Contagious Beliefs*

Thomas Graeber **Christopher Roth** **Constantin Schesch**

October 13, 2023

Abstract

When ideas circulate on the marketplace of ideas, both truths and falsehoods can spread. We study whether exposure to others' verbal explanations improves or worsens performance across fifteen canonical financial reasoning tasks. In our experiments, one group of participants records explanations for each of their choices. Another group either only observes the orator's choice or additionally listens to one of the over 3,000 verbal explanations before providing their own choice. Listening to verbal explanations sharply increases choice accuracy on average, whereas merely observing another's choice does not. The benefit of explanations is entirely driven by truths spreading more quickly; in fact, falsehoods do not become less contagious. Guided by a simple model, we measure and document a corresponding pattern in how well perceptions of others' accuracy are calibrated, and examine the underlying mechanisms. We find that the supply side of explanations is responsible for the differential effect of explanations, rather than systematic differences in interpreting them: explanations for correct choices contain more features that permit inference about their accuracy than those for incorrect choices. Our findings suggest that the contagion of truths and falsehoods depends on the degree to which people's explanations are diagnostic of their accuracy.

*We thank Simon Cordes, Jindi Huang, Paul Grass, Nicolas Röver, Gabriel Saliby and Georg Schneider for outstanding research assistance. We thank Peter Andre, Kai Barron, Stefano DellaVigna, Benjamin Enke, Nicola Gennaioli, Ryan Oprea, Jesse Shapiro, and Florian Zimmermann for helpful comments and suggestions. We thank the seminar audiences at the Max-Planck Institute for Collective Goods in Bonn 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. Human subjects approval was granted by Harvard IRB and the University of Cologne. Graeber: Harvard Business School, tgraeber@hbs.edu. Roth: University of Cologne and ECONtribute, roth@wiso.uni-koeln.de. Schesch: Harvard Business School, c.schesch@gmail.com

1 Introduction

We obtain most ideas, news and knowledge from listening to others. Some of the information we receive from others is accurate, while some of it is flawed. Whether learning from others improves or impairs our decisions therefore critically depends not only on the actual quality of the information, but also 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*:[1] the truth will emerge and prevail in an environment where thoughts and opinions are freely exchanged, much like how the best products or services thrive in a free market. Yet, an abundance of cases suggests that, instead, misbeliefs too can spread rapidly, because individuals fail to identify falsehoods (Lazer et al., 2018; Pennycook and Rand, 2021).

When people learn from others' verbal explanations, does the truth spread or do falsehoods become contagious? We conduct large-scale experiments in which respondents receive verbal information from others before solving canonical financial reasoning tasks. We focus on the effect of verbal explanations because much of the information that spreads through the economy is conveyed through people speaking and listening to each other. Unlike in canonical models of communication in economics, such information is not purely quantitative, but largely qualitative and delivered in natural language. In the context of financial reasoning in particular, insights, reasons, and justifications are routinely disseminated through word of mouth from peer to peer (Duflo and Saez, 2003; Brown et al., 2008).

Our experimental design comprises two separate experiments with different sets of respondents.We start with an Orator experiment: Respondents complete fifteen canonical financial reasoning problems, which include questions on nominal illusion, the net returns of active and passive investing, the relationship between interest rates and bond prices, and compound interest. For each question, respondents indicate their response in an incentivized manner. They then record a voice message in which they explain their answer to randomly matched other participants. Orators are incentivized in a way that induces aligned interests with the listeners of their recording: the Orator's likelihood of receiving a bonus payment increases in the accuracy of the listener's response. The Orator experiment allows us to characterize the supply side of explanations in an ecological way, using a large and heterogeneous set of respondent-generated voice recordings.

To study the effect of explanations on imitation, we conduct a Receiver experiment, in which respondents face the same fifteen tasks. They first make their own incentivized choice,

[1]The marketplace of ideas is a foundational concept underlying freedom of speech and open discourse which is routinely attributed to U.S. Supreme Court Justice Oliver Wendell Holmes Jr.

providing us with a prior. In the *Choice Only* condition, they subsequently learn about the choice of a randomly chosen respondent in the Orator experiment. In our main treatment arm (*Explanation*), 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 prior. 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 answer. The *Choice Only* condition is critical to control for the direct effects of the respondents' confidence in their prior answers, measurement error in priors and other confounders, such as experimenter demand effects.

We begin by analyzing reduced-form treatment effects on optimality and imitation rates. Observing someone else's choice does not change the average frequency of optimal choices in the sample, whereas explanations do create substantial improvements: accuracy rates increase by 4.8 percentage points ($p < 0.01$), corresponding to a 10 percent increase from the prior optimality rate. Intuitively, changes in the average optimality rate should be caused by those receivers who are exposed to *conflicting* advice, because those with confirming advice have no reason to change their answer. Conflicting advice comprises two cases: receivers with incorrect priors can be exposed to correct choices, creating *learning opportunities*, and receivers with correct priors can be exposed to incorrect orators, which we refer to as *unlearning opportunities*. We find that the aggregate effect of explanations is entirely driven by learning opportunities: in those cases, the imitation rate is 50% in *Explanation* but only 35% in *Choice Only* ($p < 0.01$). This corresponds to a staggering 43% increase in the successful use of learning opportunities induced by verbal advice. The *Explanation* treatment does not, on the other hand, decrease the frequency with which receivers switch to a wrong answer in unlearning opportunities. In roughly one fourth of all unlearning opportunities, receivers switch from accurate to inaccurate answers in both the *Explanation* and *Choice Only* conditions. Why then do verbal explanations improve outcomes in learning but not unlearning opportunities?

To guide our subsequent examination of mechanisms, we propose a simple conceptual framework that models imitation based on Bayesian updating from a signal with subjectively perceived diagnosticity. Specifically, imitation is shaped by two forces: first, an individual's confidence in their prior answer, often referred to as *meta-cognition*; and second, the subjective perception of whether the orator is accurate. The conceptual framework disciplines our analysis of optimality and imitation rates: under random matching of orators and receivers, there is an equal share of learning and unlearning opportunities among all matches. As a consequence, the sign of the difference between learning and unlearning

rates is a sufficient statistic for whether there are aggregate improvements or not. The model then shows that our main reduced-form finding of treatment differences between *Choice Only* and *Explanation* cannot be explained by meta-cognition (unlike in, e.g., Enke et al., 2023), because the distribution of prior answers and confidence is naturally the same across conditions. The treatment effect has to arise, instead, from the systematic effects of explanations on the perceived accuracy of orators' explanations. Intuitively, the calibration of perceptions of who is right and who is wrong is shaped more favorably by explanations in the case of learning than unlearning opportunities.

To investigate the mechanisms underlying the differential imitation patterns in learning and unlearning opportunities, we exploit an experimental measurement of receivers' perceptions of others' accuracy. In particular, respondents in the receiver experiments stated their degree of certainty about whether the choices they saw or the explanations they listened to were accurate. Perceptions of accuracy are indicative of actual accuracy on average: they are 72% for correct and 52% for incorrect orators in the *Explanation* treatment, and 69% and 55%, respectively, in the *Choice Only* treatment. A central prediction of our model is that learning and unlearning dynamics are governed by the correlation between the perceived accuracy of orators and their actual accuracy. We find calibration coefficients of $r = 0.24$ in *Choice Only* and $r = 0.33$ in *Explanation*, highlighting that, on average, verbal advice improves people's perceptions of who is right and who is wrong. Consistent with our findings on imitation, however, these differences in calibration are driven by learning opportunities. There is a large and significant difference in the calibration of receivers with incorrect priors, at $r = -0.35$ in *Choice Only* versus $r = -0.15$ in *Explanation* ($p < 0.01$). For receivers with correct priors, on the other hand, calibration is similar and not statistically different between *Choice Only* and *Explanation* ($r = 0.62$ and $r = 0.67$, respectively). This set of findings strongly indicates that the pattern of the effect of explanations on imitation is directly related to corresponding effects on perceptions of others' accuracy.

