Explanations*

Thomas Graeber Christopher Roth Constantin Schesch

September 15, 2025

Abstract

When people exchange knowledge, both truths and falsehoods can proliferate. We study the role of explanations for the spread of truths and falsehoods in 15 financial decision tasks. Participants record the reasoning behind each of their answers with incentives for accuracy of their listeners' responses, providing over 6,900 unique verbal explanations in total. A separate group of participants either only observe one orator's choice or additionally listen to the corresponding explanation before making their own choice. While listening to explanations somewhat improves average accuracy, there is substantial heterogeneity: explanations enable the spread of truths, but do not curb the contagion of falsehoods. To study mechanisms, we extract every single argument provided in the explanations, alongside a large collection of speech features, revealing the nature of financial reasoning on each topic. Explanations for truths are richer and contain higher argument quality than explanations for falsehoods. These content differences in the supply of explanations for truths versus falsehoods account for 60% of their asymmetric benefit, whereas orator and receiver characteristics play a minor role.

Keywords: Explanations, Social Learning, Speech Data, Financial Knowledge

---

* We thank Simon Cordes, Pietro Ducco, Maximilian Fell, Paul Grass, Jindi Huang, Milena Jessen, Julian König, Malte Kornemann, Maximilian Müller, Nicolas Röver, Gabriel Saliby, and Georg Schneider for outstanding research assistance. We thank our discussant at the Cognitive Economics Conference, Kirby Nielsen, and seminar audiences at the MPI Bonn, Duke, CESifo behavioral, the MiddExLab, Innsbruck, the University of Pittsburgh, the University of Zurich and Harvard Business School for useful feedback. The research described in this article was approved by the Institutional Review Board at Harvard Business School and the ethics committee of the University of Cologne. Roth: Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1-390838866. This work was partially supported by the RCN through its CoE Scheme, FAIR project No 262675. The data collections were pre-registered on AsPredicted (#155323; #157147; #159277; #224365; #234443). Graeber: University of Zurich, thomas.graeber@econ.uzh.ch. Roth: University of Cologne, IZA, ECONtribute, CEPR, NHH, and Max-Planck Institute for Research on Collective Goods, roth@wiso.uni-koeln.de. Schesch: Sciences Po, constantin.schesch@sciencespo.fr.

1 Introduction

We obtain most ideas, news, and knowledge from listening to others (Hirshleifer, 2020; Schotter, 2023). Some of the information we receive is accurate, while some of it is flawed. Whether learning from others improves or impairs our decisions therefore critically depends on our ability to discern what is right and what is wrong. The epitome of a welfare-improving social aggregation of information is the *marketplace of ideas*: the truth will emerge and prevail in an environment where thoughts, insights and knowledge are freely exchanged.[1] Yet, misbeliefs can spread rapidly, too, whenever individuals systematically fail to identify falsehoods (Lazer et al., 2018). Indeed, some argue that recent technological advances catalyze the spread of falsehoods, marking the onset of a "post-truth era."

At the center of any knowledge exchange are explanations: people share justifications for and reasoning behind their beliefs and choices, often conveyed by word of mouth from peer to peer (Shiller, 2017). Unlike in canonical economic models of social learning, people do not only learn from observing *what* someone else does or believes, but also from *why* they do so. To examine how explanations affect the contagion of truths and falsehoods, we conduct large-scale experiments in which respondents solve canonical financial decision tasks, receive one of over 6,900 explanations recorded by other respondents and are then allowed to update their answer. We focus on financial decisions because they are known to be shaped by information that circulates through social networks (Duflo and Saez, 2003) and include topics where false narratives are widespread, such as cryptocurrencies or stock-picking.

We run a series of pre-registered experiments using a design comprising two separate stages. We start with an Orator experiment that characterizes the *supply of explanations*. Respondents complete 15 canonical financial decision problems with an objectively correct answer, allowing us to characterize mistakes. These include questions on nominal illusion, the net returns of active and passive investing, the relationship between interest rates and bond prices, compounding of interest and other topics. In each task, respondents first indicate their choice, with incentives for accuracy. Then, they record a voice message in which they provide an explanation for their answer to randomly matched participants in a separate study. Orators' incentives induce aligned interests with the listeners of their recording: an orator's likelihood of receiving a bonus payment

[1]The marketplace of ideas is a foundational rationale for freedom of speech and open discourse, frequently attributed to U.S. Supreme Court Justice Oliver Wendell Holmes Jr.

increases with the accuracy of the listener's subsequent response to the question.[2] The faithful exchange of knowledge pervades real-world interactions in education (e.g., classroom teaching), business (e.g., collaboration across business functions), healthcare (e.g., patient-doctor consultations, public health campaigns), the legal system (e.g., advice on contracts) and finance, including peer-to-peer information sharing.

To study the *interpretation of explanations* and its consequences for social learning, we then conduct a Receiver experiment with 1,103 respondents, in which respondents face the same 15 tasks. They first make their own incentivized choice. For each task, we randomly assign respondents to the *Choice Only* or the *Explanation* condition. In *Choice Only*, they learn about the choice of a randomly chosen respondent in the Orator experiment. In the *Explanation* treatment arm, participants additionally listen to the respondent's explanation. In both conditions, respondents then again select their own best choice, which may now differ from their initial choice, yielding a dataset with more than 30,000 incentivized choices. The comparison between *Explanation* and *Choice Only* allows us to identify the specific effect of listening to a verbal explanation on imitation, above and beyond the mere observation of another respondent's choice. The *Choice Only* condition provides a natural benchmark that captures learning in the absence of an explanation, and further allows us to control for the direct effects of the respondents' confidence in their prior answers, measurement error in priors and other factors, such as experimenter demand.

Do explanations matter? We begin by analyzing how explanations shape average optimality rates. The prior optimality rate across tasks is 55.3%, meaning participants perform far better than the 33.3% rate implied by choosing randomly in a three-option task. In *Choice Only*, just observing someone else's answer increases the frequency of optimal choices to 59.5% ($p < 0.01$). Explanations somewhat increase the aggregate improvement from social learning: across all tasks, the *Explanation* treatment raises the accuracy rate to 62.7% ($p < 0.01$). This modest aggregate treatment effect, however, conceals that a large fraction of receivers encounter confirmatory advice (58.5%), in which case little or no improvement should be expected. We decompose the aggregate effects by separately analyzing different combinations of prior receiver and orator accuracy. First, in "learning situations," a receiver with an incorrect prior receives information from a correct orator. Second, in "unlearning situations," a receiver with a correct prior receives information from an incorrect orator. Third, we study receivers who are either correct or incorrect and receive a confirmatory signal.

[2]Our *explanations* thus differ from *persuasive messages*, which can reflect misaligned incentives.

We find that the aggregate benefit of explanations over merely observing someone's answer is predominantly driven by learning opportunities: in those cases, the imitation rate is only 42.8% in *Choice Only* but 55.8% in *Explanation*. This corresponds to a 13.0 p.p. ($p < 0.01$) increase in seizing learning opportunities. The *Explanation* treatment does not, however, decrease the frequency with which receivers switch to a wrong answer in unlearning opportunities. Receivers switch from accurate to inaccurate answers in 23.1% and 22.9% of unlearning situations in *Choice Only* and *Explanation* respectively ($p = 0.87$). Finally, we turn to receivers that receive confirmatory signals and, perhaps unsurprisingly, document muted effects of explanations. In situations where the orator and receiver are both correct, the posterior optimality rate is 99.1% and 99.2% in both treatment groups ($p = 0.77$). When both the receiver and the orator begin with an incorrect answer, the *Explanation* condition yields a slightly higher optimality rate (3.8%) than *Choice Only* (2.0%; $p < 0.01$) that contributes only 12% to the aggregate treatment effect.

Average effects of explanations also shroud heterogeneity at the task level. Explanations have a positive effect in 14 out of 15 tasks, though it is only statistically significant for 6 tasks. At the same time, asymmetric effects of explanations for learning and unlearning situations emerge in 14 out of 15 tasks, though only in a statistically significant manner for 7 of these.[3]

As a starting point for understanding these findings, we turn to the role of confidence. Prior work indicates that confidence in one's choice can shape both the supply of explanations and the likelihood that others choose to imitate. We first analyze heterogeneous treatment effects by orators' confidence levels. We show that imitation in learning situations is significantly higher when respondents are exposed to confident orators in the *Explanation* condition, suggesting that confident orators supply more credible explanations. We then examine the heterogeneous treatment effects by receiver confidence. Our data show that receivers with high prior confidence—who typically do not change their response based on the orator's choice alone—become more likely to change their mind when in the *Explanation* condition. This evidence hints at an important role of confidence and motivates the following more systematic unpacking of our findings.

Why do explanations help? We develop a simple model to interpret and decompose how explanations affect social learning. Receivers observe another person's answer but are uncertain about its reliability. Without explanations, all answers are treated as equally informative. With explanations, however, receivers obtain an imprecise signal about the underlying strength of

[3]Our experiments were not designed with sufficient statistical power for task-level analyses. We cannot reject the null hypothesis that the variation across tasks reflects pure sampling variation.

the signal. The agent's optimal response to this imprecision creates a *pivot* in belief movements, leading to overinference from weak and underinference from strong signals, as in Augenblick et al. (2025). In our model, explanations may also uniformly *shift* beliefs—making correct answers more persuasive regardless of confidence—or change how strongly receivers rely on their *priors*.

Taking the model to the data to decompose the treatment effect into these three components, we estimate regressions leveraging elicited confidence (Grether, 1980). First, the belief data corroborate the patterns found in the choice data. Second, the decomposition of the average treatment effect reveals that only a modest share of it reflects greater sensitivity to orator beliefs (pivot) with explanations than without, while explanations in fact reduce reliance on informative priors. The dominant force is a sizable shift effect: correct explanations receive a uniform credibility boost, which drives the entirety of aggregate gains. These asymmetric gains point to additional informational content embedded in correct explanations.

This model-guided evidence raises the question of what information explanations transmit. Is the impact of explanations akin to observing the orator's numerically expressed confidence, which is, on average, higher in learning than in unlearning opportunities? To test this hypothesis directly, we conduct an additional Receiver experiment, in which respondents observe both the orator's choice and their stated confidence. Compared to only observing the orator's choice, additionally learning about their confidence does not significantly shift imitation in learning and unlearning opportunities on average. Therefore, an explanation is a signal of accuracy that is distinct from the orator's numerical confidence.

Which features of oral explanations convey the additional information? Both the content—*what* is said—and its delivery through the speaker's voice—*how* it is said—may shape the receivers' behavior. To distinguish between these two channels, we design an additional Receiver experiment, in which respondents read the transcript of the orator's explanation instead of listening to the corresponding recording. This preserves the exact verbal content of the explanation while eliminating the role of the speaker's delivery and voice, effectively reducing the dimensionality of the message space. A strongly asymmetric effect on learning and unlearning persists in the *Transcript* treatment, suggesting the effect of explanations is primarily driven by the supply and interpretation of their *content*. While we exactly replicate the average null effect of explanations in unlearning situations of the *Transcript* treatment, there is an 8.1 p.p. increase of the imitation rate in learning opportunities (relative to *Choice Only*), which equals approximately two thirds of the effect size induced by *Explanation*. The effect size difference across *Transcript*

and *Explanation* is statistically significant ($p = 0.03$). This difference points to the possibility that the effectiveness of explanations depends on the dimensionality of the message space.[4]

Having established that the impact of explanations is distinct from confidence statements and primarily operates through content, the question remains as to why the asymmetric effect of explanations emerges. To address this, we conduct a mechanism analysis that systematically examines content differences in the supply of explanations.

Mechanisms: The content of explanations. In the first step of our mechanisms analysis, we characterize the content of the more than 6,900 explanations and study its effects on imitation. Analyzing the content of speech recordings is non-trivial due to the high-dimensional nature of language data: each sentence has innumerable features and interpretations. We pursue a two-pronged approach based on the following distinction. On the one hand, explanations are characterized by their substantive content: they provide *arguments*, which tend to be domain-specific and directly relate to a specific question and answer. We identify and code every argument provided across the universe of our explanations, delivering the first dataset of its kind for studying their effect on social learning and unlearning in a controlled setting. On the other hand, explanations are characterized by a large number of text features: they exhibit various classes of expressions (e.g., certainty phrases, hedges or questions), many linguistic and rhetorical attributes, and can be described by speech and text metrics, among others. We code a broad set of text features from the existing literature. These domain-general features apply across explanations for different questions and answers. To analyze transcripts, we combine human coding, a large language model, and machine learning methods, ensuring robustness and replicability.

Our argument annotation yields, for each of the 15 tasks, the full set of arguments, their frequency in correct versus incorrect explanations, and estimates of their effects on imitation. To illustrate, consider the task on whether actively managed funds outperform passive ones. The dominant argument—that active funds can quickly adapt to market changes—appears in 57.2% of incorrect explanations but only 4.0% of correct ones. By contrast, the claim that active funds charge higher fees occurs in 22.7% of correct explanations but just 3.2% of incorrect ones. In total, we identify 11 distinct arguments. Some are particularly influential: for instance, "active

[4]One interpretation of this lower treatment effect might be that people are less attentive in the *Transcript* treatment than the *Explanation* treatment. Yet, our decomposition exercise displayed in Table E10 shows that 74% of the differential effect of explanations in the *Transcript* treatment is explained away by our content-based measure of explanation richness. This, in turn, suggests that respondents are highly attentive to differences in the content of the scripts.

funds can react to the market" (in unlearning situations) and "passive funds target long-term growth" (in learning situations) both raise imitation by more than 20 p.p. relative to *Choice Only*. Across tasks, we document heterogeneity in the number of arguments, the degree of consensus, and the strength of their marginal effects on imitation. This perspective provides a useful lens on how individuals reason and how their reasoning persuades others. It does not, however, isolate the mechanisms underlying the asymmetric treatment effect, because argument content is not directly comparable across choices and tasks.

To compare explanations for correct and incorrect choices across tasks, we classify arguments into four categories of increasing quality: (i) no argument, (ii) irrelevant (off-topic) arguments, (iii) fallacious arguments with false premises or invalid reasoning, and (iv) sound arguments with true premises and valid conclusions. This classification matters since respondents can sometimes arrive at the right answer even with irrelevant or fallacious arguments. We find striking differences across contexts: in unlearning situations, respondents more often face no argument (21.8% vs. 16.0% in learning), irrelevant arguments (22.6% vs. 17.7%), or fallacious ones (51.0% vs. 10.5%), whereas in learning situations they are much more likely to encounter sound arguments (55.8% vs. 4.7%).

Are argument classes linked to different imitation effects? Indeed, higher-quality arguments yield stronger effects. The absence of an argument or the presence of an irrelevant one tends to reduce imitation, while even fallacious arguments raise it. Yet, conditional on argument class, imitation remains far higher in learning than in unlearning. In fact, the argument gap explains at most 25% of the asymmetry.

What content features beyond argument quality distinguish explanations for correct and incorrect answers? An analysis of domain-general speech and text markers (e.g., certainty) shows that explanations for correct answers are substantially *richer*: they include more of most features, even controlling for length. In fact, 24 of 31 features appear more often in correct explanations. A pre-registered LLM-based measure confirms this pattern: richness scores are on average 0.77 SD higher for correct than for incorrect explanations.

Can differences in explanation features account for the asymmetric treatment effect? We first test which features predict imitation and find that richness is by far the strongest predictor: a one–standard deviation increase in richness is associated with a 13.0 p.p. increase in imitation ($p < 0.01$), even after controlling for length and other features. This benefit diverges from *Occam's Razor*: rather than preferring simplicity, listeners are more persuaded by comprehensive,

detailed explanations, perhaps because they are taken as signaling credibility. Next, we show that the richness gap explains about 60% of the differential effect between learning and unlearning opportunities. We investigate the key drivers of the richness effect and find that linguistic richness—the complexity, variety, and uniqueness of vocabulary—plays a pivotal role. Finally, we systematically test and rule out differences in participant characteristics (both as orators and receivers) as the source of the asymmetry.

Motivated by this correlational evidence, we conduct an additional receiver experiment to isolate the causal effect of linguistic richness. We exogenously vary linguistic richness while holding factual content and arguments constant, by instructing a large language model to generate two versions of each transcript: a rich version with structured, coherent language, and a sparse version with fragmented sentences. Participants were randomly assigned to one version at the task level. Richness significantly increases imitation: from 40% to 58% in learning scenarios and from 15% to 26% in unlearning. These effects remain when restricting to transcripts where both versions contain exactly the same arguments, highlighting the importance of linguistic richness.

Taken together, our mechanism evidence suggests that content differences are the key determinant of imitation decisions in our setting. We leverage novel methods that were unavailable to the earlier literature on advice and social learning (see, e.g., Schotter, 2023), providing direct access to *what* people communicate in unconstrained natural language and how it affects imitation. Our data allow us to compare and emphasize the role of content over that of the identity of speakers on imitation, the focus in much of the previous literature (Cialdini, 2001). The overall beneficial effect of explanations on optimality directly hinges on the positive association between richness and truth. This relationship may be specific to settings that have, like ours, aligned incentives between speakers and listeners to identify the truth in the exchange of knowledge rather than opinions.

Finally, we note that our main financial decision-making tasks involve high-dimensional signals that combine arguments, reasoning, and knowledge. To further benchmark the effects in this rich setting, we leverage a more abstract, self-contained task in which communicable information is plausibly lower-dimensional: the classic balls-and-urns problem (Edwards, 1968). In contrast to our financial tasks, explanations have more muted effects on imitation and posterior accuracy. Together, these results suggest that explanations may be more effective in environments in which communicable information is high-dimensional.

Literature. Our paper contributes to an emerging literature on learning from qualitative information, e.g., in the form of stories and narratives (e.g., Graeber et al., 2024; Andre et al., 2025; Kendall and Charles, 2022; Hüning et al., 2022; Schwartzstein and Sunderam, 2021; Eliaz and Spiegler, 2020; Bursztyn et al., 2023). Barron and Fries (2024) study strategic communication of model parameters as a persuasive tool when financial advisors hold incentives that differ from those of the individuals they are advising. Thaler et al. (2025) provide evidence that senders with incentives to directionally persuade are more likely to communicate using language rather than numbers. Graeber et al. (2025) examine how verbal transmission distorts the supply of qualitative economic information and show that information about signal reliability gets lost in transmission more than information about signal values because it fails to come to mind. We differ from existing work in our focus on characterizing the supply and interpretation of explanations for people's choices in canonical financial decision problems.

We relate to an interdisciplinary literature on explanations (Lombrozo, 2006; Langer et al., 1978) and arguments (Sloman et al., 1998). Our contribution lies in providing a characterization of the supply of qualitative explanations and in estimating their consequences for economic decisions in a controlled setting. Our experimental manipulations of explanations highlight that linguistic richness shapes social learning, above and beyond arguments and facts.

We further contribute to a literature that studies how social learning among non-experts—labeled "naive advisors" by Schotter (2003)—affects the prevalence of biases and misinformation.[5] In a setting that—unlike ours—features incentives for deception, Serra-Garcia and Gneezy (2021) show that individuals fail to detect others' lies when they are shown a video message of another respondent paid to invent a news story. A series of papers has examined social learning in the context of motivated beliefs (Oprea and Yuksel, 2022; Thaler, 2025). Grunewald et al. (2024) study whether biases are contagious in a setting with motivated beliefs. They find that communication of personal opinions via a written text message amplifies belief biases relative to a setting of observational learning. Conlon et al. (2025) show that in the context of a balls-and-urns updating task, people are less sensitive to information others discover than to equally relevant information they receive themselves.

[5]The literature on social learning (typically defined as observational learning from others' actions, our control condition) is vast (Jackson and Yariv, 2007; Galeotti et al., 2010), especially in the context of financial decisions (Ambuehl et al., 2022; Bursztyn et al., 2014; Akçay and Hirshleifer, 2021; Hirshleifer et al., 2023). The present paper builds on an earlier literature on advice-giving (Schotter and Sopher, 2003; Çelen et al., 2010; Schotter, 2023), which has focused on pre-structured messages and not yet studied the nature and causal effect of explanations in natural language.

Finally, by characterizing the spread of truths versus falsehoods through social learning, we contribute to a long-standing literature on whether individual-level biases matter for aggregate market-level outcomes (e.g., Fehr and Tyran, 2005). Enke et al. (2023) show that awareness about biases reduces the impact of individual-level biases on aggregate outcomes through institutions that rely on self-selection, while Amelio (2024) studies how meta-cognition shapes social learning. Unlike those findings, ours cannot, by design, be explained by meta-cognition. Instead, we examine how explanations affect perceptions of others' accuracy and thereby the proliferation of truths and falsehoods.

2 Experimental Design

2.1 Overview

Our experimental design studies 15 canonical financial decision problems and consists of two stages. In the *Orator* experiment, respondents record an explanation for their answer for each of the tasks. In the subsequent *Receiver* experiment, respondents first provide their choice. Then, they either only see another respondent's choice (from the Orator experiment) or additionally listen to that respondent's explanation, before providing their answer to the same task again.

2.2 Financial Decision Problems

We select 15 financial decision tasks based on three criteria. First, we aim for a collection that is broadly representative of the reasoning and decision biases studied in the finance literature. This spans behavioral phenomena like exponential growth bias and nominal illusion, but also more specific knowledge about different asset classes and investment decisions, for example expected returns under active versus passive investing. Many of the problems we study are tightly linked to common high-stakes financial decisions, such as whether to invest in active or passive funds. Second, we restrict our attention to questions with an objectively correct answer or ones where a broad consensus exists in the financial economics literature, excluding those that rely on tastes. Moreover, participants in our studies are made aware that a "correct" solution exists.[6] Third, the questions should be reasonably short. Some tasks have a logically correct answer, for example the following question about the concept of inflation, with the correct answer underlined:

[6]This is motivated both by our focus on the exchange of knowledge—rather than opinions—and a precondition for incentivizing answers. The effect of explanations on imitation choices might be different in settings where people think that no correct answer exists.

