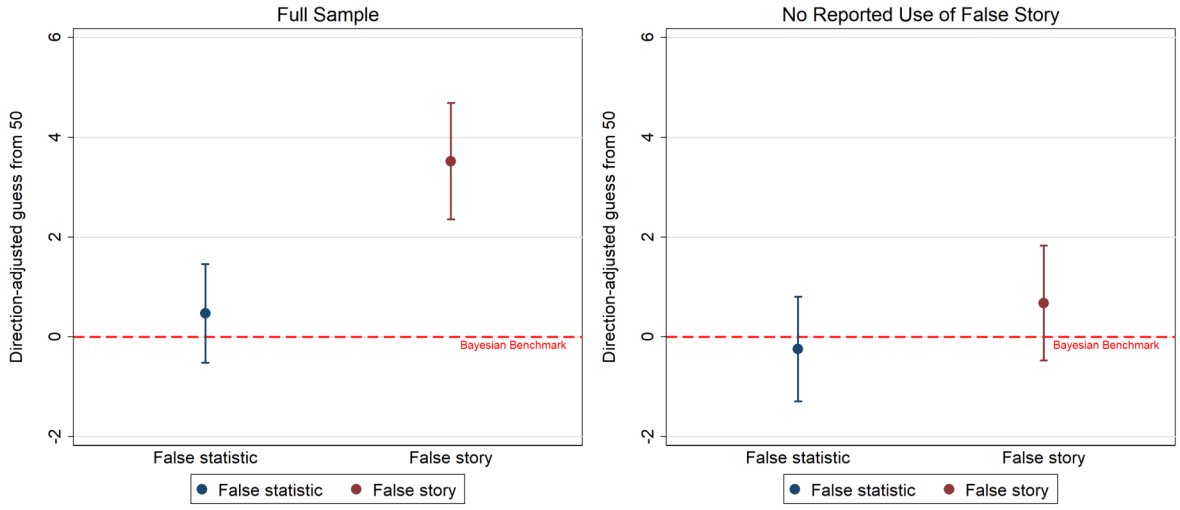
Appendix Table A4: Posterior belief movement by self-reported usage of false story

	No Relevance Uncertainty		No Interim Beliefs		Pooled	
Used false story	No	Yes	No	Yes	No	Yes
False statistic	–0.24	3.03^{**}	0.48	2.10^{**}	0.09	2.46^{***}
	(0.54)	(1.28)	(0.62)	(0.96)	(0.41)	(0.77)
False story	0.68	13.72^{***}	1.33^{**}	8.92^{***}	0.98^{**}	10.80^{***}
	(0.59)	(1.54)	(0.66)	(1.13)	(0.44)	(0.92)
N	782	218	663	337	1445	555

Notes: Each cell reports the mean direction-adjusted posterior belief movement from 50 (in points on the 0–100 belief scale). The left panel reports results from the *No Relevance Uncertainty* experiment, the middle panel from the *No Interim Beliefs* experiment, and the right panel pools both experiments. Standard errors clustered at the participant level in parentheses. Stars denote significance of a test against the null of zero movement: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.



Appendix Figure A1: POSTERIOR BELIEFS AFTER IRRELEVANT SIGNALS. This figure is based on the *No Interim Beliefs* experiment. Left panel: full sample. Right panel: respondents who report not having used the irrelevant false story. Direction-adjusted posterior belief movement from the prior of 50. Dots are sample means; bars are 95% confidence intervals. Blue: false statistic; red: false story. The dashed line marks the Bayesian benchmark of zero.



Appendix Figure A2: POSTERIOR BELIEFS AFTER IRRELEVANT SIGNALS. This figure is based on the *No Relevance Uncertainty* experiment. Left panel: full sample. Right panel: respondents who report not having used the irrelevant false story. Direction-adjusted posterior belief movement from the prior of 50. Dots are sample means; bars are 95% confidence intervals. Blue: false statistic; red: false story. The dashed line marks the Bayesian benchmark of zero.

A.4 Results for Uncertain Relevance and Relevant Information

Appendix Table A5: Secondary outcomes: interim and relevant-signal posterior beliefs

Dependent variable	(1) Interim belief movement	(2) Posterior belief movement
Story	8.797***	8.750***
	(0.393)	(0.501)
Constant	9.260***	16.729***
	(0.312)	(0.395)
N	6,000	3,000

Notes: This table uses data from the Baseline Experiment. OLS regressions. Column (1): Dependent variable is the direction-adjusted interim belief (stated before relevance is revealed); sample includes all four conditions; baseline is the statistic format. Column (2): Dependent variable is the direction-adjusted posterior belief; sample restricted to relevant conditions (statistic and story); baseline is the statistic condition. Standard errors clustered at the participant level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A.5 Attrition

Table A6 reports participant characteristics for the analysis sample of each experiment. Demographics are broadly stable across studies, consistent with uniform Prolific recruitment criteria. Samples are majority female (51–61%), with a mean age in the mid-to-late thirties, and are relatively well-educated, with roughly 60% holding at least a bachelor's degree.