Building on this evidence about the potential micro-foundation of imitation decisions, we investigate the sources of the asymmetric benefits of explanations. There are two potential forces forces at play. The first concerns the supply side of explanations: explanations associated with correct and incorrect choices might systematically differ in terms of the frequency of features that are informative about their actual accuracy. Intuitively, the nature of verbal explanations may be such that it is easier to tell that a correct explanation is right than it is to tell that an incorrect explanation is wrong. The second concerns the receiver side: listeners with inaccurate priors might make systematically different inferences from the speech features of explanations than listeners with accurate priors. To investigate

these channels, we transcribe all voice messages and analyze their content using machine-learning methods. Our evidence establishes that the nature of explanations for correct and incorrect choices is indeed different: explanations associated with correct choices exhibit a higher frequency of features that are indicative of their accuracy, such as modal verbs, certainty adverbs, longer exposition and certainty markers. From the content alone, it is harder to tell that an incorrect explanation is wrong. At the same time, the interpretation of the textual features of explanations does not differ strongly between receivers with correct and incorrect priors. Taken together, our mechanism findings show that the virality of truths and falsehoods depends on the degree to which people's explanations exhibit features that are diagnostic of their accuracy.

Finally, we investigate belief contagion in settings where orators may not act in the best interest of listeners. In our main experiment, orators and receivers had aligned incentives to find out the ground truth. In practice, the person delivering an explanation often benefits from the receiver taking a specific action. In the final step of our analysis, we examine how strategic incentives shape the nature of explanations and the propagation of correct and incorrect beliefs. We design an additional treatment in which the orators' bonus payment depends not on the listener's accuracy but on whether the listener imitates them (*Imitation Incentives* condition). Comparing receiver behavior for the two orator treatments reveals three main findings: first, and perhaps surprisingly, imitation incentives do not affect average optimality rates, relative to aligned incnetives. Second, the absence of an average treatment effect shrouds substantial heterogeneity: compared to aligned incentives, explanations under imitation incentives increase *both* the learning and unlearning rates. This means that imitation incentives increase the dispersion in net learning rates. Third, we find that the increase in imitation rates is driven by an increased frequency of expressions of certainty and fewer uncertainty markers in voice recordings among orators with both correct and incorrect choices.

This paper contributes to several literatures. First, it relates to work on social learning (Mobius et al., 2015; Weizsäcker, 2010; Conlon et al., 2021; Eyster and Rabin, 2014; Mobius and Rosenblat, 2014; Banerjee, 1992; Bikhchandani et al., 1992), specifically in the context of financial decisions (Ambuehl et al., 2022; Haliassos et al., 2020; Hvide and Östberg, 2015; Bursztyn et al., 2014), as well as the literature on advice giving (Schotter and Sopher, 2003; Schotter, 2003; Çelen et al., 2010). Conlon et al. (2022) show that people are less sensitive to information others discover than to equally relevant information they receive themselves. We differ from the previous literature in three ways: first, we provide evidence on learning from qualitative explanations in natural language about the optimal choice in a series of

financial decisions. Second, we characterize the nature and implications of explanations for correct or incorrect choices. Third, our main evidence comes from a setting with aligned incentives where some people erroneously hold mistaken beliefs. Our results suggest that individuals are, on average, able to discern truths from falsehoods, especially when provided with explanations. Previous work on deception suggests that individuals fail to detect others' lies in a political context (Serra-Garcia and Gneezy, 2021).

We also contribute to an emerging literature on learning from qualitative information, e.g., in the form of stories (Graeber et al., 2023a) and narratives (Andre et al., 2022; Kendall and Charles, 2022; Barron and Fries, 2023; Bursztyn et al., 2023; Hüning et al., 2022; Shiller, 2017, 2020). Graeber et al. (2023b) examine how the process of human information transmission distorts the supply of qualitative economic information. We depart from existing work in our focus on the role of explanations in shaping the spread of truths and falsehoods.

Finally, by characterizing the contagiousness of truths versus falsehoods through social learning, we contribute to a long-standing literature on whether individual-level biases tend to matter for aggregate market-level outcomes (Russell and Thaler, 1985; Sonnemann et al., 2013; List, 2003). Enke et al. (2023) shows that awareness about biases reduces the impact of individual-level biases on aggregate outcomes through institutions that rely on self-selection, while Amelio (2023) studies how meta-cognition shapes social learning. Unlike those findings, ours cannot, by design, be due to meta-cognition. We instead examine people's capacity to discern truths from falsehoods and how this depends on the features of explanations. Our evidence on strategic incentives shows how the design of institutions influences the contagion of truths and falsehoods.

Our paper proceeds as follows: Section 2 describes the experimental design of our Orator and Receiver experiments. Section 3 presents our main reduced-form findings on improvements in aggregate outcomes and imitation rates in learning and unlearning opportunities. Section 4 develops our conceptual framework. We then turn to a model-guided examination of mechanisms: Section 5 introduces perceptions of others' accuracy as a driver of imitation decisions. Section 6 examines the nature of explanations as a source of our reduced-form findings. Section 7 examines how strategic incentives affect the contagion of truths versus falsehoods. Section 8 concludes.

2 Experimental Design

2.1 Overview

Our experimental design studies fifteen canonical financial reasoning 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 Reasoning Problems

We study financial reasoning because it is a domain in which people frequently share explanations and narratives (Shiller, 2020). Conversations about financial investment decisions are highly ecological and often involve large stakes.

We select fifteen well-known financial reasoning problems, based on the following criteria. First, we aim for a collection that is broadly representative of the universe of reasoning biases studied in the finance literature. This spans behavioral phenomena such as exponential growth bias and nominal illusion but also more specific knowledge about different asset classes and investment decisions, such as the expected returns under active versus passive investing. Second, we restrict our attention to questions with an objectively correct answer. Third, the questions should be reasonably short to describe.

To provide an example, consider the following simple question about the concept of inflation, with the correct answer underlined:

> *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?*
>
> *(i) More than today*
>
> *(ii) Exactly the same as today*
>
> *(iii) Less than today*

Other questions relate to typical investment decisions that people might make:

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

Some questions are more technical, such as:

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.

Finally, we include questions that require specific technical knowledge related to recent technologies, for example:

Thanks to the blockchain, Bitcoin can process hundreds of transactions per second.

(i) True

(ii) False

Table 1 provides an overview of all tasks we employ. Our tasks also vary in the number of response options. Some have two options, such as whether actively or passively invested funds yield higher net returns. Others have more than two answers such as the question about the disposition effects. One other question has elicited historical stock returns on a continuous response scale. For each question, we use a binary indicator for whether the chosen option is correct as our measure of optimality. Note that due to the differences across questions, one might naturally expect different optimality rates, for example because randomizing would create an optimality rate of 50% in a two-option problem but a rate of 25% in a four-option task.