Imagine that the interest rate on your savings account was 2.5% per year and inflation was 3% per year. After 1 year, how much would you be able to buy with the money in this account?

1. *More than today*
2. *Exactly the same as today*
3. *Less than today*

Other questions relate to a broad consensus in financial economics, like the following:

Do actively managed investment funds systematically outperform passively managed investment funds in terms of expected net returns, i.e., after accounting for investment fees?

1. *Actively managed funds outperform passively managed ones.*
2. *Actively managed funds do not outperform passively managed ones.*

We embrace that differences across these tasks will likely evoke structurally different explanations. A participant might give a wrong answer because they have not heard of the concept of a call option—in a sense, they may not really know what the question is about—or they fully understand the question but still do not know its answer. This difference captures two separate important features of bias in practice and may be reflected in explanations as we discuss in the following sections. Appendix Table E6 outlines the motivation and origin of the tasks, as well as the exact wording of all questions. Three tasks have two options, while the others all have three. In most analyses, we are only interested in whether the correct option is chosen.[7]

2.3 Part 1: Orator Experiment

The main objective of the Orator experiment is to obtain recordings of people's verbal explanations for each of the financial decision tasks.[8] In the beginning, participants are told that we are interested in how they would give advice in an informal conversation. They are informed that they should share an explanation behind their response and that their recording will be played to other participants who will have to answer the same question. We ask respondents not to search for answers on the internet.[9]

[7]Due to these differences, one could expect different frequencies of correct responses, because randomizing would create an optimality rate of 50% in a two-option and 33% in a three-option task. However, this is constant across conditions and thus cannot affect treatment comparisons. Appendix C.2 shows treatment effects of explanations are very similar for tasks with two and three response options.

[8]The full set of instructions is reproduced in Online Appendix L.

[9]We ask participants at the end of the study whether they searched for any answers, stressing there is no penalty for indicating that they did. As pre-registered, we exclude the 7.0% of participants indicating that they searched for answers from our data.

People typically have some time to think about explanations they give. Correspondingly, rather than forcing respondents to talk immediately upon reading the question, we show them the question first and they decide when to start their recording. An example screen is shown in Online Appendix Figure K19. After recording their explanation, respondents first select their preferred answer and then state their confidence in its accuracy by answering "*How certain are you that your above answer is correct?*" on a scale from 0 "*Not at all certain*" to 100 "*Fully certain*".

Incentives. With a 10% chance, a respondent is eligible for a bonus payment of $10. Whether a selected respondent receives a bonus is based on one randomly drawn task. The orator is matched with another randomly selected participant in the Receiver experiment, who either only sees that orator's answer or additionally listens to their voice recording. The bonus is paid if the matched receiver gives the correct answer after exposure to the orator's answer. Our experiment thus creates aligned incentives between the orator and the receiver: the orator is incentivized not to be imitated *per se*, but to induce the receiver to make the right choice. The orators' instructions emphasize that their incentives will be known to listeners ("Participants listening to your recordings will be informed that you will receive a bonus if they select the correct answer."). We confirm that orators understand the aligned incentives scheme using a control question.

Speech recordings. The Orator experiment relies on speech recordings of people's explanations. Relative to written text, speech recordings have a series of advantages for our purposes. A voluminous literature outside of economics has characterized the differences between written and spoken text production (e.g., Akinnaso, 1982). Written text tends to be more formal, structured and cognitively taxing to produce (e.g., Bourdin and Fayol, 2002). Because writing text is typically more exhausting than speaking, the transcripts of orally provided explanations are often substantially longer. Much of social learning follows from oral conversations, making speech recordings an ideal testing ground to study the effects of explanations. Second, speech data include features of natural language that plausibly affect social learning but are mostly absent from written texts, including tone, emphasis, and disfluencies such as pauses, repetitions, revisions, hesitations, or filler words. Third, writing text as opposed to spontaneously talking about one's thoughts adds another filter that may distort measured explanations compared to explanations people give spontaneously in the real world.

2.4 Part 2: Receiver Experiment

To characterize the effect of explanations on social learning, we conduct a Receiver experiment that leverages the choices and recordings from the Orator experiment.[10]

As in the Orator experiment, respondents complete the 15 decision tasks. To measure imitation rates at the individual level, we use a within-design with five steps in each round. First, respondents read the financial decision task and are incentivized to indicate their preferred choice, which provides our measure of their prior belief. Second, they indicate their confidence in the accuracy of their response in the same format as respondents in the Orator experiment. Third, they either only learn about the choice of another randomly selected respondent in the Orator experiment (*Choice Only* condition) or additionally listen to the recording of their explanation (*Explanation* condition). Fourth, the receiver again has an opportunity to select their preferred choice with incentives for accuracy. Fifth, they indicate their confidence in their posterior answer.

Treatments. In the *Choice Only* treatment, receivers may infer and adjust their belief about the optimal answer from learning what someone else chose, even absent an explanation. This same source of learning is present in the *Explanation* treatment, but the explanation provides an additional source of information. We randomize treatments between participants, at the task level. For each task, 80% of receivers are sampled into the *Explanation* condition, while the remaining 20% are assigned to *Choice Only*.[11]

The comparison between *Explanation* and *Choice Only* allows us to identify the specific effect of listening to a recorded explanation on learning and unlearning, above and beyond the mere observation of another respondent's choice. The *Choice Only* condition is critical to control for (i) the effects of prior confidence, (ii) measurement error in priors, and (iii) other confounders, such as experimenter demand effects. At the same time, this comparison captures various potential channels of learning which we disentangle through additional treatments.

Incentives. Receivers have a 10% chance of being eligible for an additional $10 bonus payment. Whether they receive the bonus is determined by the accuracy of their answer in a randomly selected reasoning task. For every task, we randomly select whether their first answer or their second answer is the decision that counts for the bonus.

[10]We provide the full set of instructions in Online Appendix M.

[11]We oversample the *Explanation* condition to obtain the statistical power needed to examine heterogeneous effects by features of the explanations.

2.5 Logistics

Respondents in both studies received a base reward of $6 for completing the study. Median completion times were 25 minutes in the Orator experiment and 26 minutes in the Receiver experiment. All experiments were conducted on the online platform Prolific, which is widely used for experiments in the social sciences. The Orator experiment was run for a total of 505 U.S. respondents in December 2023, of whom 466 provided valid responses.[12] Participants were required to have a working microphone to record their voice message. The Orator experiment yields a total of 6,910 valid recordings obtained by integrating speech recordings with Phonic into Qualtrics surveys. We use an Amazon Web Services backend to stratify and distribute recordings into our Receiver experiment. The Receiver experiment was run with 1,235 U.S. respondents in December 2023, of whom 1,103 provided valid responses. Appendix Table E7 provides an overview of all data collections and corresponding pre-registrations.

3 Explanations and Social Learning

We start by providing basic descriptives about our respondents' explanations. We then turn to the effects of explanations on imitation and optimality. We conclude with additional results on heterogeneity and robustness.

3.1 Basic Characteristics of Explanations

Our Orator experiment generated 6,910 audio recordings with a median duration of 26 seconds. We find substantial variation in length: the 10th percentile is 11 seconds and the 90th percentile 55 seconds. The audio quality of recordings is high. Analyses of the audio files show that only 1.1% of recordings are unusable, typically because of a technical microphone problem or because respondents submitted it too early by mistake.[13] We use transcripts of the recordings that preserve details and nuances of spoken language, notably filler words such as "um" and "eh." The median length of the resulting transcripts is 55 words.

To parse basic features of the scripts, human coders classify them using a simple coding scheme (see Appendix B). We find that 13.1% of explanations are pure restatements of the

[12] In accordance with our pre-registration, we determine orators gave valid responses if not all their recordings were blank and if they did not indicate that they looked up answers online. We similarly drop receivers who self-reported looking answers up online.

[13] Incomprehensible messages or high background noise appear very rarely and are therefore not relevant concerns for our study.

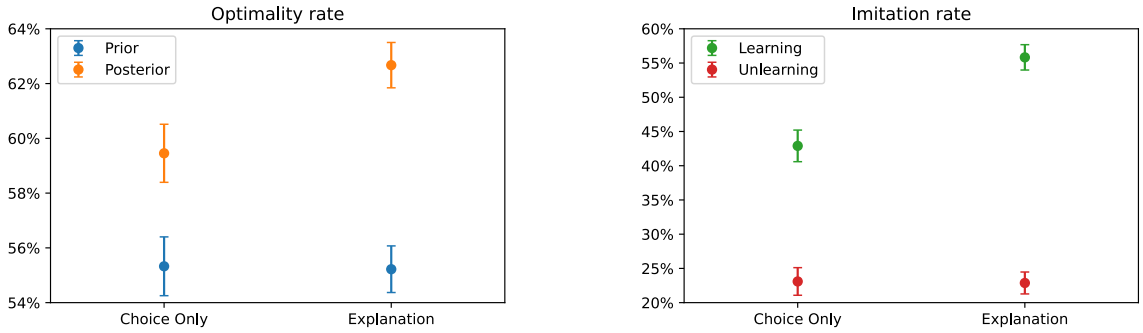
question and/or the answer, without adding any content matter. These may both be due to people not trying or not having any explanation for their answer. We characterize a negligible minority of 2.6% of recordings as nonsensical. Looking at the cases that reflect some form of actual explanation, we find that 8.6% of all explanations contain *no substantive arguments* while 74.3% of all explanations contain *substantive arguments*. Finally, 13.7% of recordings contain explicit expressions of the speaker's confidence.

This first look at basic descriptives suggests the Orator experiment provides a rich and varied database of heterogeneous explanations for our set of tasks. In Section 5, we go much further in examining their characteristics.

3.2 Explanations and Optimality

We start by analyzing the effects of our treatments on the frequency of correct choices, which we refer to as the *optimality rate*. For comparability with additional treatments and to maximize statistical power, we pool observations for *Choice Only* obtained from different between-subject collections that use this exact same control condition (see Sections 4.2 and 4.3). The prior optimality rate reflects receivers' knowledge about a task before learning from another respondent. The posterior optimality rate captures accuracy after receivers observe another respondent's answer only (*Choice Only*) or additionally listen to their verbal explanation (*Explanation*).

Figure 1a shows these optimality rates pooled across all 15 tasks. Prior to exposure, 55.4% and 55.2% of respondents provided correct answers in the *Choice Only* and *Explanation* conditions respectively ($p = 0.88$). We document two main findings on posterior optimality. First, just observing another's choice increases optimality rates by 4.1 p.p. ($p < 0.01$), creating an aggregate improvement. Second, additionally listening to another person's explanation raises the size of the improvement to 7.5 p.p. (0.15 SD, $p < 0.01$). This difference in improvement rates between the *Choice Only* and *Explanation* conditions is statistically significant ($p < 0.01$) but quantitatively modest. Importantly, this aggregate treatment effect reflects that a large fraction of receivers encounter confirmatory advice (58.5%), which plausibly creates limited scope for social learning. Moreover, it masks important variation across initially correct and incorrect listeners, different tasks and explanations, which we turn to below.



(a) Prior and posterior optimality by treatment

(b) Imitation in learning and unlearning

Figure 1: Effect of *Explanation* on optimality and imitation. *Notes:* Left panel shows share of correct receivers before and after exposure to the orator's choice or explanation. Right panel shows share of receivers picking the same answer as the orator in learning and unlearning situations. *Explanation* sample is the main Receiver experiment (1,103 receivers, 13,111 obs.), *Choice Only* is pooled from all collections (2,733 receivers, 8,232 obs.). Whiskers show 95% CIs.

3.3 Explanations and Imitation Rates

Our sample is composed of structurally distinct sets of orator-receiver matches. Specifically, pairs vary along two margins relevant for social learning: the initial accuracy of the listener and the accuracy of the orator. This creates four distinct groups, characterized by whether a receiver was initially correct, and whether they were subsequently exposed to a *confirming* or *conflicting* signal. Since the optimality rate is only modestly above 50% and matching is random, these four situations occur with broadly similar frequencies. Intuitively, situations where receivers hear conflicting advice should have the biggest scope for learning. We start by analyzing these.

Imitation in response to conflicting signals. Among respondents who receive conflicting signals, we distinguish between two distinct situations: those with initially incorrect choices who are exposed to correct ones, and those with initially correct choices who are exposed to incorrect ones. We refer to the former as *learning opportunities* and the latter as *unlearning opportunities*.[14] If the probability of being correct is p for both receivers and orators, under random matching, learning and unlearning situations both occur with frequency $p(1-p)$: indeed, in our sample, there are 21.2% learning and 20.3% unlearning opportunities.

Figure 1b displays the frequency of imitation in learning and unlearning opportunities. It illustrates two key findings. First, the unlearning rate does not differ significantly between *Choice*

[14]We will also refer to corresponding explanations as *learning explanations* and *unlearning explanations*.

Only and *Explanation*, at 23.1% vs. 22.9% ($p = 0.87$). About one of every four receivers with a correct prior confronted with another respondent's wrong answer switches away from the correct one. Thus, participants in unlearning opportunities do not, on average, infer information from explanations that systematically helps them identify the answer as wrong.

Second, we do find a quantitatively large treatment effect on the learning rate. Learning opportunities are far more likely to be seized in *Explanation*, where people imitate in 55.8% of cases, than in *Choice Only*, with 42.8%. This 13.0 p.p. ($p < 0.01$) increase in the learning rate shows that, on average, explanations help listeners identify the correct answer.

Imitation in response to confirmatory signals. To benchmark effect sizes in learning and unlearning contexts, we next examine imitation when receivers are exposed to confirming signals. Appendix Figure D4 displays optimality and imitation rates in these situations. When both the orator and the receiver are initially correct, imitation rates are extremely high, at 99.1% in *Choice Only* and 99.2% in *Explanation* ($p = 0.77$). By construction, posterior optimality rates are identical. This indicates that "destructive" arguments, agreeing with a listener's correct choice but nevertheless prompting them to switch to an incorrect one, are essentially absent in our data.

On the other hand, when both the orator and the receiver are initially incorrect, imitation rates stand at 82.6% and 82.2%, respectively ($p = 0.73$). Corresponding aggregate posterior optimality rates are quite small at 2.0% in *Choice Only* and 3.8% in *Explanation*, but do display a small 1.8 p.p. ($p < 0.01$) treatment effect. For this specific analysis, the distinction between two- and three-option tasks plays an important role and provides additional insights. In our twelve three-option tasks, explanations make receivers more likely to select the correct answer when the orator gave the same wrong answer as the receiver, but also when they supported a different wrong answer. Although these effects are small, we therefore find that arguments can play an "enabling" role, i.e., support the same or a different initial wrong choice as the receivers' but lead them to switch to the correct answer.[15]

Decomposing the aggregate effect. The 1.8 p.p. effect when both are wrong is an order of magnitude smaller than the 13.0 p.p. learning effect, while both situations occur similarly often (23.5% vs. 20.3%). Decomposing the 3.2 p.p. increase in posterior optimality in *Explanation* relative to *Choice Only*, we find that 80% of it is driven by learning, 12% by situations where both are wrong and 8% by unlearning.

[15]See Online Appendix C.2 for details.

Result 1. *Listening to another respondent's explanation somewhat increases the average optimality rate, relative to just observing their answer. The treatment effect is strongly asymmetric: explanations increase imitation in learning opportunities but do not decrease it in unlearning opportunities.*

Robustness. In Online Appendix C, we find that the asymmetric effects of explanations are robust to excluding the shortest and longest recordings, to distinguishing only by orator optimality, and to separating tasks with two or three options.

3.4 Cross-Task Heterogeneity and Robustness

The previous analysis pooled data across tasks, potentially obscuring task-level heterogeneity. Appendix Figure D6 shows optimality and imitation rates across tasks. Optimality rates vary from more than 90% for "Nominal Illusion" to less than 20% for "Disposition effect", with the difference in optimality after receiving advice and between treatments being almost an order of magnitude smaller than cross-task heterogeneity. We see this variability as a useful feature of our selection of tasks from the literature, since both harder and easier questions naturally occur in economic decision-making. In turn, imitation rates range from around 60% in learning situations to around 10% in unlearning situations, both for "Nominal Illusion". Tasks with higher optimality rates exhibit a larger gap between learning and unlearning, showing that, in both treatments, easier tasks make it easier to change incorrect beliefs and harder to change correct ones.

Turning to the effect of explanations, Figure 2 illustrates cross-task variation in our main findings. The left panel shows a positive treatment effect of *Explanation* on optimality in 14 of the 15 tasks.[16] All 95% confidence intervals contain the aggregate effect of 3.2 p.p., though the treatment effect is statistically significant in only 6 out of 15 tasks, reflecting lower power at the individual task level.

To separate the share of variation reflecting true heterogeneity rather than sampling variation, we perform Cochran's Q-test which yields a statistic of 10.69 with 14 degrees of freedom, failing to reject the null of homogeneity (p=0.71). The corresponding I^2 index is 0%, suggesting all variability should be attributed to sampling error. However, with only 15 tasks, power to detect modest variation is limited. Despite our large data collection of nearly 50,000 individual decisions, the noisy nature of social learning makes precise estimation at the task level difficult: though it should be stressed we are not very well-powered to measure it, we find no evidence to

[16]This improvement rate is not significantly correlated with the prior optimality rate in a given task ($r = -0.07$, $p = 0.81$), and neither is the net learning effect ($r = 0.13$, $p = 0.65$).

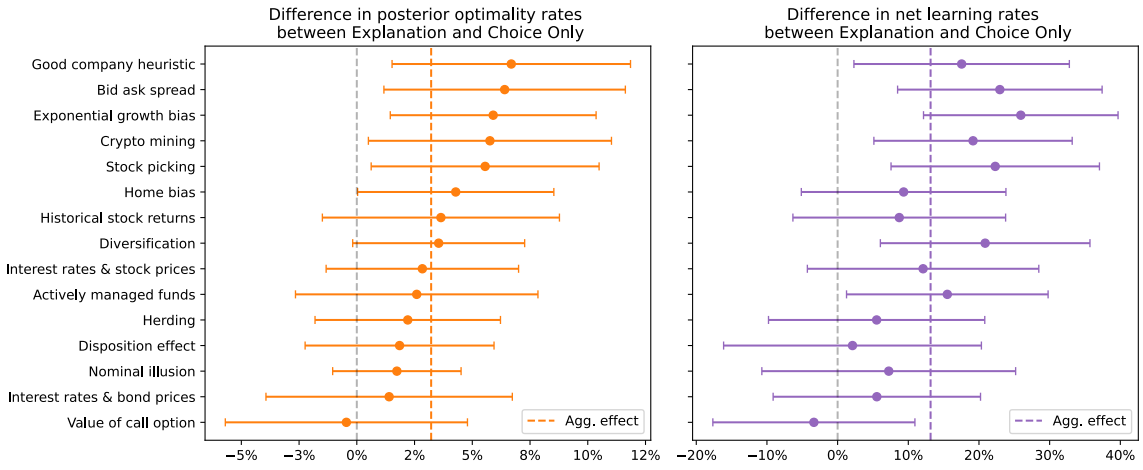


Figure 2: Difference in posterior optimality rates between *Explanation* and *Choice Only* by task, and difference in net learning rates between *Explanation* and *Choice Only* by task. *Notes:* The net learning rate is defined as the difference in imitation rates between learning and unlearning situations. Dashed vertical lines show aggregate effects across tasks, of 3.2 p.p. and 13.3 p.p. respectively. *Explanation* sample is the main Receiver experiment, *Choice Only* is pooled from all collections. Whiskers show 95% CIs.

suggest strong heterogeneity in the effects of explanation on optimality.

This conclusion is more nuanced for the asymmetric benefit of explanations. The right panel of Figure 2 shows the net learning effect of explanations, defined as the difference in learning rates between *Explanation* and *Choice Only* minus the corresponding difference in unlearning rates, is similarly pervasive. It is positive in 14 tasks, with confidence intervals also containing the 13.3 p.p. aggregate effect in 14 tasks, though it is statistically significant in only 7 tasks. The figure displays substantial heterogeneity, with estimates ranging from a maximum of +25.9 p.p. in "Exponential growth bias" to a median of +12.1 p.p. in "Interest rates and stock prices" to a minimum of -3.4 p.p. in "Value of call option." For net learning, Cochran's Q stands at 18.38 ($p = 0.19$) and the I^2 index is 24%, suggesting that roughly one-quarter of observed variation likely reflects genuine task-level heterogeneity.

3.5 Confidence and Social Learning

Our findings so far are based on actual imitation decisions and the implied optimality rates only. Our design also provides more granular data from orators' and receivers' confidence statements. We now leverage this confidence data to shed light on who imitates and who gets imitated.

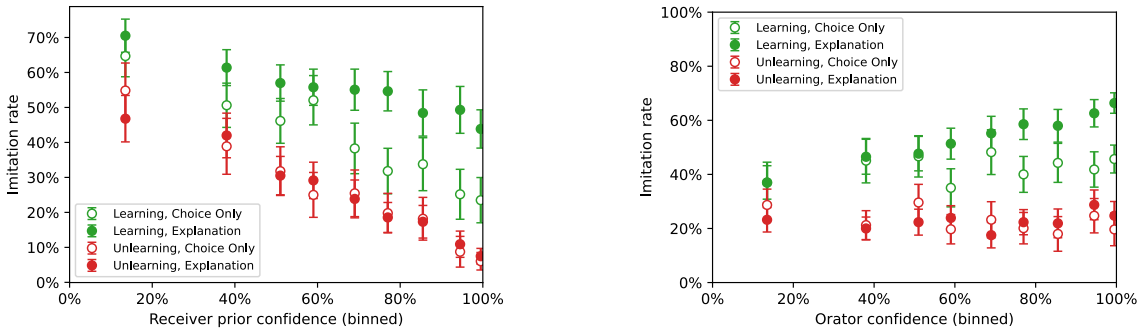
Confidence data. In our experiments, we asked both orators and receivers to report their confidence in their answer on a scale from 0% to 100%. It stands at 73.0% for correct and 62.0% for incorrect ($p < 0.01$) receiver prior choices. It is very similar at 75.0% and 63.0% ($p < 0.01$) respectively for orator choices.[17]

Heterogeneity by receiver confidence. We investigate how heterogeneity in receiver confidence interacts with the effect of explanations. One intuitive possibility is that explanations reduce the confidence threshold below which individuals are willing to imitate others. Figure 3a displays the imitation rate by bins of receiver prior confidence in *Choice Only* and *Explanation*, separately for learning and unlearning situations. As one would expect, imitation generally slopes down in receiver prior confidence. Just as there is no average effect of *Explanation* in unlearning situations, there is no statistically significant effect within any of the confidence bins. On the other hand, in learning situations, effects are relatively weak in the first bins up to 60% confidence, after which the treatment effect becomes much stronger. More precisely, splitting the sample by the overall median confidence of 74%, the net learning rate of explanations is 10.0 p.p. ($p < 0.01$) for unconfident receivers and about two times larger at 19.2 p.p. ($p < 0.01$) for confident ones, a significant 9.1 p.p. difference ($p = 0.02$). The learning benefit of explanations is therefore even greater for confident but wrong receivers.