Appendix Table A6: Participant demographics across experiments

	Baseline	Same Screen	Rel. Confirm.	No Rel. Unc.	Neutral Stories	Mental Sim.	No Int. Beliefs
N	1,500	750	750	1,000	750	1,000	1,000
Age							
Mean (SD)	41.2 (13.8)	39.4 (13.0)	39.7 (12.7)	39.3 (12.8)	40.8 (13.3)	36.3 (12.3)	38.9 (13.0)
Sex (%)							
Male	46.1	43.6	38.1	44.6	40.8	41.3	47.5
Female	53.1	54.3	60.7	54.3	57.9	57.3	51.5
Other	0.9	2.1	1.2	1.1	1.3	1.4	1.0
Education (%)							
Less than high school	0.6	0.7	0.4	0.3	0.7	0.9	1.4
High school graduate	11.7	11.7	8.9	12.1	9.3	9.4	11.8
Some college, no degree	18.5	18.9	16.4	17.0	18.4	18.4	15.9
Associate degree	9.9	9.3	7.3	10.8	9.2	8.3	6.6
Bachelor's degree	39.0	40.7	40.7	39.9	40.3	40.9	40.5
Master's degree	15.7	16.1	19.9	15.3	16.3	17.4	19.3
Professional degree (JD, MD)	2.1	1.1	2.8	1.9	3.2	2.4	1.4
Doctoral degree	2.5	1.5	3.6	2.7	2.7	2.3	3.1
Employment (%)							
Working (paid employee)	59.3	58.0	67.6	63.8	61.5	62.3	65.3
Working (self-employed)	16.6	14.8	11.6	16.3	13.7	18.9	13.9
Not working	21.6	24.1	18.9	17.4	22.5	16.1	17.3
Prefer not to answer	2.5	3.1	1.9	2.5	2.3	2.7	3.5

Notes: Demographics are reported for the analysis sample of each experiment. Percentages may not sum to 100 due to rounding or missing values.

Table A7 reports recruitment and completion rates for each of the seven data collections. Across studies, completion rates range from 54% to 67%, reflecting differences in study design and comprehension check difficulty. Non-completion reflects failure of pre-registered attention or comprehension checks administered before participants encounter any treatment conditions, and cannot generate differential attrition across the story-statistic comparison since that comparison is within-subject. In the Baseline experiment, where presentation modality is randomized between subjects, completion rates do not differ across modalities ($\chi^2(1) = 1.23$, $p = 0.268$). Attrition therefore does not threaten internal validity.

Appendix Table A7: Completion rates across experiments

Experiment	Started	Completed	Analysis sample	Completion rate (%)
Main experiment				
Baseline	2,625	1,503	1,500	57.3
Robustness experiments				
Same Screen Information	1,318	769	750	58.3
Relevance Confirmation	1,405	751	750	53.5
No Relevance Uncertainty	1,516	1,016	1,000	67.0
Mechanism experiments				
Neutral Stories	1,341	750	750	55.9
Mental Simulation	1,864	1,004	1,000	53.9
No Interim Beliefs	1,592	1,002	1,000	62.9

Notes: Started refers to all participants who opened the survey. Completed refers to participants who finished the full session; non-completion reflects failure of pre-registered attention or comprehension checks, which are administered before any treatment conditions are encountered. The analysis sample is the pre-registered target *N* for each study; surplus completers are excluded. Since all key treatment variation is within-subject, non-completion cannot generate differential attrition across conditions. Note the completion rates observed here are in line with those of other papers featuring studies conducted on Prolific. E.g. Enke and Graeber (2023) screen out slightly more than 50% of participants based on attention and comprehension checks.

A.6 Heterogeneity Tables

The tables in this subsection support the heterogeneity analyses of Section 3.3.

Heterogeneity by demographics. Table A8 reports the qualitative residue separately by gender, age and education.

Heterogeneity by order of presentation. Table A9 reports the qualitative residue separately for each scenario position. It is positive and significant in all four positions; a joint test of equality does not reject ($p = 0.542$), indicating no carryover, learning, or fatigue effects. Table A10 further splits the sample by whether the false story scenario was encountered before or after the

Appendix Table A8: Heterogeneity by demographics: effect of false story on posterior belief movement

Dependent variable	Posterior belief movement					
	Gender		Age		Education	
	Male	Female	Below median	Above median	College	No college
Story	3.087***	2.831***	2.356***	3.520***	2.440***	3.689***
	(0.748)	(0.710)	(0.702)	(0.751)	(0.615)	(0.892)
Baseline	1.593**	0.991*	1.826***	0.732	1.349***	1.152
	(0.617)	(0.576)	(0.576)	(0.612)	(0.522)	(0.702)
N	1382	1618	1472	1528	1778	1222

Notes: This table uses data from the Baseline Experiment. Each column reports OLS estimates from a regression of direction-adjusted posterior belief movement on an indicator for the false story condition and a constant; baseline is the false statistic condition. Sample restricted to irrelevant conditions (false statistic and false story). Age split at the sample median (39 years). College includes Bachelor's, Master's, doctoral, and professional degrees. Standard errors clustered at the participant level in parentheses. Joint test of equality of the false story coefficient across all six subsamples: $F = 0.909$, $p = 0.436$. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

false statistic scenario within the same participant. The residue is virtually identical across the two groups ($p = 0.818$), ruling out carryover effects between the two irrelevant conditions.