2.3 Part 1: Orator Experiment

The objective of the Orator experiment is to obtain recordings of people's verbal explanations for each of the financial reasoning tasks. This allows us to generate a heterogeneous

Table 1: Overview of Financial Reasoning Questions

Task	Explanation
A. Understanding of interest rates	
Nominal illusion (NI)[a]	Failing to assess purchasing power in real terms. Taken from Lusardi and Mitchell (2007).
Exponential growth bias (EGB)[a]	Underestimate the exponential effects of compounding. Taken from Lusardi and Mitchell (2007).
Interest rates and bond prices (IRB)[a]	Assessing the interaction between interest rates and bond prices. Taken from Lusardi and Mitchell (2007).
Interest rates and stock prices (IRS)[a]	Assessing the interaction between interest rates and stock prices. Adaptation from Lusardi and Mitchell (2007).
B. Understanding of market efficiency	
Stock picking (STP)[b]	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 market participants.
Disposition effect (DEF)[a]	Failure to account for the random walk movement of stock prices. Investors have a stronger tendency to sell assets at a profit than to sell at a loss.
Actively Managed Funds (AMF)[b]	Overestimating the return (after fees) of actively vs. passively managed funds. Adaptation from Haaland and Næss (2023).
Good company heuristic (GCH)[a]	Failing to consider that market prices reflect available information, including growth prospects.
Home bias (HB)[a]	Believing that firms headquartered close to home outperform better investments.
Herding (HER)[a]	Being influenced by the crowd, e.g. stories of friends, when investing.
C. Other Financial Topics	
Diversification (DIV)[a]	Assessing how investing in several different asset classes affects risk. Taken from Atkinson and Messy (2012).
Historical stock returns (HSR)[c]	Estimating average historical returns of the S&P500.
Value of a call option (VCO)[a] addlinespace Bid-ask spread (BAS)[a]	Inferring how uncertainty affects the value of financial derivatives. Assessing the role of market frictions in financial transactions.
Crypto mining (CM)[b]	Testing knowledge of the structure and operations of the Bitcoin network.

[a] Answer space is given by 3 options.
[b] Answer space is given by 2 options (either *Yes/No* or *True/False)*
[c] Answer space is given by a real number; tolerance for correct answer is +/- 1%

and large set of verbal explanations justifying the optimal action in these tasks. Our focus on speech recordings arises from the important role of oral communication in information flows in the economy. The full set of instructions is reproduced in Appendix A.3. In the beginning, participants are told the following:

> *We are interested in how you would give advice in an informal conversation:*
>
> - *You should share an explanation behind your response.*
> - *Your recording will be played to a few other participants who will have to respond to the same question.*

We ask respondents to not search for answers on the internet.[2]

In the real world, people typically have time to think about their explanations. Correspondingly, rather than forcing respondents to talk immediately upon reading the question, we show them the question first and they can start their recording once they are ready. An example screen is shown in Figure 1.

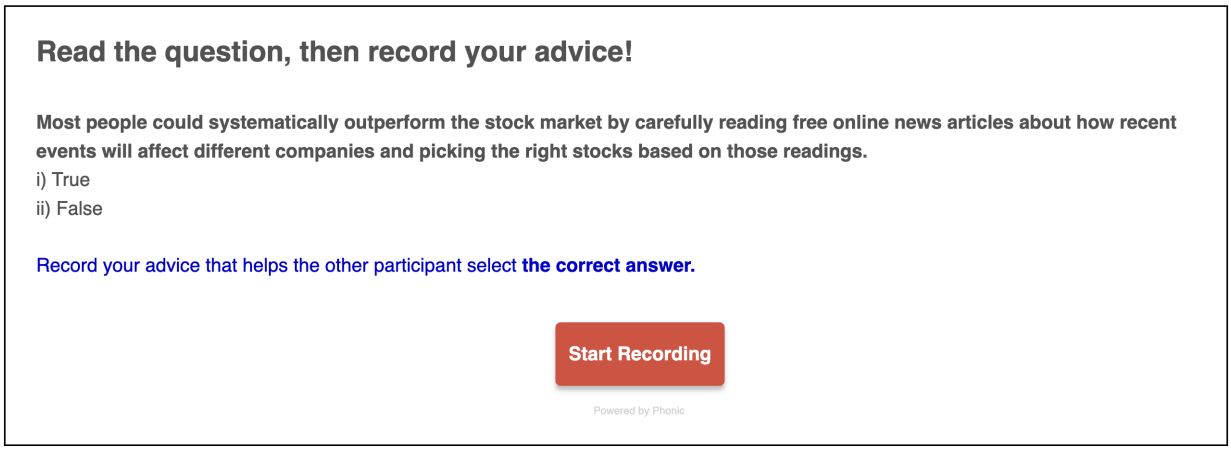


Figure 1: Recording screen from the Orator experiment.

After recording their explanation and selecting their answer, we elicit respondents' confidence in their answer by asking the question "*How certain are you that your above answer is optimal?*" on a scale from 0 "*Not at all certain*" to 100 "*Fully certain*".

Following that, we ask "*Suppose you can now choose how many additional participants your recording on this question will be shared with. Which number of other respondents would you like your recording to be shared with?*" and respondents can select a number between 0

[2]We ask participants at the end of the study whether they searched for any answers. There was no penalty for indicating that they did, but we exclude those observations from our analysis. 3.79% of participants indicated that they searched for answers.

and 10. We clarify that their decision will be played to some respondents in either case – such that they cannot opt out of their recording being shared altogether – but that for this question they should assume that they would receive an additional $1 for every additional listener who provides the right answer after listening to their answer, yet lose an additional $1 for every listener giving a wrong answer. This sharing decision provides a measure of the degree to which Orators perceive their explanation to be likely to induce Receivers to make the right decision.

Timeline. Respondents (i) read computerized instructions; (ii) are required to pass a comprehension check; (iii) read the first question and record their explanation; (iv) select their answer to the question; (v) state their confidence; (vi) make their sharing decision; (vii) and repeat (iii)-(vi) for the second task etc.

Incentives. Respondents received a reward of $6 for completing the study, which had a median completion time of 30 minutes. Moreover, 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 reasoning task. The Orator is matched to 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 main 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. We confirm that Orators understand these incentives using a control question. In Section 7, we explore the implications of an alternative incentive scheme, where the Orator is paid based on whether the Receiver imitates them or not, irrespective of the answer's accuracy.

Speech Recordings. Our Orator experiments rely on speech recordings of people's explanations. Relative to written text, speech recordings have a series of advantages: First, they are well-suited to study social learning which is often based on conversations. Indeed, as emphasized by Shiller (2020), much of the information we encounter comes in the form of orally transmitted narratives. Second, speech data allows us to capture critical features of explanations, including emotions and uncertainty markers, which cannot be measured as richly in text data. Third, writing text as opposed to spontaneously talking about one's thoughts adds another filter that may distort the measured explanations compared to the explanations appearing on top of people's minds.

2.4 Part 2: Receiver Experiment

To quantify the extent of learning and unlearning from verbal advice, we conduct a Receiver experiment, which uses the choices and recordings provided in the Orator experiment. We provide the full set of instructions in Appendix A.4.

As in the Orator experiment, respondents go through the fifteen reasoning tasks. We employ a within-design that proceeds in four steps in each round.[3] First, respondents read the financial reasoning problem and are incentivized to indicate their preferred choice, which we use as a measure of their prior belief. They then indicate their confidence in the accuracy of their response in the same format as respondents do in the *Orator Experiment*. Then, they either only learn about the choice of another randomly selected respondent in the Orator experiment (*Choice Only* treatment) or listen to a recording in which the Orator additionally provides an explanation (*Explanation* treatment). After that, the Receiver is again incentivized to provide their preferred choice, *posterior belief*, and then indicate their confidence. Moreover, Receivers evaluate the perceived accuracy of the Orator's answer. We discuss this measurement in Section 5.2.

Treatments. In the *Choice Only* treatment, Receivers may (rationally) infer and adjust their belief about the optimal answer from learning what someone else chose, even absent an explanation. The same channel is present in the *Explanation* treatment, but the explanation itself provides an additional source of learning. We randomize treatments between subjects, at the task level. For each task, 80% of Receivers are sampled into the *Explanation* condition, while the remaining 20% are assigned to the *Choice Only* condition. We oversample the *Explanation* condition, to have a sufficient amount of statistical power to examine heterogeneous effects by features of the explanations.