Orator confidence. We now turn to the role of orator confidence in shaping social learning. Figure 3b shows that imitation generally slopes up in orator confidence for learning situations, though it is almost flat in unlearning situations. Moreover, among these, there is no explanation effect at any confidence level. Finally, although learning effects appear weaker up to 60% confidence, they are strong and growing above this threshold. For listeners exposed to orators with confidence below the median, the learning effect is 8.1 p.p. ($p < 0.01$), but two times larger at 16.2 p.p. ($p < 0.01$) above the median, also a significant 8.8 p.p. difference ($p = 0.05$). The learning benefit of explanations is thus amplified by confident and correct orators. Online Appendix G provides additional results on heterogeneity by confidence.

Taken together, these findings point to a central role of orator and receiver confidence in shaping imitation decisions. In the next section, we conduct a more formal decomposition of the effects of explanations into underlying forces that leverage the richness of the confidence data.

[17]Because of an error in the survey which allowed respondents to skip the question, prior confidence is missing for 0.6% of the sample. We drop these observations in this Section.



(a) Imitation by prior confidence

(b) Imitation by orator confidence

Figure 3: Effect of prior and orator confidence on imitation in learning and unlearning situations. *Notes: Explanation* sample is the main Receiver experiment (1,103 receivers, 13,111 obs.), *Choice Only* is pooled from all collections (2,733 receivers, 8,232 obs.). Whiskers show 95% CIs.

4 Interpreting the Treatment Effect of Explanations

Having documented a strongly asymmetric treatment effect of explanations as our main reduced-form finding, we now attempt to better understand the drivers of this treatment effect. To structure our analyses, we first illustrate the role of explanations in a canonical belief formation framework. We then present additional experiments that help characterize the treatment effect.

4.1 A Model-Based Decomposition

How can explanations be represented in a canonical updating framework? We present a simple model that directly motivates an empirical specification to decompose the treatment effect.

Model summary. We outline the setup and intuition of our framework and relegate all details and derivations to Appendix A. We directly build on Augenblick et al. (2025), who propose a Bayesian model of over- and under-inference due to imprecise information about, or imperfect integration of, signal strength.

In our setting, respondents face questions with correct answer $\theta \in \{-1, 1\}$, where 1 denotes the correct answer (without loss of generality). Their own reasoning about a question is modeled as a draw of a signal strength $S > 0$ and a direction d with $\mathbb{P}(d = \theta) = \Lambda(S)$, where Λ is the logistic function, so that stronger signals are more likely to be correct.

Receivers then observe the answer of an orator, which is an additional directional signal d, but only have imprecise information about its strength S. In *Choice Only*, the receiver does not get any information about signal strength, so they must rely on its unconditional distribution. In

Explanation, the verbal explanation delivers a noisy signal $\tilde{S}$ about signal strength S, which the receiver uses to modulate the magnitude of their belief update. We model this as standard log-normal Bayesian updating, where $\log S \sim \mathcal{N}(\log \bar{S}, \sigma^2)$ and $\log \tilde{S} = \log S + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, so that the receiver forms a posterior that shrinks the imprecise signal toward their prior about signal strength:

$$\log S \mid \tilde{S} \sim \mathcal{N}\left(\log \bar{S} + \lambda(\log \tilde{S} - \log \bar{S}),\ (1-\lambda)\sigma^2\right), \qquad \lambda = \tfrac{\sigma^2}{\sigma^2+\sigma_\varepsilon^2}.$$

How can this model shed light on the treatment effect of explanations? Its basic prediction is that, while receivers are entirely insensitive to the signal strength in *Choice Only*, they become responsive to it when given an explanation: they respond more to strong signals than to weak ones. However, the imprecision makes this response imperfect, leading to an attenuation pattern relative to the benchmark of observing signal strength *without noise*: receivers tend to overinfer from weak explanations and underinfer from strong explanations (as in Augenblick et al., 2025).

Appendix Figure A1 illustrates the receiver's predicted belief update as a function of the orator's signal strength, for *Choice Only* as well as high- and low-noise *Explanation*. All curves approximately intersect at the average signal strength $\mathbb{E}(S)$, mirroring Figure 1a in Augenblick et al. (2025). Crucially, as signals about signal strength become more informative, the slope of the curve steepens while their level at $\mathbb{E}(S)$ stays approximately constant, so that the curves pivot.

We therefore refer to this basic Bayesian effect of explanations in the model as the *pivot* effect. To the extent that explanations carry some information about true signal strength, the *Explanation* treatment makes receivers more sensitive than in *Choice Only* to the true knowledge that is shared, which translates into aggregate improvements.

We consider two potential deviations from this Bayesian framework that may contribute to the treatment effect. First, explanations could uniformly *shift* belief movement, e.g., by causing receivers to imitate correct answers more irrespective of the underlying signal strength. Such a "boost" to the perceived signal strength of correct answers in *Explanation* could account for the asymmetric treatment effect we observe. Second, explanations could affect the weight receivers put on their own *prior*, e.g., by causing receivers to neglect it more or less when listening to an explanation. Depending on average accuracy of priors in the population, this, too, could contribute to an aggregate treatment effect.

Our model offers a structured framework to think about belief updating from explanations, and helps us identify three channels that govern their effect on optimality: *pivot*, *shift* and *prior*.

Estimation. We propose to estimate our model via a canonical Grether (1980) regression on log-odds beliefs. In particular, Proposition 4 in Appendix A shows that a regression of the Bayesian update on orator signal strength S captures the pivot mechanism, i.e., has a level at $\mathbb{E}(S)$ that stays constant between *Choice Only* and *Explanation* with different levels of noise. In turn, this allows us to decompose the average treatment effect (ATE) into the *pivot*, *shift* and *prior* components.

While our previous analyses of treatment effects where based on binary imitation data, this analysis requires belief data. Without loss of generality, we define a respondent's belief as the subjective probability they assign to the correct answer. We recover beliefs by leveraging our confidence data. We convert it to belief probabilities $\pi \in [0,1]$ by mapping confidence from 0% to 100% to $[0.5,1]$ for correct answers. Conversely, confidence from 100% to 0% is mapped to $[0,0.5]$ for incorrect answers.[18] Finally, we transform belief probabilities into belief log-odds $\ell = \text{logit}(\pi) \in \mathbb{R}$.[19] Our key outcome is then log-odds assigned to the correct answer ℓ, with $\ell > 0$ for correct choices, $\ell < 0$ for incorrect choices, and larger $|\ell|$ reflecting greater confidence.

It should be stressed that, unlike in controlled belief updating tasks, the true signal strength is unobservable in our setting, and instead estimated from participants' confidence statements. In the questions we study, higher confidence is predictive of accuracy: prior confidence is 61% on average for incorrect receivers and 72% for correct ones. However, in line with typical experimental patterns of overconfidence (see, e.g., Moore and Healy, 2008), our data also show that confidence is far less predictive of optimality than implied by the logit link structure.

In our model, we account for this by assuming that respondents' beliefs inflate their signal strengths as $\ell = \omega \cdot d \cdot S$, with ω a common overconfidence parameter. In our data, we estimate $\omega \approx 10$, reflecting significant overconfidence. Since this is a linear re-scaling, the model-derived decomposition that applied to the Bayesian update carries over directly to belief movement. Importantly, the pivot for the empirical decomposition is then $\mathbb{E}(|\ell|)$, a very simple, model-free sample moment. This robustness reflects the fact that our decomposition only relies on pivots and shifts in the updating function across treatments, not on its overall shape or scale.

Moreover, the confidence statements underlying our approximations are possibly subject to noise, which can cause attenuation bias in our analysis (see, e.g., Gillen et al., 2019). In

[18] Robustness checks using alternative mappings yield qualitatively identical results. In tasks with three options, the two wrong options are effectively grouped. Dropping cases where a receiver with a wrong answer hears an orator with a different wrong answer likewise has no effect (cf. Tables A1–A2).

[19] To avoid dropping 0 and 1 beliefs, we winsorize probabilities to their 5th and 95th percentiles.

our estimation, these concerns forbid strong conclusions about over- vs. under-inference, i.e., statements suggesting a comparison to a normative Bayesian benchmark. Instead, we focus on comparisons of more vs. less inference *across treatments*.

Results. Our Grether regressions are shown in Appendix Table A1, with our main specification in column (3). It shows receiver posteriors regressed on orator beliefs and receiver priors, all interacted with orator optimality and a dummy for the *Explanation* treatment. Since beliefs are truth-signed log-odds, and we find an ATE of +0.185, explanations increase posterior optimality. This belief–based analysis replicates the core choice–based findings: explanations generate an asymmetric treatment effect, with little impact when orators are incorrect but substantial gains when they are correct. This lends credibility to the use of confidence–based beliefs as an outcome.

Turning to the decomposition, we document three main findings. First, the *pivot* effect—reflecting the Bayesian response to imprecise information about signal strength as in Augenblick et al. (2025)—accounts for 20% of the treatment effect. Explanations thus make receivers somewhat more sensitive to our proxy for signal strength, though the effect is modest in size. Second, the *prior* effect contributes –36%: explanations make receivers less sensitive to their own priors, which reduces optimality given that priors are, on average, informative. Finally, the *shift* effect accounts for 116% of the treatment effect: correct explanations are uniformly more likely to be imitated regardless of orator confidence. This large boost dominates the decomposition and effectively explains the entire aggregate treatment effect.

To illustrate the findings on belief movement (pivot and shift), Figure 4 plots receiver belief movements, i.e., the change in log-odds between a receiver's prior and posterior answer, as a function of orator beliefs, also expressed as log-odds. Lines are fitted using column (3) of Appendix Table A2, which repeats our analysis but for belief movement, by subtracting the prior from the posterior on the left-hand side instead of using it as a covariate on the right-hand side. Dots represent a bin-scatter and show that the prediction lines up quite well.

It makes clear that the belief–based analysis replicates our main choice–based result. In *Choice Only*, receiver beliefs move by moderate, offsetting amounts when the orator is correct and incorrect, while in *Explanation* they shift far more when the orator is correct, driving an average treatment effect. Beyond replicating this asymmetry, the Figure also illustrates its decomposition. The slope difference between treatments reflects the *pivot* effect—receivers become more responsive to signal strength when given an explanation. At the same time, the vertical displacement between treatments for correct orators captures the large *shift* effect: even the least confident

correct explanations cause more belief movement than choices. Together, these patterns visually confirm the decomposition into a modest pivot effect and a dominant shift effect.

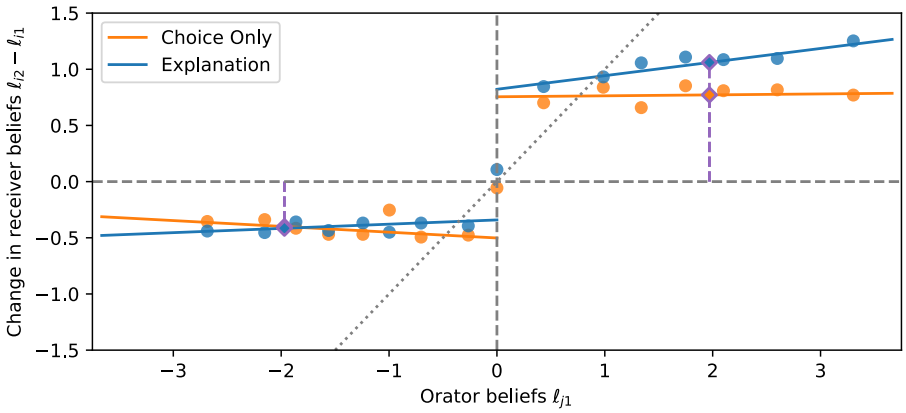


Figure 4: Change in receiver beliefs $\ell_{i2} - \ell_{i1}$ as a function of orator beliefs ℓ_{j1}. *Notes:* Lines plot results from column (3) in Table A2 and dots the corresponding bin-scatter. Purple dashed lines show $\pm\mathbb{E}|\ell_1| \approx \pm 1.97$ and diamonds the predicted change in beliefs.

Taken together, our model-guided analysis shows that explanations make listeners more sensitive to the orator's signal strength in all cases, and that, more importantly, correct answers benefit from a substantial discrete "boost" to perceived signal strength. This suggests there is additional informational content embedded in correct explanations, over and above their role in transmitting belief strength.

Before turning to our analysis of the substance of explanations in Section 5, we validate this conclusion with two additional experiments. In particular, we examine whether the learning benefit of explanations can be replicated with a one-dimensional numerical confidence signal, or whether it instead reflects the unique features of verbal argumentation.

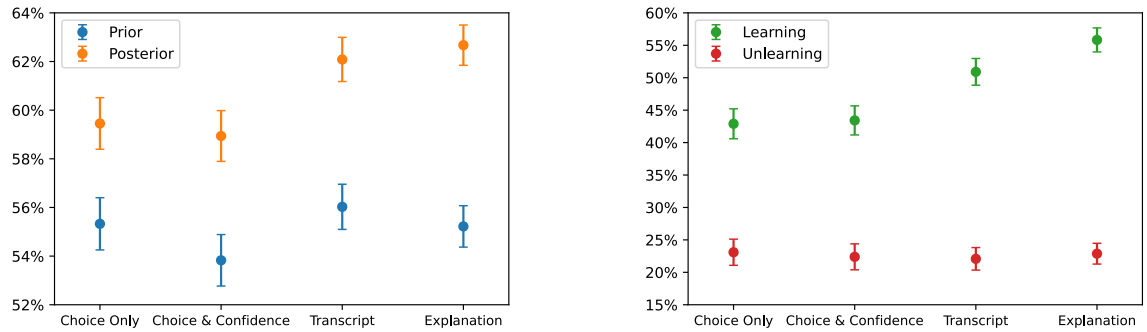
4.2 Are Explanations Equivalent to Numerical Confidence Statements?

To answer this question empirically, we conduct an additional Receiver experiment that allows us to benchmark the effect of explanations against directly observing the orator's confidence.

Design. This additional experiment closely follows the baseline Receiver experiment and also relies on the orator data collected in the baseline Orator experiment. Condition *Choice Only* is identical to its baseline version. Condition *Choice & Confidence* is identical to *Choice Only* except that the listener also sees the level of the orator's stated posterior confidence, a number between 0% and 100%. Example screens from this experiment are provided in Online Appendix Figure

K20. In each task, we randomly assign respondents to the *Choice Only* (20%) treatment or the *Choice & Confidence* treatment (80%).

Logistics. This experiment was run with 860 U.S. respondents in January 2024, of whom 713 provided valid responses.



(a) Prior and posterior optimality by treatment

(b) Imitation in learning and unlearning

Figure 5: Effect of *Choice & Confidence* and *Transcript* on optimality and imitation rates. *Notes:* Left panel shows share of correct receivers before and after exposure to the orator's choice, confidence & choice, explanation transcript or explanation speech. Right panel shows share of receivers picking the same answer as the orator in learning and unlearning. Samples are the corresponding Receiver experiments for *Explanation* (1,103 receivers, 13,111 obs.), *Choice & Confidence* (713 receivers, 8,522 obs.) and *Transcript* (917 receivers, 10,964 obs.), while *Choice Only* is pooled from all collections (2,733 receivers, 8,232 obs.). Whiskers show 95% CIs.

Results. We compare the treatment effect of the *Choice & Confidence* treatment on optimality and imitation rates to the treatment effect of *Explanation*. The results are visualized in Figure 5. Panel (a) shows that the *Choice & Confidence* treatment also induces a substantial, 5.1 p.p. ($p < 0.01$) improvement of the average optimality rate. Yet, adding confidence does *not* create a significant treatment effect on the posterior optimality rate, which is at 58.9% in *Choice & Confidence* compared to 59.5% in *Choice Only* ($p = 0.50$).[20] Results for optimality across treatments are summarized by regressions in Appendix Table E8. Turning to imitation, we find that *Choice & Confidence* has virtually no effect on both the learning and unlearning rate. At 22.4% and 23.1%, respectively, the unlearning rates of *Choice & Confidence* and the control *Choice Only* are

[20]The prior optimality rate in *Choice & Confidence*, at 53.8%, lies marginally below that in *Choice Only* at 55.4% ($p = 0.05$) and *Explanation* at 55.2% ($p = 0.05$). We attribute these small differences to sampling noise across data collections. Our conclusion of a non-significant treatment effect in *Choice & Confidence* also holds when analyzing the improvement (difference between posterior and prior, thereby accounting for the variation in priors), instead of the posterior optimality rate ($p = 0.10$).

virtually identical ($p = 0.61$). Learning rates are similarly close at 43.4% and 42.9% ($p = 0.75$), especially in comparison to the 55.8% learning rate in *Explanation*. Appendix Table E9 regresses imitation on all treatments and learning.

From the absence of treatment effects in *Choice & Confidence* we conclude that explanations operate differently from merely conveying a quantitative signal of the orator's confidence. There are various possible reasons. For example, explanations can convey information above and beyond a confidence level: they may provide objective justifications for an answer that the listener can evaluate independently. This confidence result suggests that the high-dimensional nature of messages may play an important role, motivating our mechanism analyses in Section 5.

Result 2. *The effect of explanations on social learning differs from that of merely observing a sender's confidence. Unlike explanations, confidence observations (i) do not have a treatment effect on the optimality rate, and (ii) do not affect learning and unlearning rates.*

4.3 The Effect of Explanations: What Is Said Versus How It Is Said

The drivers of the treatment effect can be broken down into two factors (e.g., Mehrabian, 1971): the content of the explanation (*what* is said) and its delivery through the speaker's voice (*how* it is said). This distinction matters because it speaks to whether the benefits of explanations for social learning are likely to be limited to oral conversations or may occur similarly, for example, in written exchanges. To distinguish the effects of content and delivery, we run an additional experiment in which receivers read the transcript of an explanation, instead of listening to it. This effectively shuts down the effect of oral delivery, while keeping the content channel constant. This treatment can be thought of as lowering the dimensionality of the message space, while keeping the substantive content identical.

Design. We transcribe the explanations from our Orator experiment in a way that preserves the nuances of the spoken text, including filler words such as "um" and "eh". The design is identical to our baseline Receiver experiment except that *Explanation* is replaced with a *Transcript* treatment, in which participants read the transcript of a recording rather than listening to it. In each task, we randomly assign respondents to *Choice Only* (20%) or *Transcript* (80%). To keep the *Explanation* and *Transcript* treatments as comparable as possible, the text is displayed progressively. Example screens are provided in Online Appendix Figure K21.

Logistics. This experiment was run with 1,266 U.S. respondents in January 2024, of whom 917 provided valid responses.

Results. Panel (a) of Figure 5 shows that explanation transcripts also strongly increase optimality rates relative to the *Choice Only* condition, with nearly similar effect sizes as the corresponding voice recordings. *Transcript* induces a posterior optimality rate of 62.1%, significantly above *Choice Only* (59.5%, $p < 0.01$) and not significantly different from *Explanation* (62.7%, $p = 0.37$). Looking at improvements, which net out the minor across-treatment differences in the prior optimality rate, *Transcript* produces a 2.0 p.p. ($p < 0.01$) larger increase from prior to posterior than *Choice Only*. This corresponds to approximately 59% of the size of the additional improvement in *Explanation* (at 3.4 p.p.), a significant difference ($p < 0.01$, cf. Table E8).

Panel (b) shows that a strong asymmetric effect between learning and unlearning also emerges in the *Transcript* treatment. While transcripts have a strong effect of 8.1 p.p. on the learning rate (at 50.9%, $p < 0.01$), the unlearning rate is virtually unaffected relative to *Choice Only*, at 22.0% ($p = 0.43$). The size of the treatment effect of *Transcript* on the learning rate corresponds to 62% of the treatment effect in *Explanation*, also a significant gap ($p = 0.03$, cf. Table E9). This evidence shows that listening to a spoken explanation leads to somewhat more imitation than just reading the same explanation in learning opportunities, though the asymmetric treatment effects qualitatively emerge in both *Transcript* and *Explanation*. The gap between these treatments highlights that the dimensionality of the message space also matters for social learning.[21]

Result 3. *Both substantive content features and the oral delivery of explanations matter for social learning, with content driving the majority of the effect.*

5 The Supply and Interpretation of Explanations

Our findings on explanations so far rely on experimental variation rather than on the content of the explanations themselves. In this section, we open this black box to better understand why explanations increase imitation in learning opportunities but do not decrease it in unlearning ones. We organize the analysis around an annotation of domain-general features, substantive arguments and an aggregate measure of richness (Section 5.1), discuss content-analysis results (Sections 5.2 and 5.3), and conclude by summarizing additional results on the role of respondent characteristics (Section 5.4) and the dimensionality of communicable information (Section 5.5).

[21]Note that the *Transcript* treatment still relies on text that was originally produced in *spoken* format. In Section 2 we reviewed the systematic differences between written and spoken text production.