Appendix Table A9: Stability of qualitative residue across rounds

Dependent variable	Posterior belief movement			
	1st	2nd	3rd	4th
	(1)	(2)	(3)	(4)
Story	3.591***	3.068**	1.331	3.850***
	(1.221)	(1.315)	(1.238)	(1.285)
Baseline	0.481	1.256	2.184**	1.153
	(0.850)	(0.782)	(0.857)	(0.887)
N	780	741	760	719

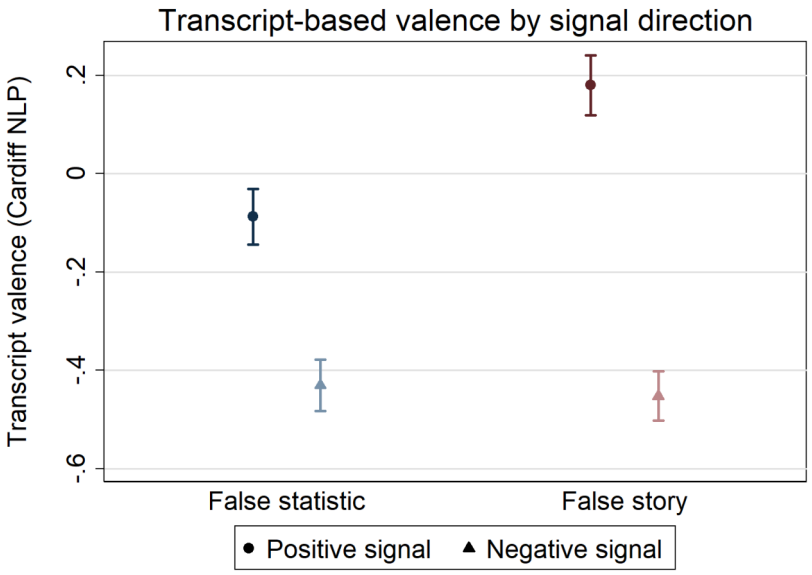
Notes: This table uses data from the Baseline Experiment. Each column reports OLS estimates from a regression of direction-adjusted posterior belief movement on an indicator for the false story condition and a constant, restricted to observations in the indicated round. Baseline is the false statistic condition. Standard errors clustered at the participant level in parentheses. Joint test of equality of the false story coefficient across all four positions: $F = 0.714$, $p = 0.544$. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Appendix Table A10: Stability of qualitative residue by within-participant scenario order

Dependent variable	Posterior belief movement	
	Story seen first (1)	Stat seen first (2)
Story	2.829***	3.066***
	(0.742)	(0.715)
Baseline	1.325**	1.213**
	(0.609)	(0.583)
N	1482	1518

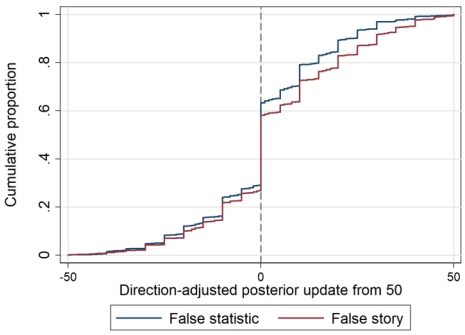
Notes: This table uses data from the Baseline Experiment. Each column reports OLS estimates from a regression of direction-adjusted posterior belief movement on an indicator for the false story condition and a constant. Column (1) restricts to participants who encountered the false story scenario before the false statistic scenario; column (2) restricts to those who encountered the false statistic first. Baseline is the false statistic condition. Standard errors clustered at the participant level in parentheses. Interaction test (difference in false story coefficient across columns): $p = 0.818$. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A.7 Transcript-based Valence Coding

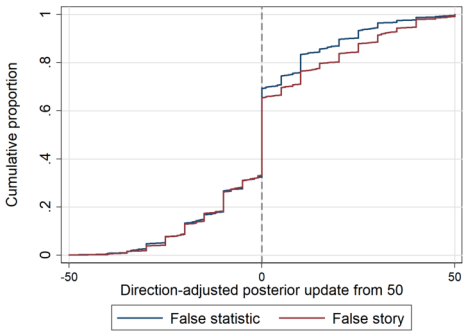


Appendix Figure A3: TRANSCRIPT-BASED SENTIMENT VALENCE BY SIGNAL DIRECTION AND INFORMATION FORMAT. This figure is based on the *Mental Simulation* experiment. The vertical axis plots the mean sentiment valence of participants' unscripted verbal descriptions, scored using a RoBERTa model fine-tuned for sentiment classification (Camacho-Collados et al., 2022). Valence is computed as $P(\text{positive}) - P(\text{negative}) \in [-1, 1]$. Filled circles denote positive-signal observations; triangles denote negative-signal observations. Lighter shading indicates negative-signal observations within each condition. Data are restricted to irrelevant-signal conditions (false statistic and false story). The gap in transcript valence between positive- and negative-signal observations is substantially larger in the false story condition than in the false statistic condition (gap-of-gaps $= 0.343$, $p < 0.001$), replicating the pattern obtained from the audio-based Odyssey classifier (Figure 6) using lexical content alone.