The comparison between *Explanation* and *Choice Only* allows us to identify the specific effect of listening to a recording providing an explanation on learning and unlearning, above and beyond the mere observation of another respondent's choice. The *Choice Only* condition is critical for controlling for (i) the effects of confidence, (ii) measurement error in priors, and (iii) confounders, such as experimenter demand effects.

Timeline. Respondents (i) read computerized instructions; (ii) are required to pass a comprehension check; (iii) provide their best answer to the first question and state their confidence; (iv) see the answer of a respondent in the Orator experiment in *Choice Only* and

[3]The within-design is critical as it allows us to measure persuasion rates at the individual level.

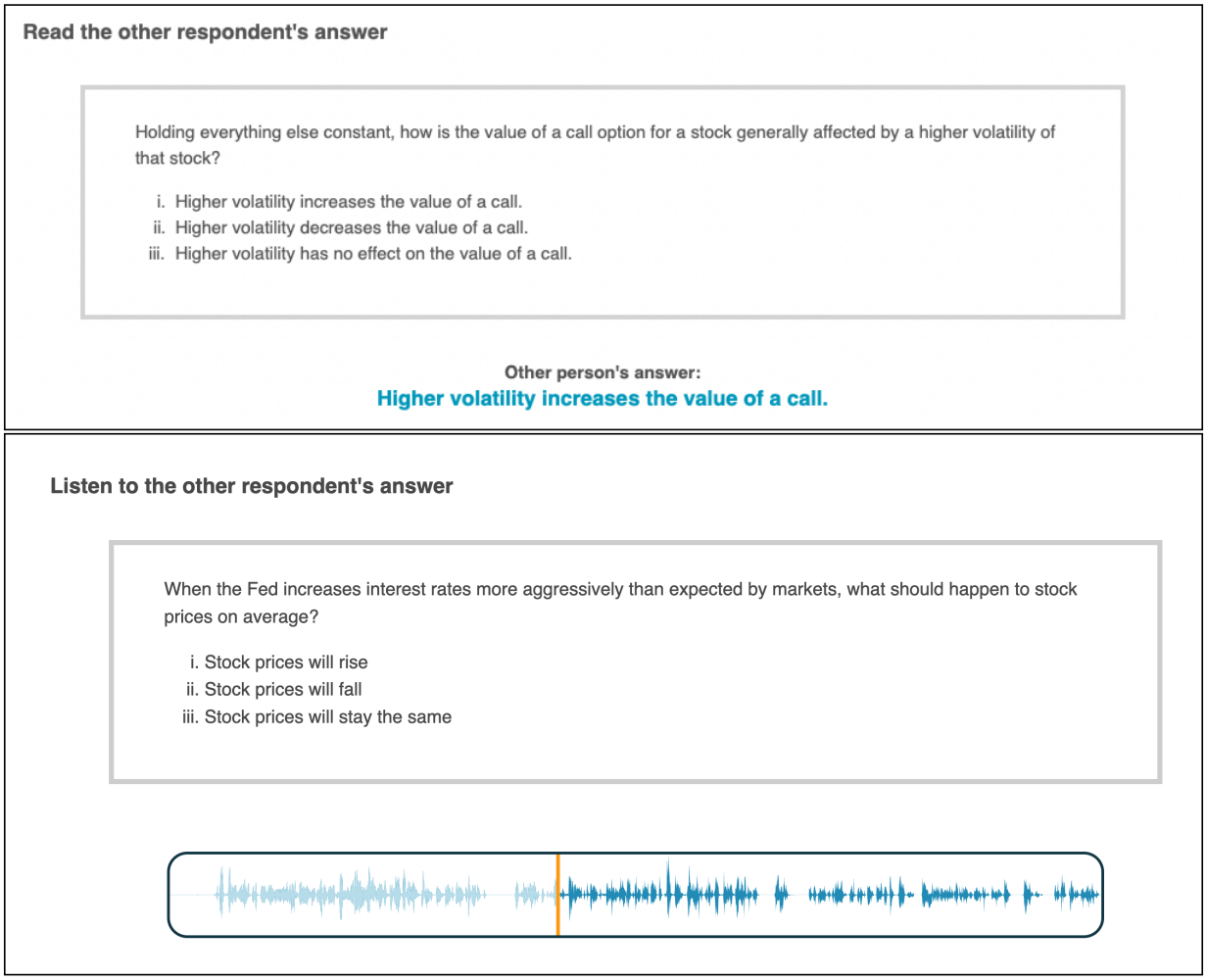


Figure 2: Other respondent's answer screens in the Receiver experiment, in the *Choice Only* treatment (upper panel) and *Explanation* treatment (lower panel).

listen to their explanation in the *Explanation* condition; (v) provide their best answer and state their confidence again, (vi) state the perceived accuracy of matched Orator's answer, and repeat (iii)-(vi) for the second task etc. We randomize the order of steps (v) and (vi) between respondents.

Incentives. Respondents received a reward of $6 for completing the study, which had a median completion time of 28 minutes. Moreover, respondents have a 10% chance of being eligible for an additional $10 bonus payment. Whether they would receive the bonus or not is determined based on 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.

2.5 Logistics

All experiments were conducted on the online platform Prolific, which is widely used for experiments in the social sciences (Eyal et al., 2021). The Orator experiment was run for a total of 240 U.S. respondents in July 2023, out of which 201 provided valid responses. Participants were required to have a working microphone to record their voice. The listener experiment was run with 542 U.S. respondents in July 2023, out of which 435 provided valid responses.[4]

Our Orator experiment yields a total of 3,105 valid recordings obtained by integrating speech recordings with *Phonic* into *Qualtrics* surveys.[5] Our transcription of recordings preserves natural language elements, such as hesitation markers ("um", "eh", pauses...), which we leverage to analyze mechanisms. Next to the transcriptions, the speech recordings lend themselves to the analysis of additional speech features, including emotions and voice features.

3 Explanations and the Contagion of Beliefs

Our analysis proceeds in four steps. In this section, we investigate the reduced-form effect of explanations on optimality rates and imitation patterns. In Section 4, we put forward a simple model that provides an explanatory framework for our reduced-form results and

[4]For both experiments we exclude participants who indicated that they looked up answers to the financial reasoning questions online.

[5]We rely on an Amazon Web Services (AWS) backend to stratify and distribute recordings into our Receiver experiment.

guides our mechanism analyses in Sections 5 and 6. Section 7 extends our setting to examine non-aligned incentives between Orators and Receivers.

3.1 The Effect of Explanations on Optimality Rates

We start by analyzing the effects of receiving information on others' choices on the average accuracy of choices in our sample. To do so, we compare how the posterior optimality rate differs from the prior optimality rate across the different treatments.

The prior optimality rate reflects the Receiver population's average knowledge of a task before being exposed to another's choice. The posterior optimality rate captures average accuracy after Receivers observe another respondent's choice only (*Choice Only* condition) or an additional verbal explanation (*Explanation* condition).

Figure 3 shows these optimality rates pooled across all 15 tasks. Before exposure, slightly more than half of the respondents provided correct answers (55% and 57% in *Choice Only* vs. *Explanation*, $p = 0.25$). We document two core findings on posterior optimality rates. First, observing another's choice does not affect average optimality (difference 0.81pp, $p = 0.77$). This suggests that the mere observation of someone else's choice does not have any noticeable average impact across a wide range of reasoning tasks in our study. Second, listening to another person's explanation does significantly increase the optimality rate by 4.8pp (0.14 SD, $p < 0.01$), corresponding to a 10 percent increase.