5.1 Dissecting Explanations

To study the role of explanations' content, we examine the transcripts of the recordings. Such data are difficult to analyze comprehensively because language is high-dimensional: each sentence and its oral delivery have innumerable features and possible interpretations.

Our analysis follows a two-pronged approach based on the following distinction: on the one hand, explanations are characterized by the substantive content of the answer to a question. Specifically, explanations frequently invoke *arguments*, defined as a series of statements with the purpose of establishing a conclusion. This content tends to be domain-specific, i.e., it directly relates to a specific question and the chosen answer. On the other hand, an explanation is also characterized by text features: it exhibits speech and text metrics, markers of certainty and linguistic features, etc. We think of such features as domain-general, i.e., applying to explanations for different questions and answers. These two ways of analyzing the data provide complementary perspectives that may shed light on central questions, such as whether imitation is affected more by the substantive content of an explanation or by specific linguistic features.

5.1.1 Argument Annotation

Identifying arguments. To identify arguments from the unstructured text data, we develop a coding scheme detailed in Appendix B. First, we provide a state-of-the-art Large Language Model (LLM), OpenAI's GPT-4, with all explanations for a given task and make it identify all distinct arguments. This extraction encompasses any type of argument: not only valid or sound ones, but also fallacious and irrelevant ones. Second, based on the initial list of arguments identified by the LLM, we manually fine-tune the categories, e.g., to avoid duplicates or distinguish between variants. Third, a team of six graduate-level research assistants annotates 100 responses in each of the different tasks; whenever they encounter arguments not captured by our scheme, we add them to it. This yields the final list of arguments.

Annotating explanations. The team of research assistants then annotates the presence of all arguments in the scheme across all 6,910 explanations. To assess the quality of the main annotation, the manual annotation is then performed again, with each task allocated to a new research assistant who is blind to the previous results. Inter-rater reliability is high. If one coder identifies a specific argument, there is a 72% chance the other coder does so as well. If one coder does not identify an argument, there is a 95% chance the other coder does not either. Cohen's κ is 0.67, indicating "substantial agreement" (Cohen, 1960; Landis and Koch, 1977).

Inter-rater reliability is even higher for argument categories. When one coder identifies "any argument", there is a 100% chance that the other coder does so as well; when one coder does not identify "any argument", there is a 0% chance the other coder does so as well. These chances are 82% and 89% for "any fallacious argument" and 83% and 89% for "any sound argument". As a second test, we performed the annotation again using GPT-4. When a human coder identifies an argument, there is a 79% chance GPT-4 does so as well. If a human coder does not identify an argument, there is a 90% chance GPT-4 does not either. Cohen's κ is 0.61, again indicating "substantial agreement". Agreement on categories is similarly high. We view these benchmarks as validating our annotation approach.

5.1.2 Feature Annotation

The second part of our approach leverages 31 domain-general text features that apply across all explanations, independent of the specific question or answer. Using GPT-4, we annotate 25 features drawn from prior work. These include explicit uncertainty markers (modal verbs such as could, might; epistemic stance markers such as I believe; hedges such as probably, perhaps), relative and absolute language (almost, always), and certainty references (definitely, certainly). We also capture implicit markers of uncertainty (hesitations, filled pauses, repetitions, self-corrections), as well as mentions of sources, personal experience, authority, direct addresses, and apologetic phrases. In addition, we compute six simple text and speech metrics, including word count, delivery speed, and the Flesch–Kincaid complexity score. Appendix Table B3 provides an overview of all features.

5.2 The Content of Explanations

We first discuss the descriptive results from our annotation approach. We then estimate how content differences affect imitation rates among receivers. We begin with task-specific arguments before turning to the domain-general text features of explanations.

5.2.1 Arguments: The Substance of Explanations

Illustration of results: Actively managed funds. To illustrate the argument annotation, consider the results on *Actively managed funds*, shown in Figure 6.[22] The left-hand panel shows the frequency of a given argument separately for the sample of explanations for correct and

[22]Analogous results for all other tasks are provided in Appendix Figure D8. The wording of the questions and answer options can be found in Appendix Table E6.

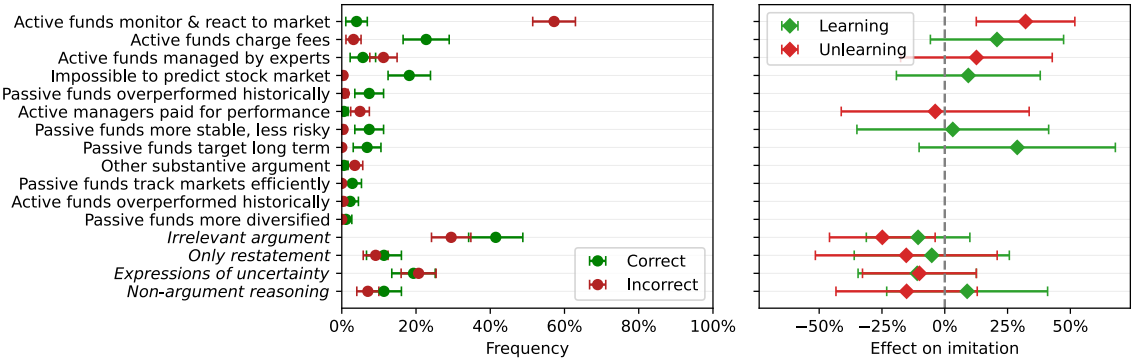


Figure 6: Argument frequency and effects for *Actively managed funds* task. *Notes:* Left panels show the frequency arguments in correct and incorrect explanations. Sample is the Orator experiment. Right panels show the difference-in-differences of the imitation rate between *Explanation* and *Choice Only*, between explanations with and without the argument, in learning and unlearning situations. Only arguments appearing in more than 5% of corresponding explanations are shown. *Explanation* sample is the main Receiver experiment, *Choice Only* is pooled from all collections. Whiskers show 95% CIs, with standard errors clustered at the orator and receiver levels.

incorrect answers, ordered by total frequency in our data. The bottom of the panel also shows the frequency of irrelevant arguments, pure restatements of the answer, any expressions of the speaker's certainty, and any non-argument reasoning. The right-hand panel displays the estimated effect of a given argument on the likelihood of imitation, separately for learning and unlearning situations. To ensure sufficient statistical power, we only show effects for arguments occurring in at least 5% of explanations in the corresponding situation.

The dominant argument is that active funds can adapt quickly to market changes—common in incorrect explanations (57.2%) but rare in correct ones (4.0%). By contrast, the claim that active funds charge higher fees is frequent in correct explanations (22.7%) and nearly absent in incorrect ones (3.2%). A third theme highlights expert management (skewed toward incorrect answers), while a fourth stresses market unpredictability (≈20% of correct, absent from incorrect). Irrelevant arguments are frequent, somewhat more so for correct (41.5%) than incorrect (29.5%) answers; restatements occur in about 10%; and certainty markers in about 20%.

In the right-hand panel, adaptability arguments raise imitation in unlearning situations by 32.1 p.p., and fee arguments raise imitation in learning situations by 20.7 p.p.. Irrelevant arguments reduce imitation in both, while restatements increase it in learning but reduce it in unlearning. Uncertainty has no systematic effects.

Variation across tasks. Explanations vary widely across tasks (see Appendix Figure D8). Some tasks generate many distinct arguments, others only a few; in some consensus is elusive, in others a single line of reasoning dominates. Arguments also differ sharply in their persuasive power, sometimes encouraging imitation even when wrong and discouraging it even when correct. These patterns highlight the variability of financial reasoning, but also the difficulty of comparing arguments across tasks, motivating us to construct a systematic way to analyze them.

Categorizing arguments across tasks. To draw more general conclusions about differences between explanations for correct and incorrect choices, we define four domain-general argument categories. First, we code the absence of any argument. Second, we define an argument as irrelevant if the premises are unrelated to the question or its answer, i.e., an argument might be entirely off-topic. Third, an argument might be relevant but fallacious: one or more of the premises are false, or the conclusion is not valid given the premises. Finally, we classify a sound argument as one with true premises and a valid conclusion. In classifying explanations, we embrace the fact that there are often several explanations for the correct answer that are sound. We also include "weakly sound" arguments in the "sound" category where, from a strict logician's perspective, the premises might not quite be sufficient for the conclusion. Online Appendix Table F1 lists all arguments and their categorization in each task.

Analyzing the prevalence of different arguments classes is important as it is *ex ante* unclear whether right or wrong answers are supported by different kinds of arguments. For example, it is in principle possible to give the right answer based on fallacious or irrelevant arguments or to give a wrong answer even if some part of the explanation relies on a weakly sound argument.

To classify each explanation, we treat these four categories as hierarchical: conditional on having any argument, an explanation will be coded based on the category of the "highest-quality" argument it contains. For example, if an explanation contains both a sound and a fallacious argument, we will assign this explanation to the sound bucket. In practice, 91.8% of explanations contain at most one argument type.

The argument gap. The left panel of Figure 7 shows the frequency of different classes of arguments encountered in learning versus unlearning opportunities. A large fraction of explanations in unlearning opportunities contain no (21.8%), irrelevant (22.6%) or, most often, fallacious arguments (51.0%). All three types of arguments are significantly less common in learning opportunities, at 16.0% for none, 17.7% for irrelevant and 10.5% for fallacious arguments, respec-

tively. This means that the three categories of "lower-quality" explanations are more frequent in unlearning situations, with the most pronounced gap in the case of fallacious arguments. Sound arguments, by contrast, are practically absent in unlearning explanations (4.7%),[23] yet they constitute the dominant category in learning explanations (55.8%). We refer to this stark imbalance in the distribution of argument types as the *argument gap*: learning explanations contain "better" types of arguments than unlearning explanations according to this taxonomy.

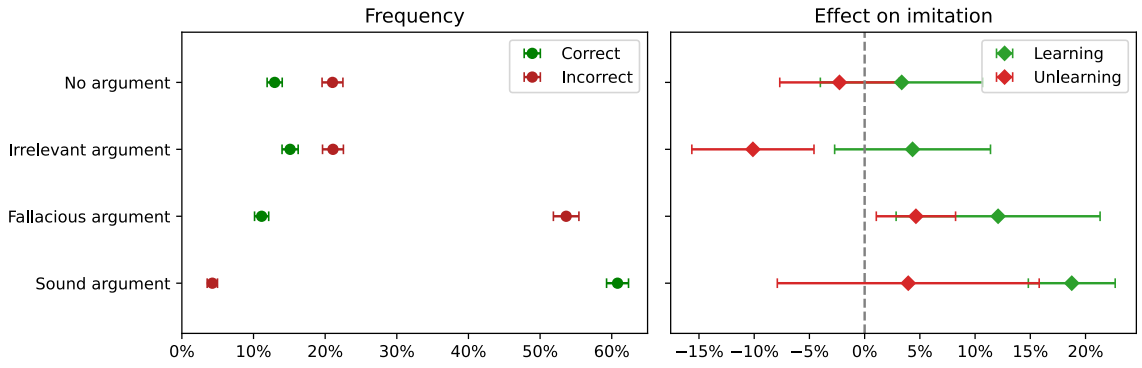


Figure 7: The argument gap. *Notes:* Left panel shows frequency of argument qualities in correct and incorrect explanations. Sample is the Orator experiment. Right panel shows difference in imitation rate between *Explanation* and *Choice Only* by argument quality in learning and unlearning situations. *Explanation* sample is the main Receiver experiment, *Choice Only* is pooled from all collections. Whiskers show 95% CIs.

Argument types and the asymmetric treatment effect. We next examine whether these different classes of arguments are associated with different effects on imitation rates. The right panel of Figure 7 displays the treatment effects associated with each category of argument separately for learning and unlearning situations. Recall that the treatment effect is calculated as the difference in the imitation rates between the *Explanation* treatment and the corresponding matches in the *Choice Only* treatment. We make the following observations.

First, consider treatment effects in unlearning situations. Figure 1b showed a precisely estimated null effect in the aggregate. This could mean that listeners do not adjust their choices in response to explanations, or that the average masks heterogeneity across argument types. Figure 7 confirms the latter: two of four categories exhibit significant effects. Fallacious arguments—which account for more than half of explanations in unlearning—increase the likelihood of

[23]This small occurrence stems from our definition of soundness that includes arguments in which the premises might not strictly be true under all circumstances or only weakly establish the conclusion.

switching to a wrong answer by 4.6 p.p. ($p = 0.01$). By contrast, irrelevant arguments reduce imitation by 10.1 p.p. ($p < 0.01$), while no argument has a small and insignificant negative effect (-2.3 p.p., $p = 0.41$). Thus, the aggregate null conceals sharply offsetting effects across explanation types.

Second, turning to learning situations, we find a positive treatment effect on imitation across all categories. It is strongest in the most common category of sound arguments (18.8 p.p., $p < 0.01$), less pronounced for fallacious arguments (12.1 p.p., $p = 0.01$), and about the same for explanations with no (3.5 p.p., $p = 0.35$) or irrelevant arguments (4.5 p.p., $p = 0.21$). This suggests that, if it supports the right answer, *any* explanation tends to help, but better arguments are more persuasive.

Third, we compare treatment effects across learning and unlearning situations. In every argument category, learning effects exceed unlearning effects, though the size of the gap varies: 14.9 p.p. ($p = 0.02$) for sound arguments, 14.7 p.p. ($p < 0.01$) for irrelevant ones, 7.4 p.p. ($p = 0.14$) for fallacious, and 5.8 p.p. ($p = 0.21$) when no argument is given.

Finally, we combine the heterogeneity results with the argument gap—the differing frequencies of argument types across learning and unlearning—to assess how much of the asymmetric effect is explained by argument composition. Comparing columns 1 and 2 of Table 1, we find that differences in argument categories account for up to 25% of the overall asymmetry.

Why does an asymmetric effect persist across the whole range of argument types? In the following subsection, we investigate whether explanations in learning and unlearning opportunities differ beyond the arguments they contain, for example, in terms of their linguistic richness.

5.2.2 The Features of Explanations

We now turn to our second perspective on the content of explanations. Figure 8 summarizes the results from our annotation of domain-general characteristics of explanations, separately for learning and unlearning situations. The left-hand panel displays the frequency with which each feature occurs in learning and unlearning explanations. We make two main observations.

First, the results confirm a number of intuitions about how explanations for correct and incorrect choices might compare. For example, low certainty markers—indicating low confidence—are more common among unlearning explanations but high certainty markers are more common among learning explanations. Low certainty markers appear more than twice as often overall as high certainty markers, plausibly reflecting that people understand the absence of confidence statements as indicating high confidence. Many features that are plausibly associated with a

Table 1: Decomposition of differential learning effects

	Dependent variable: Imitation						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Explanation	-0.002	0.042**	-0.166***	-0.128***	-0.208***	0.051***	-0.111*
	(0.013)	(0.018)	(0.046)	(0.048)	(0.065)	(0.015)	(0.065)
Learning	0.199***	0.199***	0.200***	0.128***	0.112***	0.207***	0.114***
	(0.017)	(0.021)	(0.018)	(0.016)	(0.021)	(0.018)	(0.021)
Explanation × Learning	0.133***	0.099***	0.091***	0.153***	0.101***	0.054**	0.070***
	(0.021)	(0.026)	(0.022)	(0.021)	(0.026)	(0.022)	(0.026)
Richness						-0.011	-0.007
						(0.009)	(0.010)
Explanation × Richness						0.107***	0.085***
						(0.012)	(0.014)
Argument controls		✓			✓		✓
Orator controls			✓		✓		✓
Receiver controls				✓	✓		✓
Observations	8803	8803	8803	8803	8803	8803	8803
R^2	0.093	0.100	0.107	0.166	0.186	0.112	0.195

Notes: Sample is the main Receiver experiment for *Explanation* and all collections for *Choice Only*, both restricted to learning and unlearning situations. *Explanation* is a dummy for the *Explanation* treatment, *Learning* a dummy for learning situations, *Richness* is standardized explanation richness. All controls contain the variable and an interaction with *Explanation*. *Argument controls* denotes dummies for *No argument*, *Irrelevant argument*, *Fallacious argument* and *Sound argument*. Orator and receiver controls are: *Republican*, *Higher education*, *Black*, *Working*, *Age above 35*, *Male*, *(Prior) Confidence*, *Optimality on all others tasks*. We drop the 0.6% of observations with missing receiver prior confidence from all regressions.

higher quality of explanations, such as empirical statements or indications of sources, are indeed more common in learning explanations.

Do the differences in the features of explanations we document predict imitation? The right panel of Figure 8 estimates the treatment effect of each feature on the imitation rates in learning and unlearning situations.[24] There is substantial heterogeneity in the degree to which specific features are associated with increases or decreases in imitation rates. A higher speaking pace and markers for questions are consistently associated with stronger imitation. Conversely, low certainty markers are correlated with less imitation. A substantial number of features do not significantly predict imitation. Overall, we do not find that the raw features are jointly associated with more or less imitation in learning versus unlearning situations (in a joint F-test, $F = 0.33$ and $p = 0.56$). We only find significant (albeit small) differences in the feature coefficients for

[24]These estimates are obtained from multiple regressions for learning and unlearning situations that include all features, relative to *Choice Only*. We only report estimates for features present in at least 5% of the corresponding explanations.

learning versus unlearning situations in 1 of the 32 variables.

Second, we observe that for the vast majority of features (24 out of 31, or 77.4%), explanations in learning situations exhibit *more occurrences*. Moreover, learning explanations feature higher scores in all of the quantitative text metrics, such as language complexity scores or sentence length. This pattern in the feature analysis may suggest that explanations for correct answers are *richer*, a point we examine systematically in turn.

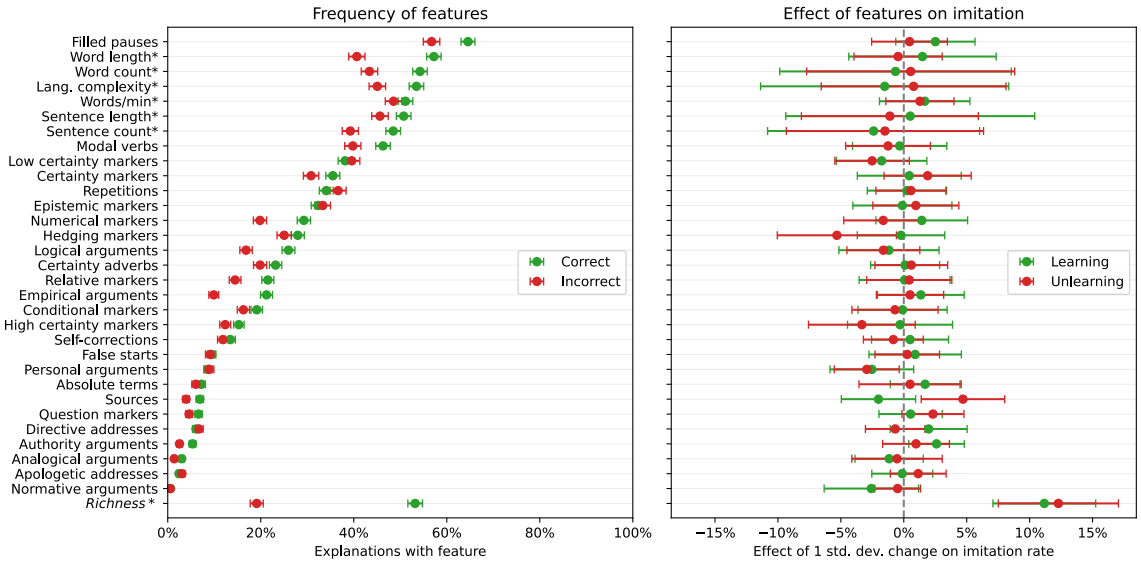


Figure 8: Frequency of explanation features and effects on imitation. *Notes:* Left panel shows share of explanations with features, split by orator optimality. Features with a * instead show the fraction above the overall median. Right panel shows the coefficients on *Explanation* × *Feature* in a multiple regression of imitation on *Explanation*, *Feature* and *Explanation* × *Feature*, for all listed features, applied separately to learning and unlearning. Whiskers show 95% CIs.

5.3 The Richness of Verbal Explanations

We now explore the concept of the richness of a verbal explanation. The motivation is that (i) if explanations in learning opportunities are indeed richer and (ii) richness itself is associated with imitation, the richness gap may partly explain the asymmetric treatment effect of explanations.

5.3.1 The Richness Gap

We set out to characterize the richness of the content of a given message in natural language.[25] We apply the following pre-registered definition in our coding instructions: "A rich explanation

[25]The term richness is used in different disciplines and economics fields, perhaps most commonly to characterize the space of numerical messages in models of communication. While various definitions of

is detailed, comprehensive, logically structured, nuanced, and tailors the argument to fit the context. A sparse explanation is basic, narrow, unclear or disorganized, presents only surface-level understanding, lacks depth or specific details and fails to clearly relate to the context." Our coding approach relies on both human and machine coding, and follows similar procedures as our main annotation (Section 5.1). We obtain richness scores on an 11-point Likert scale, ranging from 0 to 10 (both inclusive).

We document a systematic richness gap: explanations in learning situations score 0.77 SD ($p < 0.01$) higher than in unlearning, and 0.60 SD ($p < 0.01$) after controlling for transcript length. Appendix Figure D5 shows this pattern within each argument category: the gap is 0.28 SD ($p < 0.01$) for none, 0.46 SD ($p < 0.01$) for irrelevant, 0.74 SD ($p < 0.01$) for fallacious, and 0.29 SD ($p < 0.01$) for sound arguments.

What does the richness score capture? In Appendix Figure D7, we systematically investigate which attributes of a text are most predictive of its richness score. Our analysis accommodates a wide array of metrics, spanning lexical features, part-of-speech and phrase structure, readability metrics, syntactic features, complexity and cohesion measures, as well as entity metrics. While many of these features influence richness, we document that *lexical* richness—the complexity, variety, and uniqueness of vocabulary—is a key determinant.