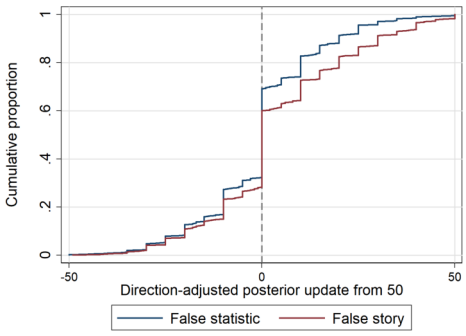
A.8 Distributions of Posterior Belief Update



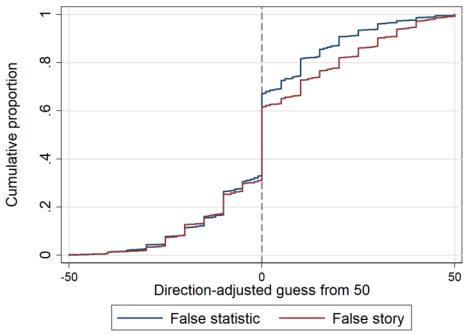
(a) Baseline



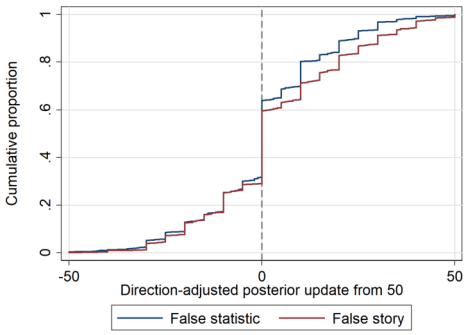
(b) Same-screen information



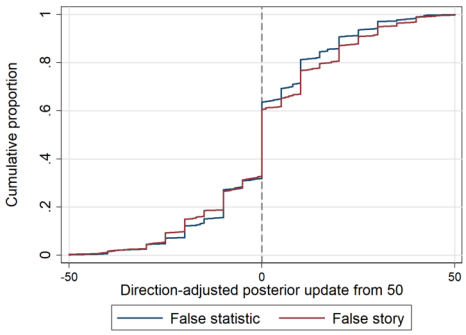
(c) Relevance confirmation



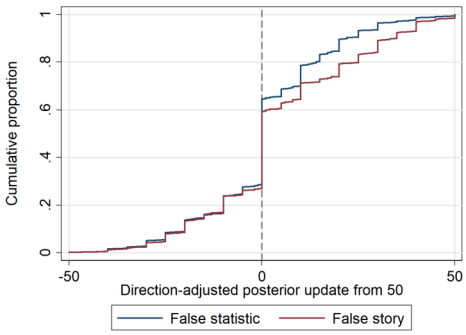
(d) No relevance uncertainty



(e) No interim beliefs



(f) Neutral stories



(g) Mental simulation

Appendix Figure A4: DISTRIBUTION OF POSTERIOR BELIEF MOVEMENT ACROSS EXPERIMENTS. Each panel shows the cumulative distribution of direction-adjusted posterior belief movement from the prior of 50, for the false statistic (blue) and false story (red) conditions.

A.9 Pre-Registration and Deviations

Each data collection in Table A1 was separately pre-registered on AsPredicted. All main-text analyses follow the pre-registered specifications, except for minor deviations documented as follows.

- The heterogeneity analyses on demographics and order of presentation presented in Sections 3.3 and A.6 were not pre-registered.
- In Section 5.4, we present a pooled analysis in addition to the pre-registered analyses in the *No Relevance Uncertainty* and *No Interim Beliefs* collections.
- In Section 5.5, the confidence elicitation was pre-registered as part of the *Mental Simulation* experiment. The analysis of the relationship between confidence and belief error ($|\hat{\pi}^{\text{post}} - 50|$) presented in Figure 8b was not pre-registered.

B Experimental Instructions

In the following, we reproduce the full instructions for the text-version of the baseline study. All stories and links to the videos can be found in the second subsection.

B.1 Instructions of the Baseline Study

Welcome Screen

Welcome!

This study is designed for computer (PC or Mac) users only (desktop, laptop, etc.). If you are accessing this study on a smartphone, a tablet or any other non-PC devices, please switch to PC and enter the study again, or return the submission on Prolific.