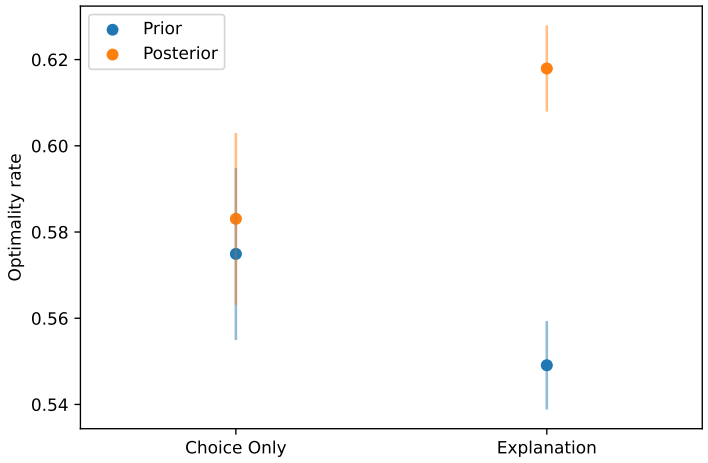


Figure 3: Prior and posterior optimality by treatment. *Notes:* Share of correct Receiver choices before and after exposure to the Orator's choice (*Choice Only*), or before and after exposure to the Orator's explanation (*Explanation*). Sample is the Receiver experiment (435 Receivers) subset to Orators with *Aligned Incentives* (100 Orators), resulting in 2967 observations. Whiskers show standard errors.

Result 1. *Only observing another respondent's choice does not improve optimality rates on average across all tasks. By contrast, additionally listening to another respondent's explanation strongly increases optimality rates on average.*

3.2 Patterns of Learning and Unlearning

Average optimality rates lump together all Receivers in the corresponding treatments, irrespective of whether a given Receiver was initially correct, and whether they were exposed to a confirming or conflicting response or explanation. However, changes in choices should be mostly driven by Receivers who obtain conflicting advice.[6] We therefore zoom into the two subgroups that drive most of the changes in overall optimality rates: people with incorrect prior choices exposed to correct choices on the one hand, and people with correct choices exposed to incorrect choices on the other. We refer to these as *learning opportunities* and *unlearning opportunities*, respectively. These two types of matches are equally frequent in our sample: there are 21.5% learning opportunities and 21.6% unlearning opportunities.[7]

Figure 4 displays the *learning rate* and the *unlearning rate* respectively. This figure illustrates two key findings. In *Choice only*, the learning rate is 32% while the unlearning rate is 21%. This suggests some moderate and marginally statistically significant learning gains arising from just observing another person's answer. In the *Explanation* condition, by contrast, learning opportunities are far more likely to be seized than unlearning opportunities: 49% of respondents imitate following a learning opportunity, yet only 21% imitate given an unlearning opportunity. This suggests a distinctive pattern of how verbal explanations shape learning and unlearning.

Result 2. *Compared to just observing another respondent's choice, additional exposure to a verbal explanation significantly increases the likelihood of seizing a learning opportunity. By contrast, exposure to a verbal explanation does not reduce the frequency of imitating those with an incorrect answer.*

In the following, we will provide a conceptual framework for these patterns and explore the underlying mechanisms.

[6] In fact, only approximately 10% of respondents who received a confirming response or explanation changed their choice.

[7] That these are equally common is not coincidental as illustrated by the model in Section 4.

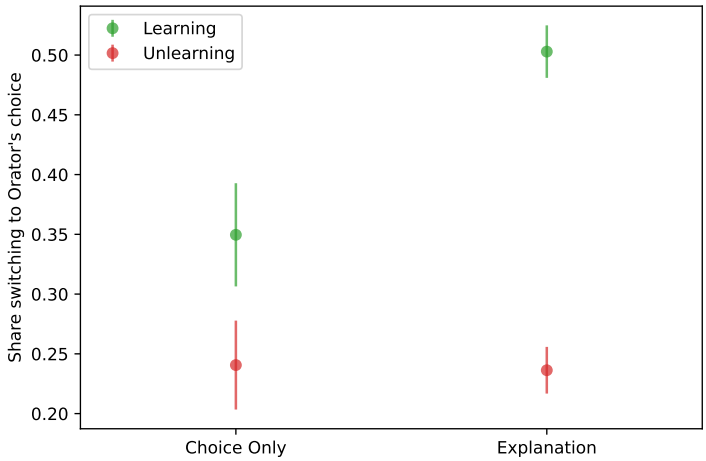


Figure 4: Share of Receivers switching to Orator's choice by treatment in learning and unlearning situations. *Notes:* In learning situations, initially incorrect Receivers are exposed to a correct Orator; in unlearning situations, initially correct Receivers are exposed to an incorrect Orator. Sample is the Receiver experiment (435 Receivers) subset to Orators with *Aligned Incentives* (100 Orators) and learning and unlearning situations, resulting in 1249 observations. Whiskers show standard errors.

4 Conceptual Framework

The purpose of this framework is to cast our experimental setup in terms of a standard belief formation setting that speaks to the existing economics literature. It serves to conceptualize our reduced-form findings and provide a guiding structure for 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 interpretations.

4.1 Setup

Consider a binary question with a correct answer $\omega = 1$ and an incorrect answer $\omega = 0$.[8] A decision-maker (DM) i enters with a prior belief p_i that the correct answer is 1 and chooses 1 if and only if that belief exceeds 0.5. The DM's prior answer is $x \in \{0, 1\}$. For simplicity, we assume that the agent's prior belief can be described by the functional form

$$p_i = \alpha_0 + \alpha_1 \mathbb{1}_{(x=1)} + \epsilon_i, \tag{1}$$

where $\mathbb{1}_{(x=1)}$ is the indicator function that the agent chooses the correct answer, and ϵ_i is a zero-mean noise term. We assume that this functional form is a probability, i.e. for all

[8]The binary setup is without loss of generality: the coding of correct vs. incorrect permits a binary classification of choices that applies to questions with multiple possible responses or a continuous scale.

realizations of ϵ_i, $0.5 \leq \alpha_0 + \alpha_1 + \epsilon_i \leq 1$ and $0 \leq \alpha_0 + \epsilon_i \leq 0.5$. To build intuition, suppose that $\alpha_0 = 0$ and $\alpha_1 = 1$. In that case, correct and incorrect Receivers would be perfectly confident in their respective answers and could never be convinced to change their mind. Consider instead a situation where $\alpha_0 = \alpha_1 = 0.5$. This corresponds to a situation where a DM taking the correct decision is perfectly confident, whereas an incorrect DM is not confident at all but is, at $p_i = 0.5$, perfectly indifferent between both actions.

The DM then observes a signal $s \in \{0, 1\}$, which is the realized answer of another respondent. To learn from the signal the DM needs to assign a *diagnosticity* to it, i.e. a belief about the likelihood that the observed answer matches the true state. We refer to agent i's perceived diagnosticity with $d_i = \mathbb{P}(\omega = s|s)$ and again assume that it can be represented by the functional form

$$d_i = \beta_0 \mathbb{1}_{ChoiceOnly} + \beta_1 \mathbb{1}_{(s=1)} \mathbb{1}_{ChoiceOnly} + \gamma_0 \mathbb{1}_{Explanation} + \gamma_1 \mathbb{1}_{(s=1)} \mathbb{1}_{Explanation} + \delta_i. \quad (2)$$

Here, $\mathbb{1}_{(s=1)}$ is the indicator function that the observed signal is answer 1 (i.e. the correct answer), $\mathbb{1}_{Explanation}$ and $\mathbb{1}_{ChoiceOnly}$ are treatment indicators, and δ_i a noise term. We assume that $d_i \in [0, 1]$ holds for all realisations of δ_i. Moreover, we assume that the DM never interprets an answer as evidence for its counterpart.