The effect of richness on imitation. The richness score of an explanation is by far the most potent predictor of imitation (Figure 8). A 1 SD increase in richness is associated with 11.3 p.p. ($p < 0.01$) and 12.4 p.p. ($p < 0.01$) increases in imitation in learning and unlearning situations, respectively, after controlling for all other features, including the length of the recording.

Does the richness gap explain the asymmetric treatment effect of explanations? Given that learning explanations are richer across the board and that richness is a strong determinant of imitation, a larger treatment effect of explanations might naturally emerge in learning situations. We examine which fraction of the asymmetric treatment effect is explained by differences in the richness of explanations encountered in learning versus unlearning situations.

Comparing regression analyses in columns 1 and 6 of Table 1 shows that 59% of the asymmetric effect is explained away by the richness gap. Richness remains a powerful determinant of the asymmetric effect once we also account for the role of argument categories. In fact, after

richness are used in the literature, they often relate to the cardinality and/or granularity of the message space as well as the mapping between messages and states. Here, we attempt to characterize the richness of a given message (rather than a hypothetical message space) in natural language.

accounting for richness, the estimated asymmetric treatment effect remains unchanged when additionally controlling for different argument types.

5.3.2 Exogenous variation in richness

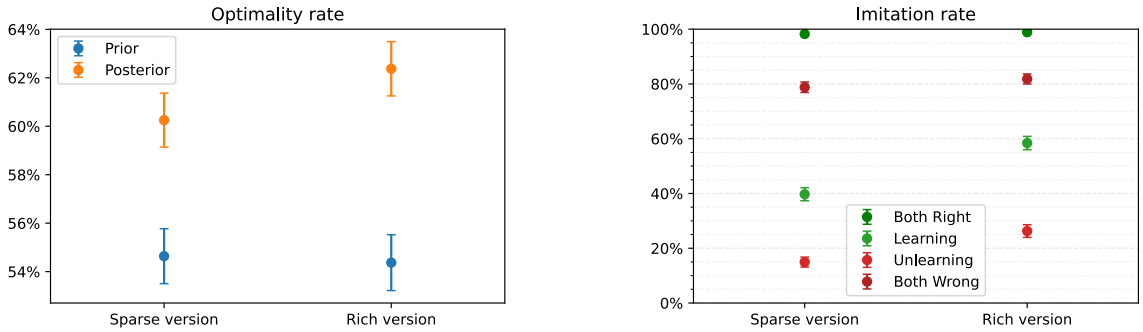
The previous section presents correlational evidence that richness accounts for some of the asymmetric effects between learning and unlearning, even conditioning on the argument gap, and points to the relevance of linguistic aspects of richness. To shed further light on this mechanism, we conduct an additional pre-registered experiment in which we exogenously vary linguistic richness, while holding the arguments and facts of the explanation constant.

Design. The experiment is identical to the *Transcript* Receiver experiment presented in Section 4.3 and relies on the same set of original explanations from the main orator experiment. The key difference is that we randomly assign respondents to either a *Rich version* or a *Sparse version* condition at the task level. We generate rich and sparse versions of each original transcript using an LLM (OpenAI's GPT-4.1), closely following the pre-registered definition of richness that we employed in Section 5.3.1. For the sparse version, the LLM is instructed to mimic conversational uncertainty by employing mid-sentence self-corrections, informal vocabulary, and fragmented sentence structures lacking clear logical progression. Conversely, for the rich version, the model uses precise terminology, explicit logical connectors (such as "because" and "therefore"), illustrative examples, and contextual framing designed to facilitate comprehension. Importantly, the LLM is instructed to preserve key arguments, factual statements, and references to external sources in both versions. Online Appendix I provides additional details, including the prompt used to create the rich and the sparse versions of the original transcript.

Results. As illustrated by Figure 9b, we document a large positive effect of being assigned to the *Rich version* condition on imitation rates. In learning situations, imitation rates increase from 39.7% to 58.4% ($p < 0.01$). In unlearning situations, imitation rates rise from 15.0% to 26.3% ($p < 0.01$). When receivers with a wrong prior receive a confirming transcript, imitation rates increase from 78.8% to 81.8% ($p = 0.03$), and remain largely unchanged at 98.2% and 98.9% when both orator and receiver are correct ($p = 0.06$). Finally, the overall optimality rate increases from 60.3% in the *Sparse version* to 62.4% in the *Rich version* ($p = 0.01$).

Robustness. A key assumption required to isolate the effect of linguistic richness is that we successfully hold argument content constant. To assess this, we employ an LLM to compare the argument content of rich and sparse transcripts and find substantial agreement. As a conservative

robustness check, we restrict the sample to transcripts with exact argument overlap and obtain significant results of similar magnitude to those in the full sample. Online Appendix I provides further evidence on the features and richness of the LLM-generated versions.



(a) Prior and posterior optimality

(b) Imitation

Figure 9: Effect of richness manipulation on optimality and imitation. *Notes:* Sample is the Richness receiver survey (972 receivers, 14,580 obs.). Whiskers show 95% CIs.

Effect size. To better interpret the effect sizes, we start by rating the richness of each version with an LLM just as for the original explanation (cf. Section 5.2.2). On a scale from 0 to 10, the richness of original explanations was 3.1 on average, while it is 2.9 for *Sparse version* ($p < 0.01$) and 5.9 for *Rich version* ($p < 0.01$). This implies rich versions are 1.71 SD (of original richness) richer than sparse versions. For comparison, the richness gap between correct and incorrect original explanations is 0.77 SD.

In turn, we find that across learning and unlearning situations, the *Rich version* condition increases imitation by 15.4 p.p. ($p < 0.01$). We can interpret the treatment assignment as an instrumental variable for richness, which implies that an experimentally induced 1 SD increase in linguistic richness has a $15.4 / 1.71 = 9.0$ p.p. ($p < 0.01$) effect on imitation. This is about two thirds of the 13.0 p.p. increase associated with naturally occurring variation in richness among original explanations.

Result 4. *Learning explanations are richer and contain higher quality arguments than unlearning ones. Richness is the key predictor of imitation and explains about 60% of the asymmetric effect. An additional experiment underscores that linguistic richness is causally related to imitation.*

5.4 The Role of Orator and Receiver Characteristics

We complement the content analyses by analyzing whether differences in orator and receiver characteristics help explain the asymmetric effect of explanations. This asymmetry reflects heterogeneity across an endogenous margin: orators in learning and receivers in unlearning situations have correct priors, while the reverse hold incorrect priors. One possibility is that individuals with correct priors are systematically different—e.g., more influential as orators or less responsive as receivers. Online Appendix H shows, however, that demographic differences across these groups do not account for the asymmetry once content differences are taken into account.

5.5 The Dimensionality of Communicable Information

Our evidence comes from tasks in which explanations convey high-dimensional information, combining facts and arguments in applied, real-world contexts. The central result is that the richness of explanations is crucial for their effectiveness. A natural question is whether explanations also matter when communicable information is low-dimensional. In other words, is high-dimensional communicable information a necessary precondition for explanations to matter?

In the final step of our analysis, we tackle this question by extending our analysis to an unfamiliar, abstract problem: a balls-and-urns belief-updating task (Edwards, 1968). In this task, the signal is plausibly one-dimensional and there might be little real-world knowledge or arguments to bring in. Under the hypothesis that there is indeed little high-dimensional information to communicate, explanations may be less effective.[26]

Design. We design a standard balls-and-urns task. Two bags contain colored balls: one holds 80% blue and 20% red, the other 20% blue and 80% red. One bag is randomly selected, with the prior probability of choosing the blue-majority bag set to either 40% or 60% (varied across participants). A single ball is then drawn from the chosen bag, and its color is revealed to the participant.[27] The participant's task is to state the likelihood with which they think the bag with more blue balls was selected. In the Orator experiment, participants are given instructions on this task with an example. They record an explanation, select their answer on a continuous scale

[26]An examination of explanations in the Balls-and-urns experiment shows that they typically contain similar, largely intuition-based arguments about inference, reflecting a more low-dimensional message space compared to our financial reasoning tasks.

[27]We chose these parameters so that (i) the two possible priors (40% or 60%) sit symmetrically around 50% and have roughly similar space to move on either side; (ii) the diagnosticity of 80% yields a strong signal, yet not so extreme that responses hit ceiling or floor effects; and (iii) limiting the evidence to one draw keeps the task one-dimensional while still illustrating a non-trivial Bayesian update.

from 0% to 100% and state their confidence. We choose a continuous outcome measure rather than a discrete one (as in our other main tasks) in order to keep our implementation of this task comparable to the experimental literature.

In the Receiver experiment, a separate sample of participants is given instructions of the task with an example. They first state their prior responses and their confidence. They then either (i) learn about another respondent's choice (*Choice Only* treatment) or (ii) *additionally* listen to a verbal explanation behind the choice (*Explanation* treatment). Subsequently, respondents again state their choice and confidence. Receivers have a 40% probability of being assigned to the *Choice Only* treatment and a 60% probability of being assigned to the *Explanation* treatment.

Results. We compare the prior and posterior accuracy rates between the *Choice Only* and *Explanation* treatments. To account for stratified assignment, we recover the population average treatment effect by reweighting observations according to their expected frequency under fully random assignment. Following our main results, we examine imitation separately for learning and unlearning opportunities. Online Appendix Figure J13 documents muted effects of explanations on various metrics of imitation: an indicator for no imitation, an indicator for full imitation and a continuous imitation measure. We find some evidence that receiving an explanation somewhat increases the likelihood of full imitation, but this effect does not differ across learning and unlearning conditions. Online Appendix J provides additional details on both the design and analysis of the balls-and-urns task. Taken together, our findings are consistent with the idea that explanations are less effective in settings where communicable information is more low-dimensional.

6 Discussion and Conclusion

We examine how explanations influence the propagation of truths and falsehoods in the context of 15 financial decision-making problems. In our first experiment, participants record an explanation for each of their answers with incentives for the accuracy of their listeners' responses. In a second experiment, a separate set of respondents either only observes an orator's choice on a question or also hears one of the over 6,900 verbal explanations before potentially updating their own decisions. We find that explanations increase aggregate optimality. Notably, this improvement is entirely driven by the greater spread of truths, whereas falsehoods do not become less contagious. Our model-guided decomposition indicates that the treatment effect is driven by a uniform increase in the credibility of explanations for correct answers. A comprehensive

analysis of underlying mechanisms reveals that explanations for truths contain fewer fallacious and more sound arguments, and are far richer than explanations for falsehoods. These richness differences account for approximately 60% of the asymmetric treatment effect. An additional experiment that exogenously varies the linguistic richness of messages, holding the underlying arguments constant, points to a causal interpretation.

Future directions. The evidence in this paper may be extended in various directions. We find that explanation richness is correlated with truth, which is a central relationship underlying the overall beneficial effect of explanations. This relationship may be specific to explanations in settings with aligned incentives. Our setup might fruitfully serve as a blueprint for studying analogous patterns in the case of *persuasive messages*, where the orator wants the receiver to take a specific action. In the case of persuasion, it is conceivable that the richness-truth association in the supply of arguments weakens, or, in some situations, even reverses.

Finally, our LLM-based annotation approach provides a blueprint for analyzing high-dimensional language data to analyze mechanisms underlying economic decisions. The two-step identification of arguments, the extraction of features and the measurement of richness could all be applied to field settings, such as social media posts explaining financial decisions. This opens up valuable opportunities to validate experimental findings in more naturalistic contexts.

References

Akçay, Erol and David Hirshleifer, "Social finance as cultural evolution, transmission bias, and market dynamics," *Proceedings of the National Academy of Sciences*, 2021, *118* (26), e2015568118.

Akinnaso, F Niyi, "On the differences between spoken and written language," *Language and speech*, 1982, *25* (2), 97–125.

Ambuehl, Sandro, B Douglas Bernheim, Fulya Ersoy, and Donna Harris, "Peer Advice on Financial Decisions: A case of the blind leading the blind?," *Review of Economics and Statistics*, 2022, pp. 1–45.

Amelio, Andrea, "Social Learning, Behavioral Biases and Group Outcomes," *Working paper*, 2024.

Andre, Peter, Ingar Haaland, Christopher Roth, Mirko Wiederholt, and Johannes Wohlfart, "Narratives about the Macroeconomy," *Review of Economic Studies*, 2025.

Atkinson, Adele and Flore-Anne Messy, "Measuring Financial Literacy: Results of the OECD/International Network on Financial Education (INFE) Pilot Study," *OECD Working papers on Finance, Insurance and Private Pensions*, 2012, *15*, 1.

Augenblick, Ned, Eben Lazarus, and Michael Thaler, "Overinference from weak signals and underinference from strong signals," *The Quarterly Journal of Economics*, 2025, *140* (1), 335–401.

Barron, Kai and Tilman Fries, "Narrative persuasion," *Working paper*, 2024.

Bordalo, Pedro, John Conlon, Nicola Gennaioli, Spencer Kwon, and Andrei Shleifer, "How people use statistics," *Review of Economic Studies*, 2025, p. rdaf022.

Bourdin, Beatrice and Michel Fayol, "Even in adults, written production is still more costly than oral production," *International Journal of Psychology*, 2002, *37* (4), 219–227.

Bursztyn, Leonardo, Florian Ederer, Bruno Ferman, and Noam Yuchtman, "Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions," *Econometrica*, 2014, *82* (4), 1273–1301.

_ , Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth, "Justifying Dissent," *Quarterly Journal of Economics*, 2023.

Çelen, Boğaçhan, Shachar Kariv, and Andrew Schotter, "An experimental test of advice and social learning," *Management Science*, 2010, *56* (10), 1687–1701.

Cialdini, Robert B, "The science of persuasion," *Scientific American*, 2001, *284* (2), 76–81.

Cohen, Jacob, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, 1960, *20* (1), 37–46.

Conlon, John J, Malavika Mani, Gautam Rao, Matthew W Ridley, and Frank Schilbach, "Not Learning from Others," *NBER Working Paper No. 30378*, 2025.

Duflo, Esther and Emmanuel Saez, "The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment," *Quarterly Journal of Economics*, 2003, *118* (3), 815–842.

Edwards, Ward, "Conservatism in human information processing," *Formal representation of human judgment*, 1968.

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

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

Fehr, Ernst and Jean-Robert Tyran, "Individual irrationality and aggregate outcomes," *Journal of Economic Perspectives*, 2005, *19* (4), 43–66.

Galeotti, Andrea, Sanjeev Goyal, Matthew O Jackson, Fernando Vega-Redondo, and Leeat Yariv, "Network games," *Review of Economic Studies*, 2010, *77* (1), 218–244.

Gillen, Ben, Erik Snowberg, and Leeat Yariv, "Experimenting with measurement error: Techniques with applications to the caltech cohort study," *Journal of Political Economy*, 2019, *127* (4), 1826–1863.

Graeber, Thomas, Christopher Roth, and Florian Zimmermann, "Stories, statistics, and memory," *The Quarterly Journal of Economics*, 2024, *139* (4), 2181–2225.

_ , Shakked Noy, and Christopher Roth, "The Transmission of Reliable and Unreliable Information," *Working paper*, 2025.

Grether, David M, "Bayes rule as a descriptive model: The representativeness heuristic," *The Quarterly journal of economics*, 1980, *95* (3), 537–557.

Grunewald, Andreas, Victor Klockmann, Alicia von Schenk, and Ferdinand A. von Siemens, "Are Biases Contagious? The Influence of Communication on Motivated Beliefs," *Working paper*, 2024.

Haaland, Ingar and Ole-Andreas Elvik Næss, "Misperceived Returns to Active Investing," *Working paper*, 2023.

Hirshleifer, David, "Presidential Address: Social Transmission Bias in Economics and Finance," *Journal of Finance*, August 2020, *75* (4), 1779–1831.

_ , Lin Peng, and Qiguang Wang, "News diffusion in social networks and stock market reactions," *NBER Working Paper No. 30860*, 2023.

Hüning, Hendrik, Lydia Mechtenberg, and Stephanie Wang, "Using Arguments to Persuade: Experimental Evidence," *Available at SSRN 4244989*, 2022.

Jackson, Matthew O and Leeat Yariv, "Diffusion of behavior and equilibrium properties in network games," *American Economic Review*, 2007, *97* (2), 92–98.

Kahneman, Daniel and Amos Tversky, "Subjective probability: A judgment of representativeness," *Cognitive psychology*, 1972, *3* (3), 430–454.

Kendall, Chad W and Constantin Charles, "Causal narratives," *NBER Working Paper No. 30346*, 2022.

Landis, J Richard and Gary G Koch, "The measurement of observer agreement for categorical data," *Biometrics*, 1977, pp. 159–174.

Langer, Ellen J, Arthur Blank, and Benzion Chanowitz, "The mindlessness of ostensibly thoughtful action: The role of placebic information in interpersonal interaction.," *Journal of Personality and Social Psychology*, 1978, *36* (6), 635.

Lazer, David MJ, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild et al., "The science of fake news," *Science*, 2018, *359* (6380), 1094–1096.

Lombrozo, Tania, "The structure and function of explanations," *Trends in Cognitive Sciences*, 2006, *10* (10), 464–470.

Lusardi, Annamaria and Olivia S Mitchell, "Financial literacy and retirement planning: New evidence from the Rand American Life Panel," *Michigan Retirement Research Center Research Paper*, 2007, *157*.

Mehrabian, Albert, *Silent Messages*, Belmont, CA: Wadsworth, 1971.

Moore, Don A and Paul J Healy, "The trouble with overconfidence.," *Psychological review*, 2008, *115* (2), 502.

Oprea, Ryan and Sevgi Yuksel, "Social exchange of motivated beliefs," *Journal of the European Economic Association*, 2022, *20* (2), 667–699.

Schotter, Andrew, "Decision making with naive advice," *American Economic Review*, 2003, *93* (2), 196–201.

_ , *Advice, Social Learning and the Evolution of Conventions*, Cambridge University Press, 2023.

_ **and Barry Sopher**, "Social learning and coordination conventions in intergenerational games: An experimental study," *Journal of Political Economy*, 2003, *111* (3), 498–529.

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

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.

Sloman, Steven A, Bradley C Love, and Woo-Kyoung Ahn, "Feature centrality and conceptual coherence," *Cognitive Science*, 1998, *22* (2), 189–228.

Thaler, Michael, "The supply of motivated beliefs," *arXiv preprint arXiv:2111.06062*, 2025.

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

Appendix for "Explanations"

Thomas Graeber Christopher Roth Constantin Schesch

A Conceptual Framework

We briefly set out a framework that casts our experimental setup as a standard belief formation model that speaks to the existing economics literature. It conceptualizes our reduced-form findings and provides a structure for our mechanism analyses. At the same time, it is not meant to be a micro-foundation of the structure of explanations in natural language and their interpretation.

Prior beliefs and choice. We model a receiver i that has to respond to a question, whose correct answer is modeled as a state $\theta \in \{-1, 1\}$. We assume, without loss of generality, that the true state is always $\theta = 1$.[28] Participants start with an ignorance prior $\pi_0 = \mathbb{P}(\theta = 1) = 1/2$, which in log-odds maps to $\ell_{i0} = \text{logit}(\pi_0) = 0$.

The receiver then thinks about their answer, which we model as a noisy signal about the state. Later on, they will hear from an orator j whose answer comes from the same signal technology. The diagnosticity of posterior beliefs, and their distribution, are then critical objects. We therefore model signals through a log-normal strength combined with a logistic link.

Assumption 1. *The respondent receives a signal comprising a direction $d \in \{-1, 1\}$ and a strength $S \in \mathbb{R}_+$. The strength is log-normal with $\log S \sim \mathcal{N}(\log \bar{S}, \sigma^2)$, while the sign depends on S via:*

$$\mathbb{P}(d = \theta \mid S) = \Lambda(S),$$

with $\Lambda(x) = 1/(1 + e^{-x})$ the logistic function.

Lemma 1. *The log-likelihood ratio (LLR) of observing (d, S) is exactly $d \cdot S$.*

Proof. Let g be the density of S and $f_\pm$ the densities of $d \cdot S$:

$$\log \frac{f_+(d \cdot S)}{f_-(d \cdot S)} = \log \frac{g(S)\Lambda(d \cdot S)}{g(S)\Lambda(-d \cdot S)} = \log \frac{\Lambda(d \cdot S)}{1 - \Lambda(d \cdot S)} = d \cdot S.$$

[28] In tasks with multiple options, we simply map all incorrect options to -1. Columns (5) in Tables A1 and A2 show our results are robust to dropping observations where a receiver with a wrong answer hears from an orator with a different wrong answer.

□

Since respondents start from $\ell_{i0} = 0$, their first choice is simply $d_i = d$. The overall probability of being correct is then:

$$p = \mathbb{P}(d = \theta) = \mathbb{E}[\Lambda(S)].$$

As expected, the optimality rate increases with median signal diagnosticity $\bar{S}$. Correspondingly, their first posterior should be $d \cdot S$. To account for the well-documented overstatement of confidence in experimental data, we assume that respondents inflate their signal strength.[29]

Assumption 2. *Respondents inflate their signal strength by a common knowledge factor ω.*

Stated beliefs are then $\ell_{i1} = \omega \cdot d \cdot S$. Calibration of variance is immediate as $\sigma^2 = \mathbb{V}(\log S) = \mathbb{V}(\log|\ell_1|) \approx 0.61$. Matching the optimality rate $p \approx 55.3\%$ then yields $\bar{S} \approx 0.16$. Finally, we get $\omega = \exp(\mathbb{E}(\log|\ell_1|))/\bar{S} \approx 1.60/0.16 \approx 10$, meaning that subjects report confidence levels that overstate diagnosticity by roughly an order of magnitude.

Learning from choices and explanations After forming their own opinion, the receiver sees the choice or hears the explanation of an orator. We make the following critical assumption.

Assumption 3. *The orator formed their belief ℓ_{j1} via an independent draw from the same signal.*

If the orator could perfectly transmit ℓ_{j1}, the receiver would naturally update to $\ell_{i2} = \ell_{i1} + \ell_{j1}$. However, under *Choice Only*, the receiver observes only the orator's choice d_j.