The study uses audio, so please make sure that you have either headphones or speakers switched on.

Please write at least 15 words describing your opinion about daylight savings in the United States. Whether you are in favor or against daylight savings does not affect your eligibility to participate in this study. However, we ask that you write at least 15 words on your thoughts about this topic.

Attention Check

Pay close attention now! One quick check to make sure you're ready.

On the next page some digits will appear in quick succession. Your task will be to enter the digits in a textbox.

[Screen: Enter the numbers, in the order shown, below.]

Study Overview

Welcome!

Thank you for participating in this study. This study will take approximately 12 minutes to complete.

To earn your reward, you have to read all instructions carefully and correctly answer the comprehension questions.

The study consists of eight payoff-relevant tasks in total.

You have a chance to win an additional bonus of up to $30 if you complete the study. One out of ten participants will be eligible for a bonus. In case you are eligible for a bonus, one of your

eight answers will be randomly selected by the computer and will determine your bonus.

Instructions

In this study you will make a number of guesses. The better your guesses, the higher the likelihood that you will win a bonus.

Questions will always be about analyst evaluations for one of four different companies. All questions, the companies and the analysts, are constructed for the purpose of this study only and are unrelated to any real-life companies or analysts. For each company, we will ask you to complete two tasks.

Companies can be either good or bad. A company is good if 50% or more of its evaluations say it is good, and is bad if less than 50% of its evaluations say it is good.

Company Guesses. For each company, we will ask you twice about the fraction of evaluations that a company has received by financial analysts indicating that the company is good (between 0 and 100%). Here is an example of how such a question might look like:

> A company has received evaluations by 16 financial analysts. What do you think is the fraction of evaluations indicating that the company is good?

For each of the four companies, the actual fraction of evaluations indicating that the company is good is randomly drawn from a uniform distribution by the computer.

For each company, you will receive additional information from one randomly drawn analyst about whether the company is good or bad. This will help you make a more informed guess about the fraction of evaluations indicating that the company is good. For each company, there is a possibility that you will also receive an explanation for the analyst's evaluation of whether the company is good or bad.

Note: Sometimes this additional information concerns the company relevant for your guess. Sometimes, by random assignment, this additional information is about another company. If this additional information is about another company, then it does not contain any useful information for your guess.

For each company, you will provide two guesses. For your first guess, you will always face uncertainty, i.e. with 50% the additional information concerns the company relevant for your guess and with 50% it is about another company that is irrelevant for your guess.

Before your second guess, uncertainty will be resolved and you will be informed whether the additional information concerns the company relevant for your guess, or whether it is about another company.

Bonus payment. We will randomly select one of your eight company guesses to determine your bonus payment.

The bonus payment is determined by the accuracy of your guess: The closer your guess is to the true fraction of evaluations indicating that the company is good, the higher the likelihood that you receive the bonus of $30.

If you click on the below triangle, the precise formula will be displayed. While this formula might seem complicated, the underlying principle is very simple: the smaller the difference between your guess and the truth, the higher the likelihood that you win the $30. It is hence in your best interest to simply state your best guess.

[Click on the triangle to see the formula.]

Comprehension Questions

You have to answer all comprehension questions correctly in order to receive your completion bonus and keep your chance of receiving the $30 bonus.

1. How many companies will be presented to you in this survey?
    - 2
    - 4
    - 15
    - 20
2. Which one of the following statements is true?
    - For each company, the computer will randomly draw the evaluations to indicate that the company is good or bad. The evaluations for a given company are determined independently of all other companies.
    - The computer will determine the correct answer to all companies by randomly drawing evaluations once. This means that all companies have the same evaluations.
3. Which one of the following statements is true?
    - The guesses I make in this study might affect my payoffs. The study involves real stakes.
    - The guesses I make in this study will not affect my payoffs. The study is purely hypothetical.
4. Is the following statement correct or wrong? “I will receive additional information from one analyst about whether a company is good or bad. I should take this additional information

into account when making my guesses if the additional information is about the relevant company for my guess."

- Correct.
- Wrong.

5. Is the following statement correct or wrong? "I will receive additional information from one analyst about whether a company is good or bad. I should take this additional information into account when making my guesses if the additional information is about another company."

   - Correct.
   - Wrong.

6. Is the following statement correct or wrong? "For some guesses it is uncertain (50-50) whether the additional information concerns the company relevant for your guess or whether it is about another company that is irrelevant for your guess."

   - Correct.
   - Wrong.

Audio Attention Check

Please enter the four numbers from the audio file without spaces inbetween the single digits (e.g. 1234).

[Audio file plays a sequence of four digits.]

Timeline

You will now be presented with the companies.

The timeline for each scenario is as follows:

1. You face uncertainty, i.e. with 50% the additional information concerns the company relevant for your guess and with 50% it is about another company that is irrelevant for your guess.
2. We will inform you about the company and the evaluation of the randomly drawn analyst.
3. We will then ask you what you think is the fraction of analyst evaluations indicating that the company is good.