Assumption 1. *For all realizations of δ_i the perceived diagnosticity d_i is greater or equal to* 0.5.

An intuitive interpretation of d_i is as Receivers' perceived precision of the answer or explanation they are exposed to. Crucially, the drivers of these perceptions as captured by the treatment indicators are entirely different between *Choice Only* and *Explanation*. In *Choice Only*, perceived diagnosticity is only affected by the very fact that this means that another respondent's best answer was such. The *Explanation* treatment nests this source of learning, but additionally provides a host of additional ways to infer the perceived precision. The richness of verbal expressions in natural language, as well as features of paralanguage such as prosody or tonal emphasis, may provide insights into the accuracy of the Orator, all of which will be reflected in the parameters of equation 2.

Given a prior belief p_i, a signal s, and a perceived diagnosticity d_i, the DM updates their

belief according to Bayes' rule, which yields their posterior belief π_i that action 1 is correct:

$$\pi_i(s=1) = \frac{p_i \cdot d_i}{p \cdot d_i + (1-p_i) \cdot (1-d_i)} \tag{3}$$

$$\pi_i(s=0) = \frac{p_i \cdot (1-d_i)}{p \cdot (1-d_i) + (1-p_i) \cdot d_i}. \tag{4}$$

As before, the DM chooses action 1 if and only if $\pi_i > 0.5$. We refer to the posterior answer using $y \in \{0,1\}$.

Our baseline setup puts structure on two central objects of interest in our experiment: people's prior confidence (or meta-cognition), p_i, and the perception of others' behavior's diagnosticity, d_i. The framework pins down the calibration of these objects, i.e. how they are related to the true state. First, the calibration of confidence is determined by α_1, which is the difference in prior confidence between initially correct and incorrect DMs. Second, the calibration of perceived diagnosticity is governed by the parameters β_1 and γ_1, which determine differences in the perceived accuracy of correct and incorrect observed answers in the context of the *Choice Only* and *Explanation* treatments, respectively.

4.2 Analysis

We characterize the effect of the different treatments on learning, unlearning and optimality rates. A first observation is that due to Assumption 1, signals that coincide with the Receiver's prior choice do not lead to a change from prior to posterior choice.

Proposition 1. *If the prior action coincides with the signal, $s = x_i$, then behavior does not change, $x_i = y_i$.*

Note that while Proposition 1 establishes no switching away from one's already preferred action when receiving a supportive signal, a DM will in this case indeed become more confident, i.e. form a more extreme belief, $|\pi_i - 0.5| > |p_i - 0.5|$.

Learning and unlearning rates. Therefore, the two cases of interest occur when the DM initially chooses the wrong answer and is presented with a correct signal, or when they initially choose the incorrect answer and are presented with an incorrect signal. In the first scenario, the signal can drive the agent from an incorrect to a correct answer. This is what we refer to as *learning*. In the second scenario, the opposite can happen and the agent can switch from a correct to an incorrect answer. We call this *unlearning*. The quantities of

interest are the rates of learning and unlearning, which are given by

$$l = \mathbb{E}[\mathbb{1}(\pi_i(s=1) > 0.5) \mid p_i < 0.5]$$
$$u = \mathbb{E}[\mathbb{1}(\pi_i(s=0) < 0.5) \mid p_i > 0.5].$$

The following result establishes how these rates depend on the parameters of the functional forms.

Proposition 2. *The learning rate l always rises in α_0, and further rises β_0 and β_1 (γ_0 and γ_1) in the Choice Only condition (Explanation condition). The unlearning rate u always falls in α_0 and α_1, and rises in β_0 (γ_0) in the Choice Only condition (Explanation condition).*

To build intuition about the drivers of the learning rate, note that a higher α_0 means that an initially incorrect DM is less confident, i.e. has a belief closer to 0.5, and it therefore takes a (perceived) less precise signal to move them over the behavioral threshold of 0.5. β_0 and γ_0 capture the baseline of perceived diagnosticity in *Choice Only* and *Explanation*, respectively; an increase in these parameters will make *all* observed signals be perceived as more convincing and make it more likely that the DM's belief is moved enough to change actions. A higher β_1 specifically makes seeing a correct signal more convincing, which is the relevant signal in learning opportunities, and similarly for γ_1 in *Explanation*.

Similarly, for unlearning opportunities, lower a_0 and a_1 mean that the initially correct Receiver is less confident in their choice, i.e. has a prior belief closer to 0.5, which implies that a (perceived) less precise signal is needed to move them below 0.5 and thus convince them to switch actions. The perceived persuasiveness of an incorrect signal increases in β_0 in *Choice Only* and in γ_0 in *Explanation*.

Optimality rates. Next, we turn to the expected rate of correct choices across the subject population. We denote the optimality rate prior to signal observations by θ^{pre}, defined as $\mathbb{E}[\mathbb{1}_{p_i \geq 0.5}] = \mathbb{P}[p_i \geq 0.5] \in [0,1]$. In correspondence to the random matching mechanism of our experimental design, we assume that each DM's signal is uniformly drawn from the pool of choices in the population. Therefore, the expected fraction of participants with a correct answer observing an incorrect signal equals $\theta^{pre} \cdot (1 - \theta^{pre})$, which is, at the same time, the expected fraction of participants with an incorrect answer observing a correct signal. This is a simple but crucial insight that may be counter-intuitive at first. Compare two tasks, with the second exhibiting a higher baseline optimality rate before exposure to others. Two forces are simultaneously at play once interaction occurs: first, on the Receiver

side, a higher baseline rate means there are more correct and fewer incorrect Receivers, implying more capacity for potential unlearning by initially correct and less unlearning by initially incorrect Receivers. Second, on the Orator's side, a higher baseline rate means that there are more correct and fewer incorrect Orators, so random matching implies less capacity for potential unlearning and more for learning. In terms of the resulting frequency of learning and unlearning opportunities, these forces exactly offset each other, so that there will always be an identical fraction of learning and unlearning opportunities, in expectation. Formalizing this observation, note that the expected fraction of Receivers with a correct posterior answer, denoted by θ^{post}, equals $\theta^{pre} + [\theta^{pre} \cdot (1 - \theta^{pre})] \cdot (l - u)$. This yields the following result.

Proposition 3. *The posterior optimality rate exceeds the prior optimality rate if and only if the learning rate exceeds the unlearning rate. The posterior optimality rate rises with the learning rate and falls with the unlearning rate.*

The implication of Proposition 3 is that the analysis of the learning and unlearning rates directly extends to the analysis of optimal rates. The first part highlights a critical reduced-form relationship between learning, unlearning and optimality rates: the sign of the difference between learning and unlearning rate determines whether there is an aggregate improvement or not, i.e. whether the posterior exceeds the prior optimality rate. This provides a simple formal justification to study learning and unlearning rates as the drivers of aggregate improvements, as we did in Section 3. Crucially, this result is entirely independent of the prior optimality rate in a given task.

The second part establishes that, conditional on the sign of $(l - u)$, the importance of learning and unlearning rates is governed by $\theta^{pre} \cdot (1 - \theta^{pre})$, which expresses the frequency of both of learning and unlearning opportunities as a function of the prior optimality rate. Intuitively, when the prior optimality rate is closer to $\frac{1}{2}$, opportunities for learning and unlearning become more frequent, and the impact of the imitation rates in these situations on the posterior optimality rate becomes greater.