Proposition 1. *If the orator only transmits their choice d_j, the receiver updates to*

$$\ell_{i2}^{C} = \ell_{i1} + \omega \cdot d_j \cdot \text{logit}(p).$$

Proof. The orator's choice is a binary signal with $\mathbb{P}(d_j = \theta) = p$, so its LLR is $d_j \cdot \text{logit}(p)$. □

In *Explanation*, the receiver additionally hears the explanation behind the orator's choice, which can convey a noisy signal about the orator's confidence. We model this learning from signals with imperfectly known diagnosticity by closely following Augenblick et al. (2025).

[29] One could alternatively model a bias in reported beliefs. This would, however, obscure the interpretation of S, which is strictly positive in our setting. Importantly, the distinction does not affect our empirical results: when outcomes are measured as belief changes before and after an explanation (see Table A2), we obtain qualitatively identical findings, namely that explanations are dominated by a uniform *shift*.

Assumption 4. *Explanations contain a noise-free choice signal d_j and a noisy belief strength signal $\log \tilde{S}_j = \log S_j + \varepsilon$, with disturbances $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ independent of ℓ_{i1} and θ.*[30]

After observing $\tilde{S}_j$, the receiver forms the posterior $\log S \mid \tilde{S}_j \sim \mathcal{N}(\log \bar{S}_{\text{post}}, \sigma^2_{\text{post}})$, with:

$$\log \bar{S}_{\text{post}} = \log \bar{S} + \lambda(\log \tilde{S}_j - \log \bar{S}), \qquad \sigma^2_{\text{post}} = (1-\lambda)\sigma^2, \qquad \lambda = \frac{\sigma^2}{\sigma^2 + \sigma_\varepsilon^2}.$$

Proposition 2. *If the orator transmits an explanation, the receiver updates to*

$$\ell_{i2}^E = \ell_{i1} + \omega \cdot d_j \cdot \text{logit}(\mathbb{E}[\Lambda(S) \mid \tilde{S}_j]).$$

Proof. Let $f_\pm$ denote the densities of $(d_j, \tilde{S}_j)$ under $\theta = \pm 1$. Since ε is state-independent,

$$\log \frac{f_+(d_j, \tilde{S}_j)}{f_-(d_j, \tilde{S}_j)} = \log \frac{g(\tilde{S}_j)\mathbb{P}(d_j \mid \theta = 1, \tilde{S}_j)}{g(\tilde{S}_j)\mathbb{P}(d_j \mid \theta = 0, \tilde{S}_j)} = \log \frac{\mathbb{E}[\Lambda(d_j S) \mid \tilde{S}_j]}{1 - \mathbb{E}[\Lambda(d_j S) \mid \tilde{S}_j]} = \text{logit}(\mathbb{E}[\Lambda(d_j S) \mid \tilde{S}_j]).$$

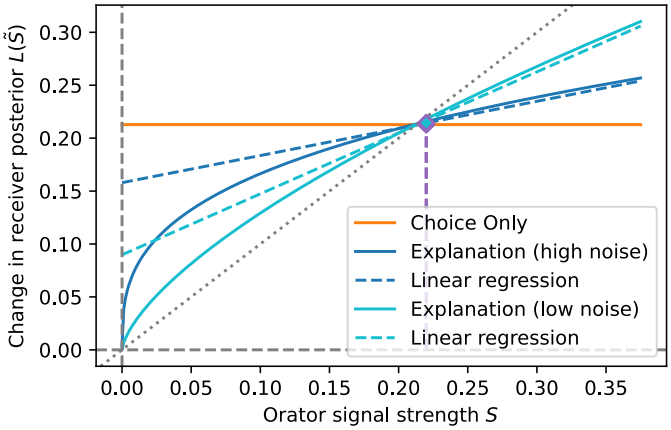
If $d_j = 1$, the result is clear. If $d_j = -1$, this is $\text{logit}(1 - \mathbb{E}[\Lambda(S) \mid \tilde{S}_j]) = -\text{logit}(\mathbb{E}[\Lambda(S) \mid \tilde{S}_j])$. □

Belief movement therefore separates into a perfectly-known direction d_j and an imperfectly-known magnitude $L(\tilde{S}) = \text{logit}(\mathbb{E}[\Lambda(S) \mid \tilde{S}])$. As $\sigma_\varepsilon^2 \to 0$, $\lambda \to 1$ and the receiver perfectly recovers S_j so that the increment tends to $d_j \cdot S_j$, i.e., the noise-free benchmark. As $\sigma_\varepsilon^2 \to \infty$, $\lambda \to 0$ and $\tilde{S}_j$ becomes completely uninformative so that the increment tends to $d_j \cdot \text{logit}(\mathbb{E}[\Lambda(S)])$, i.e., the *Choice Only* benchmark.

Pivot Figure A1 plots the belief update $L(\tilde{S})$, using our calibrated values $\bar{S} \approx 0.16$ and $\sigma^2 \approx 0.61$, for *Choice Only* (i.e., $\lambda = 0$), high-noise *Explanation* ($\lambda = \frac{1}{3}$) and low-noise *Explanation* ($\lambda = \frac{2}{3}$). It is directly comparable to Figure 1a in Augenblick et al. (2025). As they only contain a noisy signal about the orator's belief strength, receivers move too little when hearing a confident explanation, and too much when hearing an unconfident explanation. This reproduces the pattern of over-inference from weak signals and under-inference from strong signals identified in Augenblick et al. (2025).

As signals about belief strength become more precise, the curves' level at their intersection stays approximately constant, while their slope increases, so that the curves pivot. We can characterize this behavior analytically in the small signal informativeness $\bar{S} \to 0$ limit, which is

[30] Note we could equivalently have assumed explanations contain a signal reflecting (over-)stated confidence $\omega \cdot S$. Because ω is common knowledge, the update would simply subtract $\log \omega$.



Appendix Figure A1: Change in receiver beliefs $L(\tilde{S})$ as a function of orator belief strength S. *Notes:* Parameters are $\bar{S} \approx 0.16$ and $\sigma^2 \approx 0.61$, as well as $\lambda = \frac{1}{3}$ for high-noise and $\lambda = \frac{2}{3}$ for low-noise explanations. Purple dashed line indicates $\mathbb{E}(S) = \bar{S}e^{\sigma^2/2}$, blue dashed lines show output of a linear regression of $L(\tilde{S})$ on S.

appropriate in our setting where the optimality rate is 55.3%.[31]

Proposition 3. *As λ increases, the LLR curve $L(\tilde{S})$ pivots approximately around $\mathbb{E}(S)$, in the sense that the effect of λ on the level at $\mathbb{E}(S)$ is third order in $\mathbb{E}(S)$:*

$$L(\mathbb{E}(S)) = \mathbb{E}(S) + \tfrac{1}{12}\mathbb{E}(S)^3(1 - e^{3(1-\lambda)\sigma^2}) + O(\bar{S}^5).$$

Proof. Recall the Taylor expansions around $x = 0$:

$$\Lambda(x) = \tfrac{1}{2} + \tfrac{x}{4} - \tfrac{x^3}{48} + O(x^5), \qquad \text{logit}(\tfrac{1}{2} + x) = 4x + \tfrac{16}{3}x^3 + O(x^5).$$

Plugging these into $\mathbb{E}(\Lambda(S) \mid \tilde{S})$ and $L(\tilde{S}) = \text{logit}(\mathbb{E}(\Lambda(S) \mid \tilde{S}))$ yields:

$$\mathbb{E}[\Lambda(S) \mid \tilde{S}] = \tfrac{1}{2} + \tfrac{1}{4}\mathbb{E}[S \mid \tilde{S}] - \tfrac{1}{48}\mathbb{E}[S^3 \mid \tilde{S}] + O(\bar{S}^5),$$

$$L(\tilde{S}) = \mathbb{E}[S \mid \tilde{S}] + \tfrac{1}{12}\left(\mathbb{E}[S \mid \tilde{S}]^3 - \mathbb{E}[S^3 \mid \tilde{S}]\right) + O(\bar{S}^5).$$

Under the log-normal posterior,

$$\mathbb{E}[S^k \mid \tilde{S}] = \bar{S}^{k(1-\lambda)}\tilde{S}^{k\lambda}\exp(\tfrac{\sigma^2}{2}k^2(1-\lambda)).$$

[31] As shown in Section 3.4, optimality within tasks varies more widely, from 20% to 90%. The small–$\bar{S}$ characterization should thus be interpreted as capturing the average financial reasoning environment.

Evaluating at $\tilde{S} = \mathbb{E}(S)$ gives:

$$\mathbb{E}[S \mid \tilde{S} = \mathbb{E}(S)] = \bar{S}e^{\sigma^2/2} = \mathbb{E}(S), \qquad \mathbb{E}[S^3 \mid \tilde{S} = \mathbb{E}(S)] = \bar{S}^3 \exp(3 \cdot \sigma^2(3/2 - \lambda)),$$

and gathering terms then yields the stated result. □

The first term is $\mathbb{E}(S)$ and independent of λ, while the only λ-dependence is at cubic order and quite small. In Figure A1, the numerical difference in $L(\mathbb{E}(S))$ between $\lambda = \frac{2}{3}$ and $\lambda = \frac{1}{3}$ is indeed only +0.002, while the difference in intercepts $a^\star$ is only +0.001. This explains why the noise-free benchmark, low- or high-noise *Explanations* and *Choice Only* all approximately intersect at $\mathbb{E}(S)$. In turn, we can show that coefficients in a linear regression of $L(\tilde{S})$ on S also reflect and identify this pivot.

Proposition 4. *A regression of the LLR $L(\tilde{S})$ on $\{1, S - \mathbb{E}(S)\}$ yields a slope $b^\star$ and an intercept $a^\star$ where the latter only depends on λ up to third order in $\mathbb{E}(S)$:*

$$a^\star = \mathbb{E}(S) + \tfrac{1}{12}\mathbb{E}(S)^3(e^{3\lambda\sigma^2} - e^{3\sigma^2}) + O(\bar{S}^5).$$

Proof. Since we regress $L(\tilde{S})$ on a constant and a demeaned $S - \mathbb{E}(S)$,

$$a^\star = \mathbb{E}[L(\tilde{S})] = \mathbb{E}[\mathbb{E}[S \mid \tilde{S}]] + \tfrac{1}{12}\left(\mathbb{E}[\mathbb{E}[S \mid \tilde{S}]^3] - \mathbb{E}[\mathbb{E}[S^3 \mid \tilde{S}]]\right) + O(\bar{S}^5).$$

By the tower property, $\mathbb{E}[\mathbb{E}[S \mid \tilde{S}]] = \mathbb{E}[S]$ and $\mathbb{E}[\mathbb{E}[S^3 \mid \tilde{S}]] = \mathbb{E}[S^3]$. The only new moment is:

$$\mathbb{E}[\mathbb{E}[S \mid \tilde{S}]^3] = \mathbb{E}[\bar{S}^{3(1-\lambda)} \cdot \tilde{S}^{3\lambda} \cdot e^{3(1-\lambda)\frac{\sigma^2}{2}}] = \bar{S}^{3(1-\lambda)} \cdot e^{3(1-\lambda)\frac{\sigma^2}{2}} \cdot \bar{S}^{3\lambda} e^{\frac{9}{2}\lambda^2(\sigma^2+\sigma_\varepsilon^2)} = \mathbb{E}[S]^3 e^{3\lambda\sigma^2},$$

where we used $\lambda(\sigma^2 + \sigma_\varepsilon^2) = \sigma^2$. Gathering terms yields the result. □

One can further prove that, as expected, the slope b^* is increasing in λ. We have therefore shown that more informative signals, meaning a higher λ, make the belief update $L(\tilde{S})$ pivot; we have shown that this pivot is approximately at $\mathbb{E}(S)$; and we have shown that a linear regression of $L(\tilde{S})$ on S also reflects this pivot mechanism. In Figure A1, dashed lines display the result of such a regression, confirming that predictions made in the small–$\bar{S}$ limit hold true in our setting.

In our regressions, outcomes are beliefs ℓ directly so that we use $\overline{|\ell|} = \omega \cdot \mathbb{E}(S) = \mathbb{E}(|\ell_1|)$. Our model therefore motivates a test of its key effect in a canonical linear Grether regression,

using only an extremely simple sample moment. Moreover, it justifies our decomposition of the effect of explanations into a Bayesian component and deviations from this benchmark.

Deviations from Bayesian benchmark Our Bayesian model of explanations as noisy transmitters of confidence therefore predicts that, compared to *Choice Only*, belief movement as a function of confidence in the *Explanation* treatment acts as a pivot: low-confidence explanations are imitated less than choices, while high-confidence explanations are imitated more. If there is at least some knowledge to be shared, so that confidence correlates with optimality, this translates into aggregate improvements. More precisely, truth-signed log-odds beliefs should be higher in *Explanation* than in *Choice Only*. We examine deviations from this Bayesian benchmark in two directions: First, explanations could uniformly *shift* belief movement, e.g., by causing correct explanations to be imitated more or less irrespective of confidence. Second, explanations could affect weight put on *priors*, e.g., by causing receivers to neglect their prior more or less when hearing an orator speak. Since we shed light on the asymmetry between correct and incorrect explanations, we allow for different pivots, shifts and prior sensitivities when receivers face correct or incorrect answers.

Estimation We model the posterior belief of receiver i faced with the correct (+) or incorrect (−) choice ($E_i = 0$) or explanation ($E_i = 1$) of orator j as:

$$\ell_{i2} = (\delta^C_\pm + \delta^E_\pm E_i)\ell_{i1} + (\kappa^C_\pm + \kappa^E_\pm E_i) + (\gamma^C_\pm + \gamma^E_\pm E_i)(\ell_{j1} \mp \overline{|\ell|}) + \varepsilon_i.$$

The $\delta^C_\pm$ etc. identify the effect of choices while the $\delta^E_\pm$ etc. identify the additional effect of explanations. The conditional treatment effect of a correct or incorrect explanation is then:

$$\mathbb{E}[\ell^E_{i2} - \ell^C_{i2} \mid d_j = \pm 1] = \delta^E_\pm \bar{\ell}_{i1} + \kappa^E_\pm + \gamma^E_\pm(\bar{\ell}_\pm \mp \overline{|\ell|}).$$

Here, we have defined average beliefs conditional on being correct or incorrect $\bar{\ell}_+ = \mathbb{E}[\ell_{j1}|d_j = 1]$ and $\bar{\ell}_- = \mathbb{E}[\ell_{j1}|d_j = -1]$. Moreover, we call $\bar{\ell}_{i1+} = \mathbb{E}[\ell_{i1}|d_j = 1]$ and $\bar{\ell}_{i1-} = \mathbb{E}[\ell_{i1}|d_j = -1]$ the average conditional receiver beliefs.[32] Recall that $p = \mathbb{P}(d_j = 1)$. The average treatment

[32]Within a task, these are identical in population due to random matching, but may differ in finite samples. Across tasks, they could co-vary due to heterogeneity in optimality rates. We track both to keep the decomposition exact.

effect (ATE) can then be broken down into three components:

$$\begin{aligned}
\text{ATE} &= \mathbb{E}[\ell^E_{i2} - \ell^C_{i2}] \\
&= p(\delta^E_+ \bar{\ell}_{i1+} + \kappa^E_+ + \gamma^E_+(\bar{\ell}_+ - \overline{|\ell|})) + (1-p)(\delta^E_- \bar{\ell}_{i1-} + \kappa^E_- + \gamma^E_-(\bar{\ell}_- + \overline{|\ell|})) \\
&= \underbrace{p\kappa^E_+ + (1-p)\kappa^E_-}_{\text{shift}} + \underbrace{p\gamma^E_+(\bar{\ell}_+ - \overline{|\ell|}) + (1-p)\gamma^E_-(\bar{\ell}_- + \overline{|\ell|})}_{\text{pivot}} + \underbrace{p\delta^E_+ \bar{\ell}_{i1+} + (1-p)\delta^E_- \bar{\ell}_{i1-}}_{\text{prior}}.
\end{aligned}$$

In practice, to avoid opaque variable demeaning, our regression is simply specified as:

$$\begin{aligned}
\ell_{i2} = (\beta + \beta^E E_i) + (\beta_c + \beta^E_c E_i)c_j + (\beta_\ell + \beta^E_\ell E_i)\ell_{j1} + (\beta_{c\ell} + \beta^E_{c\ell} E_i)(c_j \cdot \ell_{j1}) \\
+ (\beta_p + \beta^E_p E_i)\ell_{i1} + (\beta_{cp} + \beta^E_{cp} E_i)(c_j \cdot \ell_{i1}) + \varepsilon_i,
\end{aligned}$$

where $c_j = \mathbb{1}_{\{d_j=1\}}$ is the choice with correct/incorrect coded as $\{1, 0\}$. These coefficients map to linear model parameters as $\gamma^E_- = \beta^E_\ell$, $\gamma^E_+ = \beta^E_\ell + \beta^E_{c\ell}$, $\kappa^E_- = \beta^E - \overline{|\ell|}\beta^E_\ell$, $\kappa^E_+ = \beta^E + \beta^E_c + \overline{|\ell|}(\beta^E_\ell + \beta^E_{c\ell})$, $\delta^E_- = \beta^E_p$ and $\delta^E_+ = \beta^E_p + \beta^E_{cp}$. Note that the decomposition always sums to the ATE, and that the choice of a pivot $\overline{|\ell|}$, which is motivated by our model, only affects its breakdown.

Results are reported in Table A1. Each column re-computes $\overline{|\ell|}$ for the decomposition using its variable coding and sample restrictions, for which we pool orator and receiver priors for increased precision. Our main specification is presented in column (3). It shows the following decomposition of the effect of explanations on truth-signed beliefs:

- The *pivot* effect accounts for 20%: explanations make receivers more sensitive to orator beliefs, which improves optimality since orators are correct on average,
- The *prior* effect accounts for -36%: explanations make receivers less sensitive to their prior, which hurts optimality since receivers are also correct on average,
- The *shift* effect accounts for 116%: correct explanations are uniformly more likely to be imitated, irrespective of orator confidence; moreover, this effect is very large while *pivot* and *prior* effects cancel out, so that it accounts for virtually the entire treatment effect.

Figure A1 illustrates our approach by plotting receiver belief movement as a function of orator beliefs. It effectively replicates Figure A1 with data, separately for correct (on the $\ell_{j1} > 0$ pane) and and incorrect ($\ell_{j1} < 0$) explanations. Lines are fitted using column (3) of Table A2, which repeats our previous analysis but for belief movement, by subtracting the prior from the

Appendix Table A1: Grether regression

	Dependent variable: Posterior log-odds				
	(1)	(2)	(3)	(4)	(5)
Intercept	0.701***	-0.382***	-0.485***	-0.351***	-0.411***
	(0.027)	(0.029)	(0.040)	(0.039)	(0.042)
Explanation	0.168***	0.209***	0.167***	0.100*	0.093*
	(0.031)	(0.041)	(0.054)	(0.055)	(0.056)
Prior log-odds		0.829***	0.888***	0.918***	0.943***
		(0.010)	(0.011)	(0.014)	(0.014)
Explanation × Prior log-odds		-0.096***	-0.010	-0.035*	-0.028
		(0.013)	(0.015)	(0.018)	(0.019)
Orator correct		1.251***	1.298***	1.087***	1.147***
		(0.057)	(0.058)	(0.057)	(0.060)
Explanation × Orator correct		-0.079	-0.030	0.150*	0.157*
		(0.081)	(0.081)	(0.082)	(0.083)
Orator log-odds		0.014	-0.045**	-0.033	-0.042*
		(0.012)	(0.020)	(0.023)	(0.025)
Explanation × Orator log-odds		0.114***	0.092***	0.081**	0.078**
		(0.018)	(0.028)	(0.032)	(0.034)
Orator correct × Prior log-odds			-0.104***	-0.125***	-0.150***
			(0.017)	(0.020)	(0.021)
Explanation × Orator correct × Prior log-odds			-0.151***	-0.140***	-0.146***
			(0.022)	(0.026)	(0.027)
Orator correct × Orator log-odds			0.098***	0.053*	0.062**
			(0.027)	(0.029)	(0.031)
Explanation × Orator correct × Orator log-odds			0.038	0.015	0.017
			(0.036)	(0.040)	(0.042)
ATE	0.168	0.187	0.185	0.168	0.173
% shift	100%	103%	116%	144%	147%
% pivot	0%	19%	20%	12%	10%
% prior	0%	-22%	-36%	-56%	-57%
Observations	21199	21199	21199	13213	12658
R^2	0.001	0.678	0.687	0.698	0.693

Notes: Outcome is the receiver posterior belief in log-odds ℓ_{i2}. *Prior log-odds* is ℓ_{i1}. *Explanation* is a dummy for the *Explanation* treatment, *Orator correct* a dummy for the orator being correct, *Orator log-odds* is the orator's belief in log-odds ℓ_{j1}. *ATE*, *shift*, *pivot* and *prior* report the treatment effect of explanations and its decomposition, as detailed in the main text. In columns (1) to (3), confidence from 0% to 100% is mapped to belief probabilities in $[0.5, 1]$ for correct and $[0, 0.5]$ for incorrect answers. In column (4), confidences from 5% to 100% are mapped to $[0.5, 1]$ for correct and $[0, 0.5]$ for incorrect answers, while observations below 50% are dropped. In column (5), we further drop observations where a receiver with a wrong answer hears from an orator with a different wrong answer. Sample is the main Receiver experiment for *Explanation* and all collections for *Choice Only*, excluding 0.7% of observations with missing confidence. Standard errors are clustered at the orator and receiver levels.

posterior on the left-hand side instead of adding it as a covariate on the right-hand side. The bin-scatter shows our linearized model fits the data quite well, and again illustrates the *pivot* effect. They also reflect the strong contribution of the *shift*: the least confident correct explanations cause more belief movement than *Choice Only*.