4. You will then learn whether the additional information concerns the company relevant for your guess or whether it is about another company that is irrelevant for your guess.
5. We will then ask you again what you think is the fraction of analyst evaluations indicating that the company is good.

Please pay careful attention to the information.

Example Scenario: Statistic only

Uncertainty Reminder. Note: The next additional information you will receive may or may not be relevant for your guess. There is a 50% chance that it concerns the company relevant for your guess, while there is a 50% chance that it is about another company that is irrelevant for your guess.

Information Screen. A food company has received evaluations from 16 analysts.

> Additional information:
>
> A randomly drawn analyst thinks the company is good.

Company Guess. What do you think is the **total fraction of evaluations indicating that the company is good (in percent)** for the food company?

_____%

Relevance Resolution. *[Subjects then see one of the following two screens, depending on randomization:]*

If relevant: Note: The additional information you received concerns the company relevant for your guess. On the next screen you will be able to revise your guess.

If irrelevant: Note: The additional information you received concerns another company not relevant for your guess. On the next screen you will be able to revise your guess.

Revised Guess Screen. A food company has received evaluations from 16 analysts.

> Reminder: you received the following additional information:
>
> A randomly drawn analyst thinks the company is good.

Revised Company Guess

What do you think is the **total fraction of evaluations indicating that the company is good (in percent)** for the food company?

_____%

Example Scenario: Statistic + story

Uncertainty Reminder. Note: The next additional information you will receive may or may not be relevant for your guess. There is a 50% chance that it concerns the company relevant for your guess, while there is a 50% chance that it is about another company that is irrelevant for your guess.

Information Screen. A construction company has received evaluations from 14 analysts.

> Additional information:
>
> A randomly drawn analyst thinks the company is good and said the following about the company:
>
> "My evaluation of this company is positive. The internal workings of this company reflect a level of operational excellence that is genuinely uncommon. Processes are efficient, waste is minimized, and execution is consistently reliable. The organization demonstrates an ability to deliver on its commitments with a precision that builds credibility with customers and partners alike. Cost discipline is embedded in the culture without sacrificing quality, which means margins hold up even under pressure. When a company runs this well from the inside, it creates a compounding advantage that is hard to quantify but unmistakable in outcomes. The fundamentals here are sound and resilient."

Company Guess. What do you think is the **total fraction of evaluations indicating that the company is good (in percent)** for the construction company?

_____%

Relevance Resolution. *[Subjects then see one of the following two screens, depending on randomization:]*

If relevant: Note: The additional information you received concerns the company relevant for your guess. On the next screen you will be able to revise your guess.

If irrelevant: Note: The additional information you received concerns another company not relevant for your guess. On the next screen you will be able to revise your guess.

Revised Guess Screen. A construction company has received evaluations from 14 analysts.

> Reminder: you received the following additional information:
>
> A randomly drawn analyst thinks the company is good and said the following about the company:
>
> "My evaluation of this company is positive. The internal workings of this company reflect a level of operational excellence that is genuinely uncommon. Processes are efficient, waste is minimized, and execution is consistently reliable. The organization demonstrates an ability to deliver on its commitments with a precision that builds credibility with customers and partners alike. Cost discipline is embedded in the culture without sacrificing quality, which means margins hold up even under pressure. When a company runs this well from the inside, it creates a compounding advantage that is hard to quantify but unmistakable in outcomes. The fundamentals here are sound and resilient."

Revised Company Guess

What do you think is the **total fraction of evaluations indicating that the company is good (in percent)** for the construction company?

_____%

Final Questionnaire

To complete, please fill out the following questionnaire.

- Your sex:
  - Male
  - Female
  - Other
- Your age: _____
- What is the highest level of school you have completed or the highest degree you have received?
  - Less than high school degree
  - High school graduate (high school diploma or equivalent including GED)
  - Some college but no degree
  - Associate degree in college (2-year)

- Bachelor's degree in college (4-year)
- Master's degree
- Doctoral degree
- Professional degree (JD, MD)

- Which statement best describes your current employment status?
  - Working (paid employee)
  - Working (self-employed)
  - Prefer not to answer
  - Not working

End Screen

The study is now completed. Thank you for your participation!