Linking model to data. Before delving into an empirical exercise motivated by this framework, we point out what our reduced-form findings imply in terms of the model. Our main pattern as stated in Result 2 is a differential effect of explanations (relative to mere observation) on learning versus unlearning. First, it implies that $\gamma_0 \approx \beta_0$. This means that the perceived diagnosticity of incorrect answers is similar irrespective of whether the Receiver just learned about the Orator's choice or also listened to their explanation. Intuitively, explanations associated with incorrect answers do not provide Receivers with any additional

insight that the corresponding answer is incorrect, on average. Second, our finding implies that $\gamma_1 > \beta_1$. This means that relative to the perceived diagnosticity of incorrect answers, correct answers are associated with a higher increase in perceived diagnosticity under explanations than mere observation. Put differently, explanations associated with correct answers boost perceived diagnosticity. In the following section, we bring elements of the framework to the data and experimentally study the driving forces of imitation.

5 Imitation and Perceptions of Others' Accuracy

In the following two sections, we investigate the mechanisms underlying our reduced-form findings. We first study the cognitive determinants of imitation decisions as suggested by our framework. We then move to a systematic analysis of the structure of explanations provided in the experiments, and how features of explanations and their interpretation by Receivers help understand imitation patterns.

5.1 What Drives Imitation Decisions?

Our model shows that learning and unlearning rates depend on two main constructs, the agent's prior confidence (1) and perceptions of others' accuracy (2). More specifically, Proposition 2 establishes that learning and unlearning rates decrease in Receivers' prior confidence in their choices (as reflected in the distance between p_i and $\frac{1}{2}$) and increase in the perceived accuracy of the conflicting answer they observe.

However, our main finding, Result (2), concerns the comparison between the *Explanation* and *Choice Only* treatments, i.e. the additional effect of the provision of an explanation on imitation rates. Importantly, note that the meta-cognitive part is held constant across these treatments: due to random sampling, prior confidence should be identical in both groups. As (1) clarifies, the expression of prior confidence does not depend on the treatment. This sets what happens in our experiments conceptually apart from the literature on confidence, including recent work in economics on this topic (e.g., Enke et al., 2023). Perceptions of others' accuracy, by contrast, may differ between treatments (2), and this is the only component of the model that can explain differences in imitation rates between the two treatments. Put differently, our framework shows that point differences in imitation rates between *Explanation* and *Choice Only* must be driven by differences in perceptions of others' accuracy induced by the treatments. These perceptions in turn critically hinge on the content of the explanations, as we will explore in depth in Section 6. The insight into

the central role of perceived diagnosticity motivates an analysis of the perceptions of others' accuracy underlying imitation decisions. While some work has been done on the related issue of lying detection (e.g., Serra-Garcia and Gneezy (2021)), comparably little is known about how good people are at detecting whether someone they observe holds a correct or mistaken belief about a decision problem.

5.2 Measuring Perceptions of Others' Accuracy

We elicit perceptions of others' accuracy in the listener experiment following the elicitation of the posterior action and confidence. Respondents state this perception both in the *Explanation* and *Choice Only* treatments. Note that in the *Choice Only* condition, Receivers can learn from the very fact that they see a specific choice: in canonical models of social learning, knowing another person's best response carries information (Mobius and Rosenblat, 2014). In the *Explanation* treatment, Receivers can additionally learn from the content of the explanation as well as features of the voice message. On a separate screen, we ask "*Having [listened to the respondent's recording / seen the respondent's answer], what do you think is the percent chance that the choice the other respondent gave is correct?*". Participants answer using a slider on a scale from 0% (labeled *Incorrect with certainty*) to 100% (labeled *Correct with certainty*). A screenshot of the elicitation screen is displayed in Appendix Figure A1.

5.3 The Calibration of Perceptions of Others' Accuracy

Perceived Accuracy and Imitation. We start by examining the link between perceptions of others' accuracy and imitation decisions as posited by the model. We compute correlations between perception scores and an indicator for whether the respondent imitated. We find strong relationships across the board, at $r = 0.56$ (*Choice Only*) and $r = 0.62$ (*Explanation*) for learning opportunities, and $r = 0.48$ (*Choice Only*) and $r = 0.51$ (*Explanation*) for unlearning opportunities. This indicates, not surprisingly, that the likelihood of imitation is strongly associated with the perception of whether the other person is correct. Note that accuracy perceptions are somewhat more predictive of imitation in unlearning than learning opportunities.

Explanations and Perceived Accuracy. Turning to treatment differences, Figure 5 shows average values of others' perceived accuracy across our two treatments, separately for learning and unlearning opportunities. We document a pattern that very closely mimics the pattern of imitation decisions in Figure 4. First, perceptions of accuracy are higher for correct

than incorrect answers in both *Explanation* and *Choice Only*.[9] The size of this difference is much more pronounced in *Explanation*, mirroring the larger gap in imitation rates observed in Figure 4. Second, considering point estimates we find that perceptions of others' accuracy are not statistically different between *Choice Only* and *Explanation* for unlearning opportunities, at 41.62 and 37.65 (p=0.25), respectively. For learning opportunities, by contrast, average perceptions of others' accuracy (54.82) in *Explanation* significantly exceed those in *Choice Only* (49.01, $p = 0.03$ for the treatment difference). This, again, very closely mirrors the pattern in imitation from Figure 4. Taken together, the patterns of perceived accuracy track imitation behavior in a striking fashion, strongly supporting the model insight on the central role of perceptions of others (rather than meta-cognition).

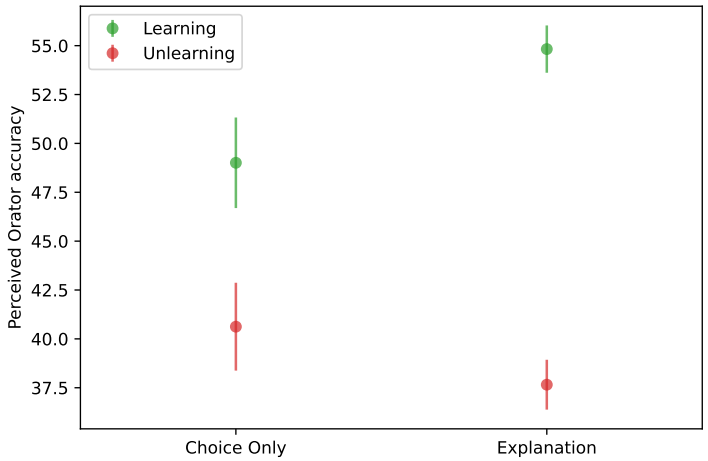


Figure 5: Perceived Orator accuracy by treatment in learning and unlearning situations. *Notes:* Sample is the Receiver experiment (435 Receivers) subset to Orators with *Aligned Incentives* (100 Orators) and learning and unlearning situations, resulting in 1249 observations. Whiskers show standard errors.

Note that this perspective focuses on subgroups in which all participants were either exposed to a correct or an incorrect choice. As a result, this analysis cannot provide any insight into how perceptions of accuracy are related to the actual accuracy of the Orator's answer.

The Calibration of Accuracy Perceptions. An alternative perspective is the relationship between perceptions of others' accuracy and the accuracy of the Orator. To this end, we

[9]There are various possible reasons for the difference in perceptions between correct and incorrect responses in *Choice Only*, where no explanation or further information is received. For example, observing an answer might prod respondents to think about a justification, and they manage to come up with better justifications for correct answers. Moreover, note that the samples of Receivers are systematically different: the perceptions of accuracy come from individuals with incorrect prior in learning opportunities, but individuals with correct priors in unlearning opportunities.

can look at how perceptions of accuracy change when exposed to the response of or an explanation by a correct or an incorrect respondent. Separately for Receivers with a correct and incorrect prior, we compute the correlation between the perceived accuracy of the Orator's choice and whether the choice was actually correct.