Appendix Table A2: Belief movement regression

	Dependent variable: Change in receiver log-odds				
	(1)	(2)	(3)	(4)	(5)
Intercept	0.257***	-0.435***	-0.502***	-0.370***	-0.434***
	(0.019)	(0.030)	(0.041)	(0.042)	(0.045)
Explanation	0.185***	0.187***	0.161***	0.083	0.073
	(0.021)	(0.042)	(0.055)	(0.057)	(0.059)
Orator correct		1.239***	1.257***	1.000***	1.064***
		(0.057)	(0.057)	(0.057)	(0.060)
Explanation × Orator correct		-0.101	-0.094	0.038	0.048
		(0.083)	(0.083)	(0.083)	(0.085)
Orator log-odds		-0.012	-0.052**	-0.037	-0.045*
		(0.012)	(0.020)	(0.023)	(0.026)
Explanation × Orator log-odds		0.104***	0.089***	0.073**	0.071**
		(0.018)	(0.028)	(0.033)	(0.034)
Orator correct × Orator log-odds			0.060**	0.015	0.024
			(0.027)	(0.029)	(0.031)
Explanation × Orator correct × Orator log-odds			0.023	0.018	0.020
			(0.036)	(0.041)	(0.042)
ATE	0.185	0.189	0.189	0.177	0.182
% shift	100%	82%	82%	90%	91%
% pivot	0%	18%	18%	10%	9%
Observations	21199	21199	21199	13213	12658
R^2	0.003	0.201	0.202	0.183	0.194

Notes: Outcome is the change in receiver belief in log-odds $\ell_{i2} - \ell_{i1}$. See notes for Table A1.

B Annotation of explanations

Our annotation starts from transcripts generated by Phonic using Amazon Transcribe. Notably, these transcripts preserve disfluencies or hesitation markers like "um" or "eh" that are typically removed by speech-to-text software. We then annotate these transcripts using a combination of human coding by a team of RAs and machine coding by a Large Language Model (LLM). For the latter, we use the state-of-the-art OpenAI GPT-4, with a temperature set to 0 for reproducibility.

We annotate four different dimensions of explanations. First, we categorize explanations into broad categories, e.g., to distinguish restatements of the answer from non-substantive or substantive argumentation. Second, we identify a large set of 31 features in the explanations, e.g., the word count, the number of uncertainty markers or of analogical arguments. Third, we

rate the general richness of explanations using a pre-registered definition. Fourth, we identify the different arguments appearing in each task and tag their presence in each explanation.

B.1 Explanation categorization

We first categorize speeches into general categories to acquire a broad overview of the different types of explanations. For that, we asked a team of RAs to identify whether an explanation fell into one of the following categories: *Only Restatement*, *Any Uncertainty*, *Non-Substantive Explanation*, *Substantive Explanation*, *Correct Explanation*, *Incorrect Explanation*, *Unclear Explanation*, *Invalid Explanation* (see Table B4 for a detailed overview). They are not necessarily mutually exclusive.

To benchmark our fully manual categorization, we then performed the same categorization with GPT-4. When the human coder identified one of the categories, GPT-4 did so too in 79% of cases; when the human coder did not identify one of the categories, GPT-4 did so too in 82% of cases. Cohen's κ is at 0.53, indicating 'moderate agreement'. These statistics are higher for the more specific categories we rely on in our analyses, e.g., they stand at 56%, 96% and 0.55 for the *Only Restatement* category. Aggregate frequencies also seem more stable, e.g., with human coder finding 13.1% of explanations to be Only Restatements while GPT-4 identifies a close 11.1%.

B.2 Feature identification

We identify 31 text features in explanations, which are domain-general and were largely taken from the vast existing research on text analysis and natural language data. We extract 25 features in five categories: language markers, disfluencies, certainty markers, reasoning content and addresses to the Receiver. Some features potentially overlap, e.g., we simultaneously extract high confidence markers, low confidence markers and any confidence markers.

We instruct GPT-4 to identify all instances of each feature and return them as a JSON dictionary of lists. The annotation can then easily be audited, and appears sensible upon inspection. Instances are then counted, and counts are then standardized (intensive margin) or turned into dummies equal to 1 if any instance has been detected (extensive margin).

We generate 6 simple textual & speech features via direct computation. To investigate which features explain richness in Figure D7, we additionally generate 19 more advanced features, notably based on the distribution of words, part-of-speech tagging, named entity recognition and syntactic structure identification. Table B3 provides an overview of all features.

B.3 Richness rating

To assess the richness of explanations, we provide GPT-4 with the following, pre-registered definition of richness: *A rich explanation is detailed, comprehensive, logically structured, nuanced, and tailors the argument to fit the context. A sparse explanation is basic, narrow, unclear or disorganized, presents only surface-level understanding, lacks depth or specific details and fails to clearly relate to the context.* We instruct GPT-4 to rate each speech's richness individually on a numerical scale from 0 to 10 (both inclusive).

B.4 Argument identification

Section 5.1.1 describes the argument identification and annotation scheme. It also provides statistics on inter-rater reliability, from a second blind human annotation and from an annotation via GPT-4, all showing substantial agreement. Online Appendix Table F1 shows all arguments appearing in the final scheme. Each has a title used to denote it in Figures D8 and a detailed description used in the annotation.

Table B5 further shows the four types of argument we have identified. Section 5.2.1 describes how each speech is then associated with a specific argument category based on the strongest type of argument it contains.

Appendix Table B3: Explanation features annotated via GPT-4 or computed directly

Feature	Description
Language Markers	
Modal verbs	Verbs indicating possibility, probability, or necessity. Example: “might”, “could”, “would”.
Certainty adverbs	Adverbs indicating certainty or doubt. Example: “possibly”, “probably”, “likely”.
Hedging language	Phrases indicating hedged claims. Example: “it seems”, “appears to be”, “to the best of our knowledge”.
Relative language	Words indicating qualifiers or comparisons. Example: “almost”, “nearly”, “more or less”.
Absolute language	Words indicating absolutes or superlatives. Example: “Always”, “Best”.
Epistemic stance markers	Phrases indicating subjective judgment. Example: “I believe”, “we assume”, “in my opinion”.
Conditional statements	Sentences indicating “If-Then” constructs. Example: “If we don’t act now, then”, “Assuming X, then Y”.
Interrogation markers	Words indicating questions or uncertainty. Example: “who”, “what”, “where”, “when”.
Numerical expressions	Phrases indicating quantitative or probabilistic information. Example: “more than 100 banks”, “95% chance that”.
Disfluencies	
Filled pauses	Instances of filled pauses. Example: “um”, “ah”, “er”.
False starts	Sentences starting but not completed. Example: “If you look at - I believe that”.
Repetitions	Instances of word or phrase repetition. Example: “I I mean”, “this is, this is wrong”.
Repairs	Instances where the speaker corrects themselves. Example: “I have two- three dogs”.
Certainty Markers	
Certainty markers	Statements indicating overall confidence. Example: “Without a doubt”, “I am certain that”.
High certainty markers	Statements indicating high confidence. Example: “I am certain that”, “I am sure that”.
Low certainty markers	Statements indicating low confidence. Example: “It might”, “I’m not sure but”.
Reasoning Content	
Indications of origin	Statements indicating information origin. Example: “According to”, “My grandmother has always said that”.
Personal experience args.	Arguments based on personal experience. Example: “I have often found that”.
External authority args.	Arguments based on external authority. Example: “My girlfriend works at a bank and said”.
Empirical args.	Arguments based on empirical facts. Example: “I remember reading a newspaper article saying”.
Analogical args.	Arguments based on analogies. Example: “Investments funds are like babies”.
Logical reasoning args.	Arguments based on logical reasoning. Example: “Since active managers put in more research”.
Normative args.	Arguments based on ethical considerations. Example: “It would not be fair if”.
Addresses to Receiver	
Directive addresses	Directives to the listener. Example: “You should definitely say that”.
Apologetic or humble addresses	Apologetic or humble addresses. Example: “I apologize for not knowing more”.
Simple Computed Features	
Word count	Total number of words.
Word length	Average length of words.
Words per minute	Average number of words per minute.
Sentence count	Total number of sentences.
Sentence length	Average length of sentences.
Language complexity	Flesch-Kincaid readability grade.
Lexical Metrics	
Lexical Diversity	Ratio of unique words to total number of words.
Entropy of Words	Entropy of distribution of words.
Hapax Legomena Ratio	Percentage of words that appear only once.
Share of long words	Percentage of words that have more than 10 letters.
Additional Readability Metrics	
Gunning Fog Index	Years of education required to understand the text, based on sentence length and percentage of complex words.
SMOG Index	Years of education required to understand the text, based on polysyllabic word counts.
Automated Readability Index	Years of education required to understand the text, based on characters per word and words per sentence.
Cohesion Metrics	
Referential Cohesion	Mean word overlap between sentence and following sentence.
Syntactic Complexity Metrics	
Part-of-Speech Tag Entropy	Entropy of parts of speech (e.g., nouns, verbs, adjectives) in the text.
Mean Length of T-units	Average length of T-units, i.e., a main clause plus any subordinate clauses.
Subordination Index	Ratio of subordinate clauses to main clauses.
Clause Density	Average number of clauses per sentence.
Named Entity Recognition	
Entity Count / Words	Ratio of number of named entities (e.g., people, organizations, locations) to total number of words.
Entity Type Count / Words	Number of different types of named entities (e.g., person, organization, location) to total number of words.
Sentence Structure and Syntax Metrics	
Number of Clauses	Total number of clauses.
Syntactic Tree Depth	Maximum depth of the syntactic dependency tree.
Syntactic Tree Branching	Average number of branches per node in the syntactic tree.
Noun Phrase Density	Ratio of number of noun phrases to total number of words.
Verb Phrase Density	Ratio of number of verb phrases to total number of words.

Appendix Table B4: Overview of explanation categories

Category	Description	Example
Only Restatement	The explanation is purely a restatement of the answer, without any arguments or elaborations.	"I think it's number one."
Any Uncertainty	The explanation contains any expressions of (un)certainty in the answer or arguments presented.	"Um, this one is more tricky. I think it's, um, I think it would be that they do not outperform passively managed ones. Um, I'm not really sure of an exact explanation because to be honest, I don't have any idea. Um, sorry"
Non-Substantive Explanation	The explanation only contains non-substantive justifications: appeals to authority, appeals to emotion, etc.	"I believe that passively managed funds perform better. And I'm gonna say that as uh uh as I remember Warren Buffett uh during an interview [...]"
Substantive Explanation	The explanation contains any substantive justification, e.g., any form of argument.	"If active funds outperformed, passive funds wouldn't exist."; "A fund is just like a plant, if you take more care of it, it will grow better."
Correct Explanation	The explanation is correct in meaning.	"I believe that actively managed funds do not outperform passively managed ones, the account for fees is too high when constantly monitoring an actively managed account."
Incorrect Explanation	The explanation is incorrect in meaning.	"Actively managed funds, do outperform, passive ones because you're actively making decisions about it and doing what makes you the most money."
Unclear Explanation	The explanation is very unclear or nonsensical.	"Passively managed funds, outperform, actively managed funds. And this is why hedge funds have a very short life spans. So question number two."
Invalid Explanation	The explanation is empty or entirely incomprehensible due to transcription errors.	'Yes, I conquer, actively managed form. I perform passively managed forms. Every time, every time I really conquer, I good choice."

Appendix Table B5: Overview of argument types

Type	Description	Example
Sound Argument	An argument that has correct premises and where the conclusion follows from the premises. The premises might not quite be sufficient for the conclusion.	"I believe that actively managed funds do not outperform passively managed ones, the account for fees is too high when constantly monitoring an actively managed account." (Active funds charge fees)
Fallacious Argument	An argument that is relevant to the question or its answer, but where one or more of the premises are false, or the conclusion is not valid given the premises.	"Actively managed funds will outperform passively managed ones because actively managed funds make more strategic decisions. While passively managed ones are kind of just going with the flow of the market. But actively managed funds can predict what the market is gonna do and make a decision based on that. So the answer is actively managed funds outperform, passively managed ones." (Active funds managed by experts)
Irrelevant Argument	An argument whose premises are unrelated to the question or its answer.	"Actively managed funds, outperform, passively managed ones because they are being actively managed. Whereas passively managed ones are being managed passively and actively sounds better than passively."
No Argument	No argument given at all.	"Um, actively managed funds outperform passively managed ones most times probably."

C Robustness checks

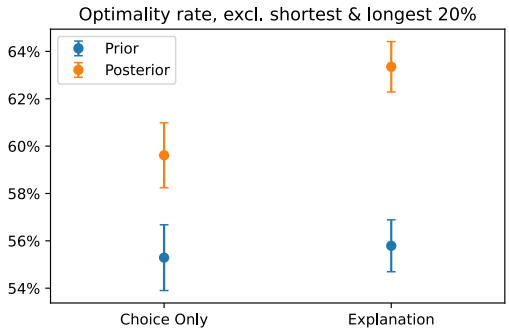
C.1 Excluding the shortest and longest recordings

To ensure our reduced-form findings are not driven by a small subsample of explanations, e.g., extremely succinct or long-winded, we verify that they are robust to excluding the shortest and longest 20% of recordings. We filter recordings based on the total duration of the audio file.

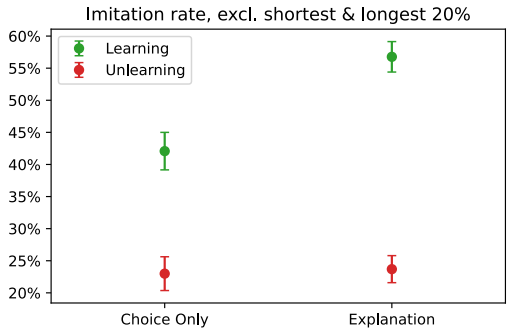
Figure C2b shows the resulting optimality rates, mirroring Figure 1b. The share of receivers giving the correct answer before having been exposed to the orator's explanation is 55.3% in *Choice Only* and 55.8% in *Explanation* ($p = 0.61$). After being exposed to the orator's explanation, the share of receivers giving the correct answer is 59.6% in *Choice Only* and 63.4% in *Explanation* ($p < 0.01$). Being exposed to an orator's explanation instead of only their choice therefore increases the optimality rate by 3.3 p.p. ($p < 0.01$).

Figure C2b repeats the analysis by learning and unlearning situations as in Figure 1b. In unlearning situations, being exposed to an orator's explanation increases the likelihood of imitating their answer an insignificant 0.7 p.p. ($p = 0.68$) relative to only seeing their choice. On the other hand, in learning situations, being exposed to an orator's explanation in addition to their choice increases the likelihood of taking over their answer by 14.7 p.p. ($p < 0.01$).

In conclusion, both findings confirm that our results from Section 3 are robust to excluding the shortest and longest 20% of explanations from the sample.



(a) Prior and posterior optimality



(b) Imitation by situation

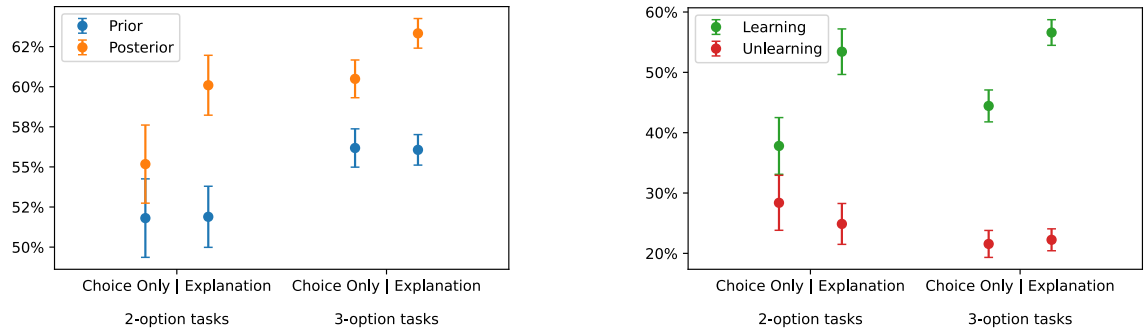
Appendix Figure C2: Robustness of main findings to dropping shortest and longest recordings. *Notes:* Analyses and starting samples are the same as in Figures 1a and 1b. We additionally drop the 20% of shortest and 20% of longest explanations. Whiskers show 95% CIs.

C.2 Distinguishing by number of options

While twelve of the tasks gathered from the literature have three answer options, three of them are effectively true/false questions and therefore only have two options: "Actively managed funds", "Crypto mining" and "Stock picking" (cf. Table E6).

It is natural to expect different optimality rates if some respondents randomize between answers. In our case, however, the prior optimality rate is 51.8% for the two-option tasks but 56.2% for the three-option tasks ($p < 0.01$), showing that cross-task heterogeneity dominates this mechanism. Prior confidence is 63.2% on average for two-option tasks, against 68.8% for three-option tasks ($p < 0.01$). Posterior confidence exhibits a similar gap at 68.8% and 74.0% respectively ($p < 0.01$).

Although the number of options is constant across conditions and should therefore not influence treatment effects, Figure C3 confirms that our main effects hold for both types of tasks. The left panel shows that *Explanation* increases posterior optimality by 4.9 p.p. relative to *Choice Only* among two-option tasks ($p < 0.01$), and by 3.0 p.p. among three-option tasks ($p < 0.01$). This difference in treatment effects between the two types of tasks is not statistically significant ($p = 0.18$). In the right panel, the effect of *Explanation* over *Choice Only* is not significant in unlearning situations for two- or three-option tasks ($p = 0.23$ and $p = 0.64$). On the other hand, in learning situations it is 15.8 p.p. for two-option tasks ($p < 0.01$) and 12.2 p.p. for three-option tasks ($p < 0.01$). As before, this difference is not statistically significant ($p = 0.32$).



(a) Posterior optimality

(b) Imitation rate

Appendix Figure C3: Effect of *Explanation* on optimality and imitation by number of options. *Notes:* See notes for Figure 1. Two-option tasks are "Actively managed funds", "Crypto mining" and "Stock picking", see Table E6. Whiskers show 95% CIs.

The fact that we use two- and three-option tasks introduces an interesting source of variation

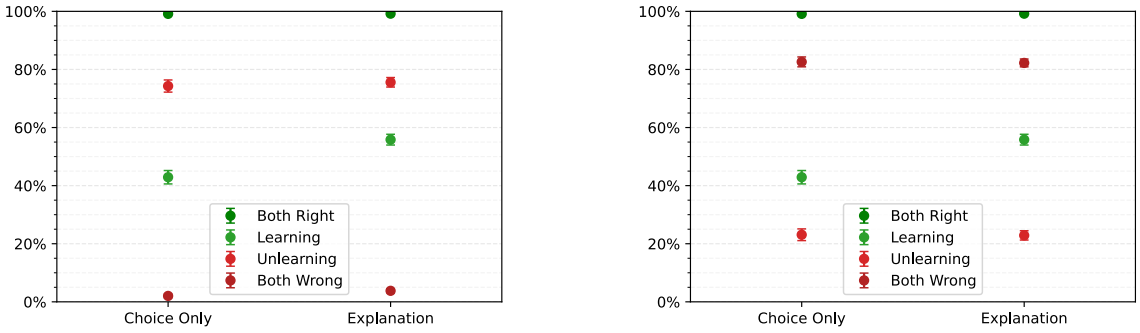
when analyzing receivers' reaction to confirmatory signals, as discussed in Section 3.3. Importantly, we define imitation as the receiver picking the exact same option as the orator. When the receiver and the orator are both right, they have necessarily chosen the same answer. However, when both are wrong, they will sometimes have chosen two different wrong answers.

When the receiver and orator are both wrong, imitation rates are 82.6% in *Choice Only* and 82.2% in *Explanation* ($p = 0.75$), substantially lower than when both are right. However, in tasks with only two options, imitation rates are 97.5% and 96.7% respectively ($p = 0.44$). They are similarly high in tasks with three options when the orator chose the same wrong option as the receiver, at 97.7% and 96.3% ($p = 0.03$). On the other hand, imitation is much lower at 36.0% and 37.8% ($p = 0.52$) in three-option tasks when the orator chose a different wrong option. Interestingly, the imitation rate for an alternative wrong option is between the learning and unlearning rates. Explanations therefore slightly reduce imitation in three-option tasks when receiver and orator initially picked the same wrong answer.

In the main analysis, we noted that aggregate posterior optimality rates display a small 1.8 p.p. ($p < 0.01$) treatment effect when both are wrong. Again breaking this down by number of options, we find an insignificant 0.8 p.p. ($p = 0.44$) effect on optimality in two-option tasks; a highly significant 1.8 p.p. ($p < 0.01$) effect in three-option tasks when the orator chose the same wrong option as the receiver; and a 2.5 p.p. ($p = 0.06$) effect in three-option tasks when the orator chose another wrong option.

Explanations therefore reduce imitation and increase posterior optimality when the orator chose the same wrong option as the receiver, meaning receivers switch away from their and the orator's choice to the correct answer. When the orator chose a different wrong option from the receiver's, explanations have no effect on imitation but increase optimality, so that they only make the receiver more likely to choose the correct answer. The coexistence of two- and three-option tasks therefore helps us identify the small but significant enabling role that arguments can play, by prompting incorrect listeners to think more deeply and switch to the correct answer.