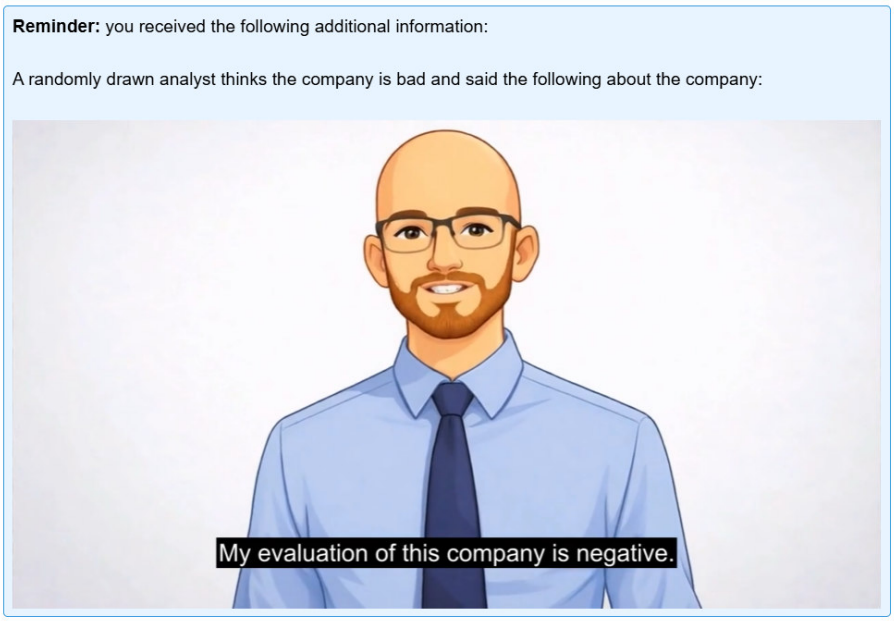
If you have any comments on the survey, please let us know.

B.2 Stories and Links to Videos

In our experiment, participants receive analyst stories that vary along two dimensions: valence (positive or negative) and vividness (baseline or neutral). Baseline stories make a persuasive case for or against the company, while neutral stories contain only factual, non-evaluative descriptions of the company's operations. All stories were generated using Claude Sonnet 4.6 (Anthropic) and presented to participants either as text or videos featuring animated analyst avatars created using HeyGen. Below we report the prompts used and the resulting stories. Videos for all stories as presented to participants are available at `https://doi.org/10.6084/m9.figshare.32114773`.

Positive Baseline Stories

Prompt. *Write 4 distinct analyst narratives, each approximately 90 words, from the first-person perspective of a financial analyst giving a positive evaluation of a hypothetical company. Each narrative should begin with the sentence "My evaluation of this company is positive." and then make a persuasive case that the company is of high quality and a strong investment. The 4 versions should each emphasize a different dimension of the company's quality — for example, leadership, competitive position, operational strength, and growth potential — so that they are stylistically and substantively distinct from one another. Keep the language generic enough to apply to any type of company, regardless of sector. Do not mention any specific industry, product,*



Appendix Figure A5: EXAMPLE FOR THE VIDEO MODALITY. The videos featured the analyst saying 'My evaluation of this company is [positive/negative]' for statistics, and the additional qualitative elaboration in the case of stories.

or company name. Do not include any numbers or statistics. Write only the narrative texts, numbered 1 to 4, with no preamble or labels beyond the numbering.

Technology Company. "My evaluation of this company is positive. What sets this company apart is the exceptional caliber of its leadership team. The executives bring a rare combination of vision and operational discipline, consistently making decisions that prioritize long-term value over short-term gains. Their track record of navigating challenging environments with composure and clarity inspires confidence across the organization. A strong culture flows from the top, attracting and retaining talented people at every level. When leadership is this capable and aligned, the entire enterprise benefits. This is a management team that earns trust, and in my assessment, that trust is well placed."

Food Company. "My evaluation of this company is positive. This company occupies a uniquely defensible position in its market. It has built meaningful advantages that are difficult for competitors to replicate, including strong brand recognition, deep customer loyalty, and established relationships that took years to develop. New entrants face significant barriers, and existing rivals have repeatedly struggled to erode this company's standing. Rather than competing on price alone, it competes on quality and reputation — a far more durable foundation. Companies that hold this kind of structural advantage tend to sustain superior performance over time, and this one is no exception."

Construction Company. "My evaluation of this company is positive. The internal workings of this company reflect a level of operational excellence that is genuinely uncommon. Processes are efficient, waste is minimized, and execution is consistently reliable. The organization demonstrates an ability to deliver on its commitments with a precision that builds credibility with customers and partners alike. Cost discipline is embedded in the culture without sacrificing quality, which means margins hold up even under pressure. When a company runs this well from the inside, it creates a compounding advantage that is hard to quantify but unmistakable in outcomes. The fundamentals here are sound and resilient."

Energy Company. "My evaluation of this company is positive. Perhaps the most compelling aspect of this company is the breadth and credibility of its growth outlook. It is well positioned to expand into adjacent opportunities, and the groundwork for that expansion has already been thoughtfully laid. Demand for what it offers is broadening, and the company has demonstrated the agility to scale without sacrificing quality or consistency. There is a clear pathway forward, supported by a capable team and a strong existing base. Companies at this stage of their trajectory, with this degree of readiness, represent precisely the kind of forward-looking opportunity worth serious attention."