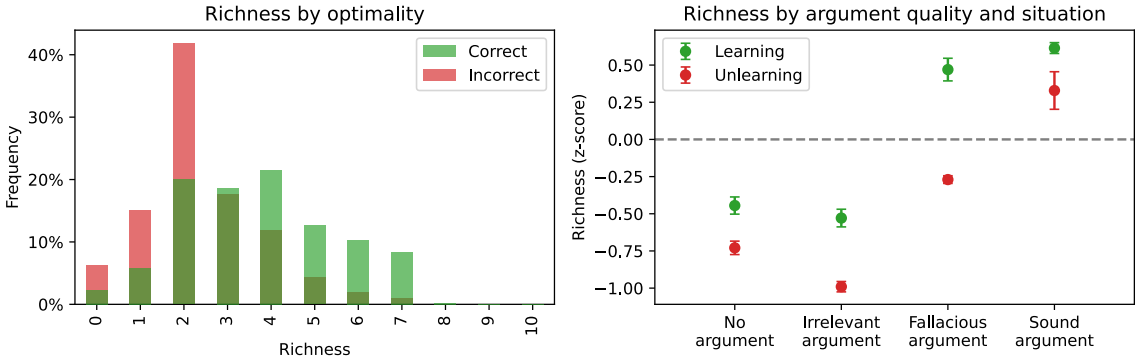
D Additional figures



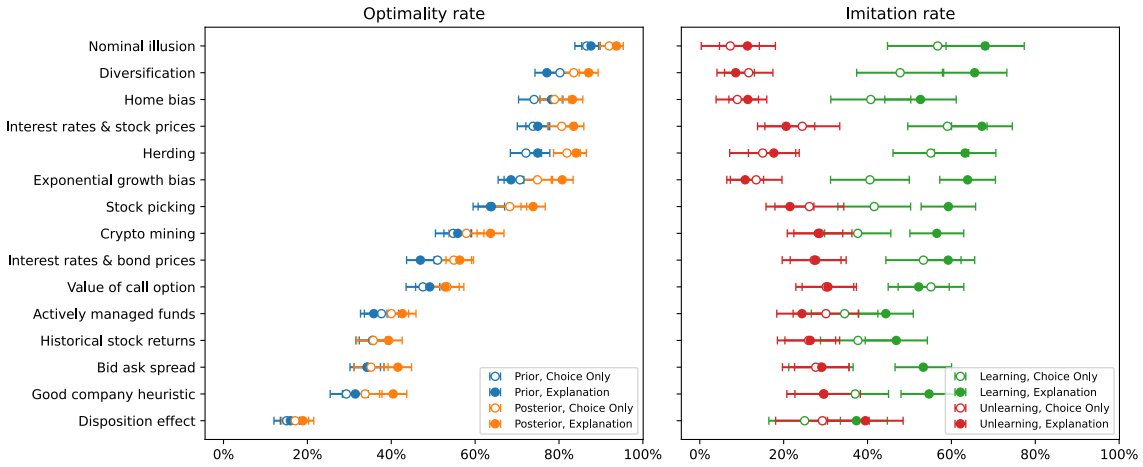
(a) Posterior optimality

(b) Imitation rate

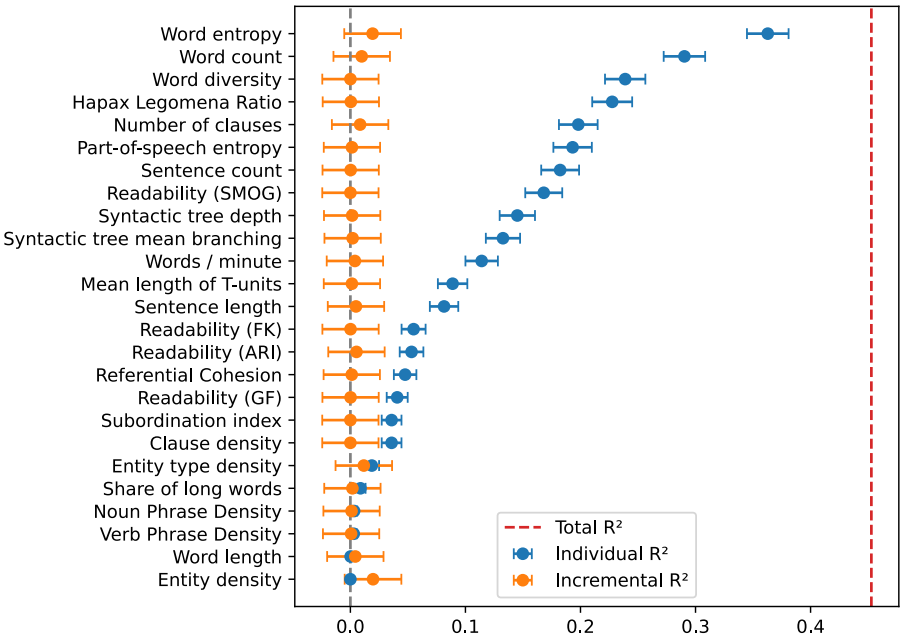
Appendix Figure D4: Effect of *Explanation* in conflicting and confirming situations. *Notes:* See notes for Figure 1. By construction, receiver prior optimality is 0% in *Learning* and *Both Wrong*, and 100% in *Unlearning* and *Both Right*. Whiskers show 95% CIs.



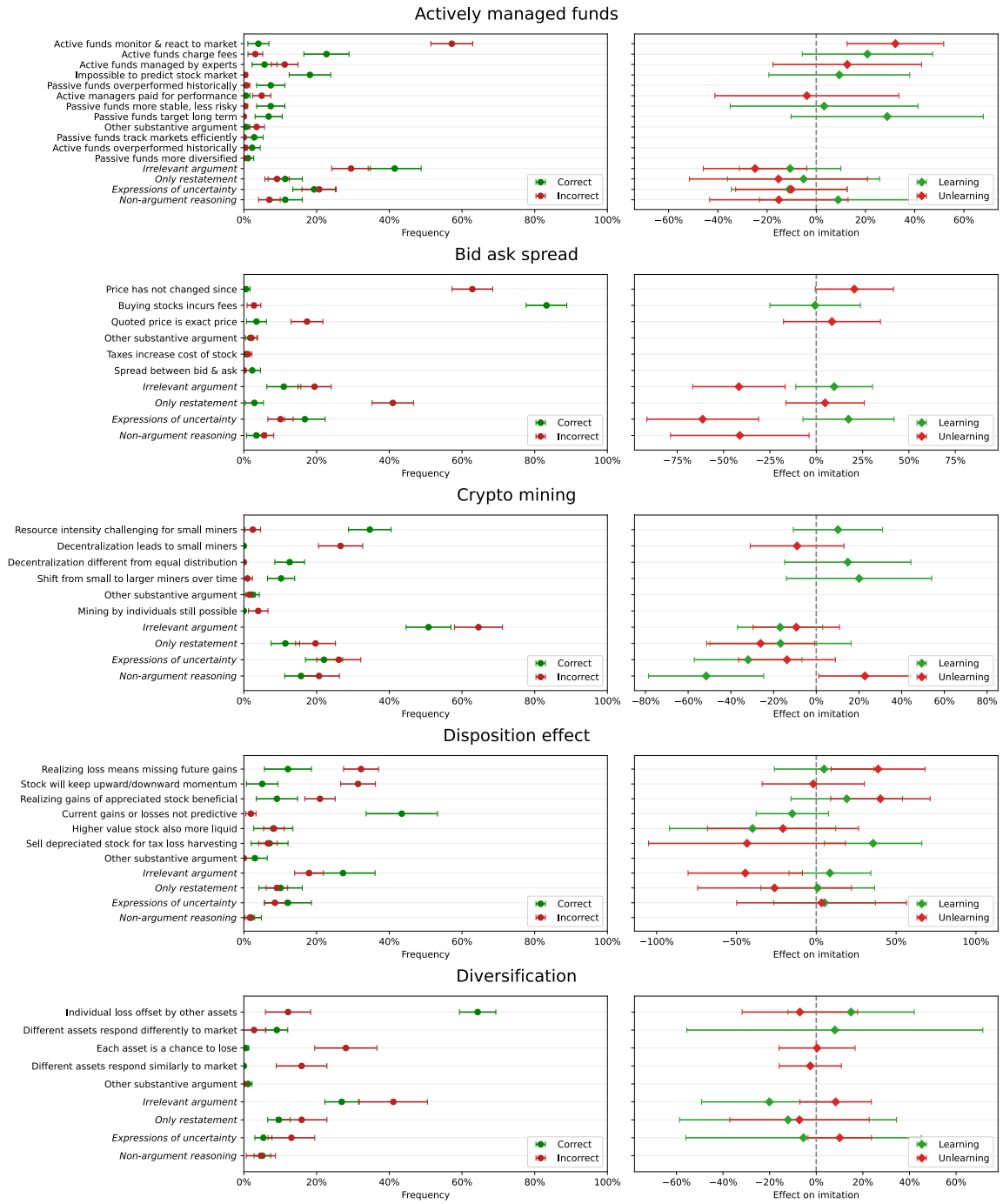
Appendix Figure D5: Richness gap by orator optimality, and by argument quality and situation. *Notes: Explanation* sample is the main Receiver experiment, *Choice Only* is pooled from all collections. See Appendix B for details on richness ratings. Whiskers show 95% CIs.



Appendix Figure D6: Effect of *Explanation* on optimality and imitation by task. *Notes:* See notes for Figure 1. Whiskers show 95% CIs.

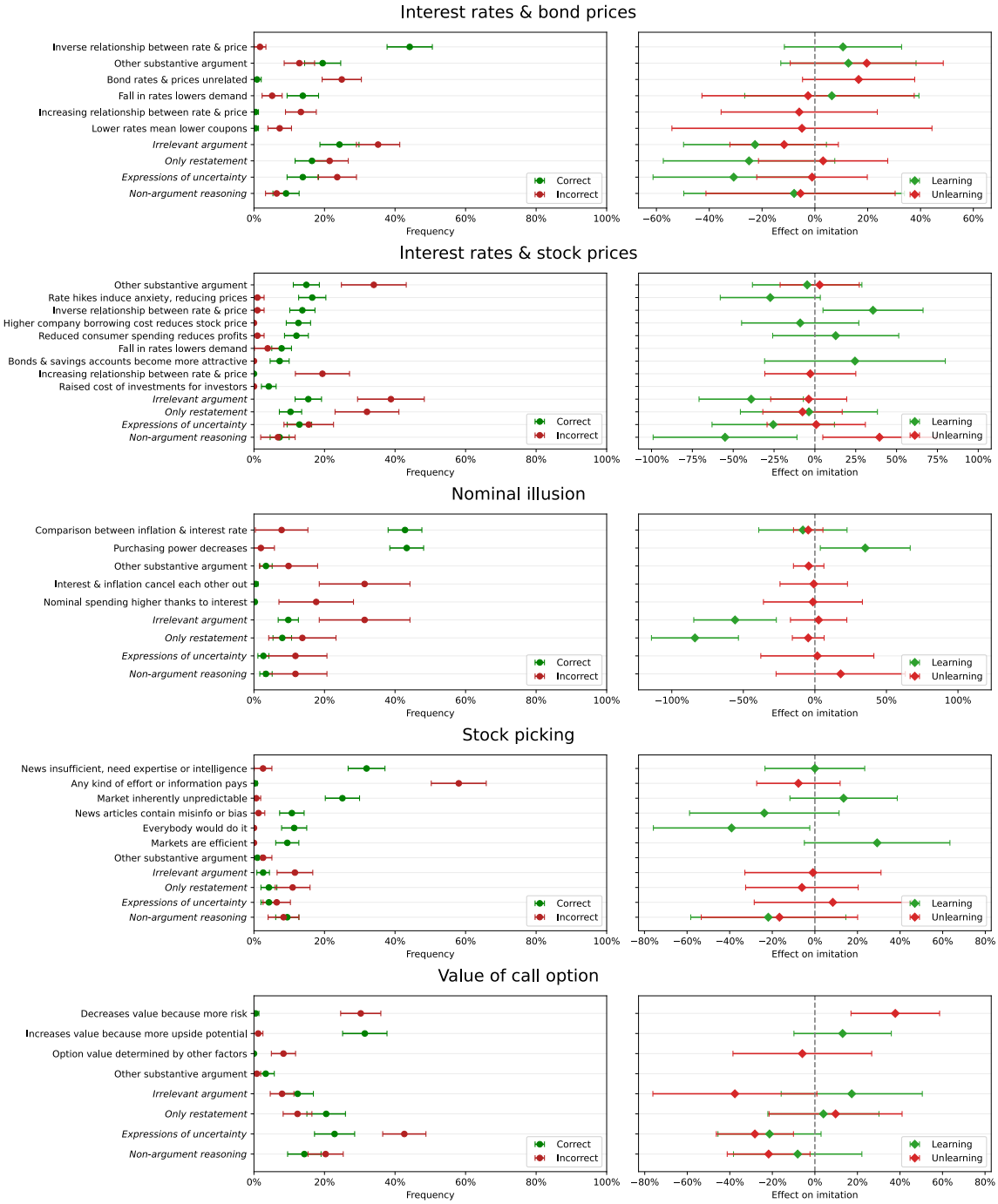


Appendix Figure D7: Explanatory power of explanation features for richness. *Notes: Individual* R^2 is the R^2 in a regression of richness on a given feature and a constant. *Incremental* R^2 shows the decrease in R^2 when removing a given feature from a regression of richness on all features and a constant. *Total* R^2 shows the R^2 in a regression of richness on all features and a constant. Whiskers show 95% CIs.



Appendix Figure D8: Arguments by task I/III. *Notes:* See notes for Figure 6.

Appendix Figure D8: Arguments by task II/III. *Notes:* See notes for Figure 6.



Appendix Figure D8: Arguments by task III/III. *Notes:* See notes for Figure 6.

E Additional tables

Appendix Table E6: Financial Decision Questions

Task	Motivation	Question
Actively managed funds	Overestimating the return (after fees) of actively vs. passively managed funds. Adapted from Haaland and Næss (2023).	Do actively managed investment funds systematically outperform passively managed investment funds in terms of expected net returns, i.e., after accounting for investment fees? (i) Actively managed funds outperform passively managed ones. (ii) Actively managed funds do not outperform passively managed ones.
Bid-ask spread	Assessing knowledge about features of financial transactions.	You look up live stock prices on the internet and see that the current trading price of a stock you're interested in buying is $30. You go to your online broker and buy that stock. Assuming the trading price hasn't changed in the meantime, how much do you have to pay for the stock? (i) Less than $30 (ii) Exactly $30 (iii) More than $30
Crypto mining	Testing knowledge of the structure of the Bitcoin network.	Since the blockchain is decentralized, most Bitcoin mining is done by many small miners. (i) True (ii) False
Disposition effect	Failing to account for the random walk of stock prices. Investors have a stronger tendency to sell assets at a profit than to sell at a loss.	You have two stocks in your portfolio: one went up a lot in value since you bought it whereas the other one lost value. You need to sell one to raise cash. Is it optimal to sell the one that has lost value since you bought it? (i) Yes (ii) No (iii) This does not make a difference
Diversification	Assessing how investing in several different asset classes affects risk. Taken from Atkinson and Messy (2012).	When an investor spreads his money among different assets, does the risk of losing money: (i) Increase (ii) Decrease (iii) Stay the same
Exponential growth bias	Underestimating the exponential effects of compounding. Taken from Lusardi and Mitchell (2007).	Suppose you had $100 in a savings account and the interest rate was 2 percent per year. After 5 years, how much do you think you would have in the account if you left the money to grow: (i) More than $110 (ii) Exactly $110 (iii) Less than $110
Good company heuristic	Failing to consider that market prices reflect available information, including growth prospects.	Imagine two hypothetical firms from the same industry, Firm A and Firm B, which have equal risk. However, Firm A has much higher growth prospects than Firm B. Imagine investing into one of the two firms. Which investment yields higher returns? (i) Firm A (ii) Firm B (iii) Need to know more information
Herding	Being influenced by "old news" from others, e.g., stories of friends, when investing.	Some of your friends with no prior experience or expert knowledge in financial markets tell you that they bought cryptocurrencies and made a lot of money with those cryptocurrencies; they mention that they bought after they came across an interesting newspaper article which describes the past price movements of cryptocurrencies. For your long-run investment strategy, how should the experience and information received from your friends influence your decision to invest (more) into cryptocurrencies? (i) Should invest more (ii) Should invest less (iii) Should not affect my decision
Historical stock returns	Estimating average historical returns of the S&P 500.	What is the average annual return of the S&P 500 stock market index over the past 20 years? (i) Less than 10% (ii) Between 10% and 15% (iii) More than 15%
Home bias	Believing that firms headquartered close to home outperform better investments.	Imagine two hypothetical companies that are identical in every possible way except that one is headquartered in your home state, whereas the other one is not. Assume you're deciding between investing in one firm or the other. Which one is the better investment? (i) The firm headquartered in my home state. (ii) The firm headquartered outside of my home state. (iii) Given the assumptions, both are equally good investments.
Interest rates & bond prices	Assessing the interaction between interest rates and bond prices. Taken from Lusardi and Mitchell (2007).	If the interest rate falls, what should generally happen to bond prices? (i) Rise (ii) Fall (iii) Bond prices are not affected
Interest rates & stock prices	Assessing the interaction between interest rates and stock prices. Adaptation from Lusardi and Mitchell (2007).	When the Fed increases interest rates more aggressively than expected by markets, what should happen to stock prices on average? (i) Stock prices will rise (ii) Stock prices will fall (iii) Stock prices will stay the same
Nominal illusion	Failing to assess purchasing power in real terms. Taken from Lusardi and Mitchell (2007).	Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, would you be able to buy: (i) More than today (ii) Exactly the same as today (iii) Less than today
Stock picking	Overconfidence in the value of free online news to "beat the market". Many investors actively pick stocks despite evidence that this leads to underperformance for most participants.	Most people could systematically outperform the stock market by carefully reading free online news articles about how recent events will affect different companies and picking the right stocks based on those readings. (i) True (ii) False
Value of a call option	Inferring how uncertainty affects the value of financial derivatives.	Holding everything else constant, how is the value of a call option for a stock generally affected by a higher volatility of that stock? (i) Higher volatility increases the value of a call. (ii) Higher volatility decreases the value of a call. (iii) Higher volatility has no effect on the value of a call.

Notes: Correct answers are underlined.

Appendix Table E7: Overview of data collections

Collection	Sample	Respondents	Treatments	Main outcomes	Pre-analysis plan
Baseline experiments					
Orator Experiment	Prolific	466	None	Choices in 15 financial decision tasks and voice recordings of explanations for choices	`https://aspredicted.org/4t5b-42ws.pdf`
Explanation Receiver Experiment	Prolific	1,103	*Choice Only* *Explanation*	Choices in 15 financial decision tasks	`https://aspredicted.org/4t5b-42ws.pdf`
Additional experiments					
Confidence Receiver Experiment	Prolific	713	*Choice Only* *Choice & Confidence*	Choices in 15 financial decision tasks	`https://aspredicted.org/fwv8-3gzs.pdf`
Transcript Receiver Experiment	Prolific	917	*Choice Only* *Transcript*	Choices in 15 financial decision tasks	`https://aspredicted.org/y2f5-8zdp.pdf`
Richness Receiver Experiment	Prolific	972	*Sparse version* *Rich version*	Choices in 15 financial decision tasks	`https://aspredicted.org/9t49-kxbc.pdf`
Balls-and-Urns Orator Experiment	Prolific	464	None	Estimate in standard balls-and-urns task	`https://aspredicted.org/p6sd-p88k.pdf`
Balls-and-Urns Receiver Experiment	Prolific	1,822	*Choice Only* *Explanation*	Estimate in standard balls-and-urns task	`https://aspredicted.org/p6sd-p88k.pdf`

Notes: Sample sizes refer to the final sample of respondents that satisfied the pre-specified inclusion criteria for each collection.

Appendix Table E8: Treatment effects on optimality

	Dependent variable: Posterior correct - Prior correct
Intercept	0.041***
	(0.005)
Choice & Confidence	0.010*
	(0.006)
Transcript	0.019***
	(0.006)
Explanation	0.033***
	(0.006)
Observations	40859
R^2	0.001

Notes: Choice Only is the omitted category. The p-value in a t-test for equality between *Transcript* and *Explanation* is $p = 0.004$. Samples are the corresponding Receiver experiments for *Explanation* (1,103 receivers, 13,111 obs.), *Choice & Confidence* (713 receivers, 8,522 obs.) and *Transcript* (917 receivers, 10,964 obs.), while *Choice Only* is pooled from all collections (2,733 receivers, 8,232 obs.).

Appendix Table E9: Treatment effects on imitation in learning & unlearning situations

	Dependent variable: Imitation
Intercept	0.231***
	(0.010)
Choice & Confidence	-0.007
	(0.016)
Transcript	-0.010
	(0.014)
Explanation	-0.002
	(0.013)
Learning	0.198***
	(0.017)
Learning × Choice & Confidence	0.012
	(0.023)
Learning × Transcript	0.090***
	(0.022)
Learning × Explanation	0.132***
	(0.021)
Observations	16850
R^2	0.083

Notes: Choice Only is the omitted category. The p-value in a t-test for equality between *Learning* × *Transcript* and *Learning* × *Explanation* is $p = 0.018$. Samples are as in Table E8, restricted to learning and unlearning situations.

Appendix Table E10: Decomposition of differential learning effect in *Transcript* treatment

	Dependent variable: Imitation						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Transcript	-0.009	0.022	-0.092**	-0.034	-0.053	0.039**	0.028
	(0.014)	(0.018)	(0.046)	(0.054)	(0.068)	(0.015)	(0.069)
Learning	0.199***	0.199***	0.200***	0.128***	0.112***	0.207***	0.114***
	(0.017)	(0.021)	(0.018)	(0.016)	(0.021)	(0.018)	(0.021)
Transcript × Learning	0.090***	0.043	0.067***	0.092***	0.044	0.023	0.017
	(0.022)	(0.027)	(0.023)	(0.022)	(0.027)	(0.023)	(0.027)
Richness						-0.011	-0.007
						(0.009)	(0.010)
Transcript × Richness						0.095***	0.078***
						(0.012)	(0.014)
Argument controls		✓			✓		✓
Orator controls			✓		✓		✓
Receiver controls				✓	✓		✓
Observations	7879	7879	7879	7879	7879	7879	7879
R^2	0.072	0.079	0.078	0.161	0.174	0.086	0.180

Notes: See notes for Table 1. *Explanation* sample is the *Transcript* survey, *Choice Only* sample is pooled from all collections. We drop the 0.3% of observations with missing receiver prior confidence from all regressions.