Negative Baseline Stories

Prompt. *Write 4 distinct analyst narratives, each approximately 90 words, from the first-person perspective of a financial analyst giving a negative evaluation of a hypothetical company. Each narrative should begin with the sentence "My evaluation of this company is negative." and then make a persuasive case that the company is of low quality and a poor investment. The 4 versions should each emphasize a different dimension of the company's weakness — for example, leadership, competitive position, operational strength, and growth potential — so that they are stylistically and substantively distinct from one another. Keep the language generic enough to apply to any type of company, regardless of sector. Do not mention any specific industry, product, or company name. Do not include any numbers or statistics. Write only the narrative texts, numbered 1 to 4, with no preamble or labels beyond the numbering.*

Technology Company. "My evaluation of this company is negative. The leadership team has repeatedly demonstrated poor judgment, prioritizing short-term appearances over long-term value creation. Strategic decisions feel reactive rather than deliberate, and management has failed to inspire confidence among employees or external stakeholders. High turnover in senior roles has created instability and a lack of institutional continuity. Without steady, visionary leadership at the helm, the organization lacks the direction necessary to navigate challenges or capitalize on opportunities. Until meaningful change occurs at the top, this company presents an unacceptable risk for any serious investor."

Food Company. "My evaluation of this company is negative. The company operates in an increasingly crowded space where rivals consistently outperform it on quality, efficiency, and customer loyalty. It has failed to carve out any meaningful differentiation, leaving it vulnerable to being undercut or displaced. Competitors with stronger brand recognition and superior execution continue to erode its market share, and there is no credible strategy in place to reverse this trajectory. A business without a defensible competitive position is one perpetually fighting for survival. This company shows no signs of reclaiming relevance in a landscape that has largely moved past it."

Construction Company. "My evaluation of this company is negative. Internally, the company is burdened by inefficiencies that permeate every layer of its operations. Processes that should be streamlined remain fragmented and inconsistent, leading to chronic underperformance relative to peers. Supply chains are poorly managed, execution is unreliable, and the organization has shown a persistent inability to deliver on its own commitments. These are not temporary growing pains but deeply entrenched structural problems. Operational weakness of this kind is difficult and costly to fix, and without dramatic internal reform, the company will continue to destroy value rather than create it."

Energy Company. "My evaluation of this company is negative. The company's prospects for future growth are deeply unconvincing. Its core business is mature and showing clear signs of stagnation, while its attempts to expand into new areas have been tentative and unimpressive. There is no compelling vision for how the business will generate meaningful value over the coming years, and its pipeline of future opportunities looks thin at best. In an environment that rewards bold, forward-thinking organizations, this company appears content to coast — a posture that, over time, is indistinguishable from decline. I see little reason for optimism here."

Neutral Stories

Prompt. *Write 4 analyst narratives, one each for a technology company, a food company, a construction company, and an energy company. For each company, write two versions of the narrative: one beginning with "My evaluation of this company is positive." and one beginning with "My evaluation of this company is negative." After the opening sentence, the text should consist entirely of neutral, factual descriptives about what the company does and how it operates — do not mention any positive or negative aspects, strengths, weaknesses, or evaluative language. The text after the opening sentence should be identical across the two versions for each company. Each full narrative (opening sentence included) should be approximately 90 words. Write from the first-person perspective of a financial analyst. Do not include any numbers or statistics. Do not mention any specific company name. Label each narrative by sector and valence.*

The neutral stories are sector-specific. The opening sentence varies by valence (positive or negative), but the remainder of each story is identical across valence conditions.

Technology Company. "My evaluation of this company is [positive/negative]. The company is in the software industry. Its employees are based in offices in several locations. The firm's products relate to data storage, internal communications, and digital record-keeping. The company charges its clients on a recurring basis. A portion of the company's budget is spent on developing new products. Its clients are organizations of varying sizes in both the public and private sector. The company's operations are organized into several internal divisions. The firm has been in operation for a number of years."

Food Company. "My evaluation of this company is [positive/negative]. The company is in the packaged food industry. It operates processing facilities and uses trucks for distribution. Its products include frozen meals, canned goods, and dry snacks, which are sold through retail stores. The company obtains its ingredients from agricultural suppliers. The size of its workforce varies by time of year. The firm's facilities are subject to periodic regulatory inspections. Its products are distributed within the domestic market. The company is organized into separate departments for production, distribution, and administration."

Construction Company. "My evaluation of this company is [positive/negative]. The company is in the construction industry, working on both commercial and residential structures. It owns construction equipment and employs both permanent and temporary workers. The firm handles multiple projects at the same time, with durations that vary depending on scope. New projects are obtained through a bidding process. The company holds certifications that are standard in the industry. It obtains building materials from a set of suppliers. The firm's headquarters and main equipment storage are at a single site."

Energy Company. "My evaluation of this company is [positive/negative]. The company is in the electricity sector, involved in both generation and distribution. It operates power plants that use different fuel sources. Its customers include households and businesses, who are billed through utility contracts. The company employs technical and administrative staff across several offices. Its infrastructure includes power lines, substations, and meters. The company is subject to regulation by governmental authorities. Equipment maintenance follows a recurring schedule. The firm's service area covers a defined geographic region set by its operating license."