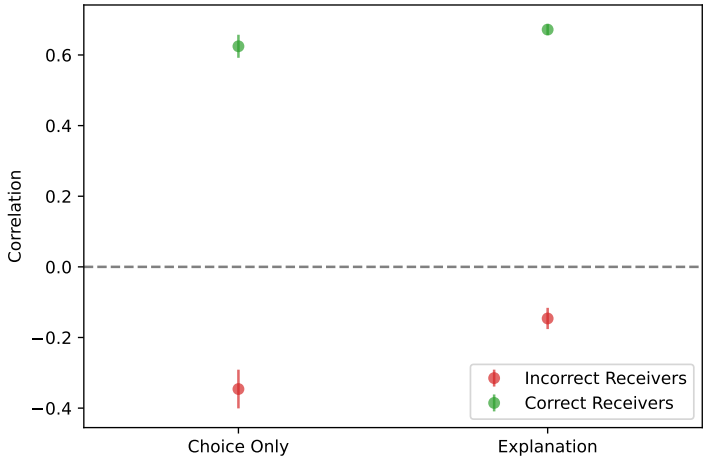


Figure 6: Correlation of perceived accuracy and Orator optimality by treatment in learning and unlearning situations. *Notes:* Sample is the Receiver experiment (435 Receivers) subset to Orators with *Aligned Incentives* (100 Orators), resulting in 2967 observations. Whiskers show standard errors.

Figure 6 shows the results. We find that among Receivers with a correct prior, exposure to correct (as opposed to incorrect) responses are associated with a similarly strong increase in perceptions of accuracy in *Choice Only* and *Explanation*, with correlations of $r \approx 0.62$ and $r \approx 0.67$ ($p = 0.18$), respectively. Among Receivers with incorrect prior choices, the correlation between perceived accuracy and true accuracy is naturally negative: Those people start out believing in a wrong answer, and they will therefore perceive the same wrong answer by someone else as similarly accurate. Seeing or hearing about a correct answer, by contrast, will typically be perceived as *less accurate*, as it differs from one's own answer, which implies a negative correlation. The less negative this correlation, the less people with incorrect priors fall prey to this issue of not recognizing correct Orators. In *Choice Only*, this results in a correlation of $r \approx -0.35$. However, this correlation becomes weaker, i.e. *better calibrated*, in *Explanation*, with $r \approx -0.15$ ($p < 0.01$). This suggests that the additional exposure to a verbal explanation allows individuals with a wrong prior to form more calibrated assessments than in the absence of such verbal explanations, but not individuals with correct priors.

Result 3. *Perceptions of others' accuracy are highly predictive of imitation decisions. Average perceptions across treatments closely mirror imitation rates in both learning and unlearning*

opportunities. Explanations make initially incorrect Receivers much more responsive to the accuracy of a response compared to just observing someone's choice, but explanations have no such effect on Receivers with correct priors.

6 The Supply and Interpretation of Explanations

The central methodological tool of our experiments, speech recordings, allows us to investigate the nature of explanations and the mechanisms underlying perceived accuracy and imitation decisions. To this end, we transcribe all voice recordings and analyze their content. We then proceed in two steps. First, we examine which attributes are predictive of whether an explanation is associated with a correct or an incorrect choice. Second, we study which attributes are predictive of a high perceived accuracy and likelihood of imitation, and whether there are any differences between learning and unlearning opportunities.

These two empirical exercises correspond to two distinct hypotheses about the origins of our main finding, Result 2. As for the first type of analysis, the underlying hypothesis is that differences between explanations of correct and incorrect choices account for the differences in treatment effects of *Explanation* between learning and unlearning opportunities. Correct transcripts may be systematically different from incorrect transcripts in a way that makes it easier for people to identify their actual accuracy above and beyond the choice itself. For example, they might contain elements that communicate the precision of the explanation, e.g., by providing justifications or arguments, whereas explanations in unlearning opportunities may tend to lack content that allows one to easily identify the answer as incorrect.

The hypothesis underlying the second approach is based on the observation that the groups of Receivers exposed to learning and unlearning opportunities are different – namely, respondents with different prior accuracy –, and that they might also differ in their capacity to tell apart correct from incorrect explanations. Specifically, for this mechanism to explain the main finding of explanations only helping in learning opportunities, Receivers with incorrect priors would need to be *relatively* better at identifying correct explanations than Receivers with correct priors are at discerning incorrect explanations. To the extent that correct priors are related to cognitive skills, this explanation may sound counter-intuitive: intuitively, the more cognitively skilled would need to be relatively worse at interpreting explanations. Our data allow us to shed light on both of the above hypotheses.

6.1 Features of Explanations

For each of our 3,215 voice messages, we obtain a text transcript. These text transcripts are coded to closely mirror the actual spoken message, i.e. they include a wide array of features of spoken language that are absent from normal written language, such as disfluencies and hesitation markers ("um", "eh"). To analyze the frequency of different attributes of explanations across conditions, we require a taxonomy of potentially relevant components.

We leverage existing work on explanations (Lombrozo, 2006, 2016) to generate a long list of 28 features of verbal explanations. We then design a coding scheme that allows us to identify and quantify the presence of each of these elements in transcripts using a popular large-language model, GPT-4. Appendix B.1 provides additional details on the annotation, Table A1 describes the features, and Appendix Table A2 provides basic characteristics in the sample.

6.2 The Features of Correct and Incorrect Explanations

The hypothesis motivating this analysis is that explanations accompanying correct answers exhibit relatively more instances of features that make it possible to infer the answer's accuracy. Notably, this hypothesis cannot be examined based on the frequency of features that are naturally more strongly associated with correct than incorrect answers. Take the example of phrases expressing high confidence. To the extent that correct answers tend to be associated with higher confidence (which they are), one would expect more high-confidence statements in explanations of correct answers. However, one would similarly expect more statements of low confidence in explanations of incorrect answers. Both of these are, in fact, true in our data. We will refer to this class as *specific* features, in the sense that they are naturally more strongly related to correct or incorrect explanations. Appendix Figure A3 displays the share of explanations exhibiting different specific features, separately for those associated with correct and incorrect answers.

Consistent with Orators having some awareness about the accuracy of their choice, explanations of correct answers on average invoke more logical arguments and more statements of high confidence but significantly fewer apologies, questions, self-corrections and statements of low confidence.

However, these features do not directly speak to our hypothesis: instead, it is the frequency of statements that are, per se, *unspecific* to correct or incorrect explanations, that matter. Put differently, we want to examine types of attributes that plausibly occur in correct explanations as much as in incorrect explanations, and that may help to infer the degree

of reliability or precision of an explanation. Such unspecific features naturally compromise expressions that can go either way: for example, a feature capturing *any expression indicating the level of confidence* would include both high and low confidence statements, and it is the total frequency of their occurrence in explanations of correct versus incorrect answers that will be telling about their structural difference.

From our total set of features, five qualify as unspecific. Figure 7 plots the share of explanations with each feature, separately for the set of correct and incorrect answers. We document strong evidence for the differential presence of these features in explanations of correct versus incorrect answers. Explanations of correct answers exhibit a higher share across all five measures, and significantly so in four of them.

Correct explanations tend to include more of those elements that allow the listener to infer the answer's accuracy. The lower frequency of such attributes in incorrect answers likely impairs the ability of Receivers to detect their inaccuracy. In terms of our main finding on the differential benefit of explanations in learning opportunities, this provides strong evidence in favor of the first hypothesis examined in this section: the nature of explanations in learning opportunities provides more ways to identify the corresponding answer as correct than explanations in unlearning opportunities provide ways to tell the associated choice is incorrect.

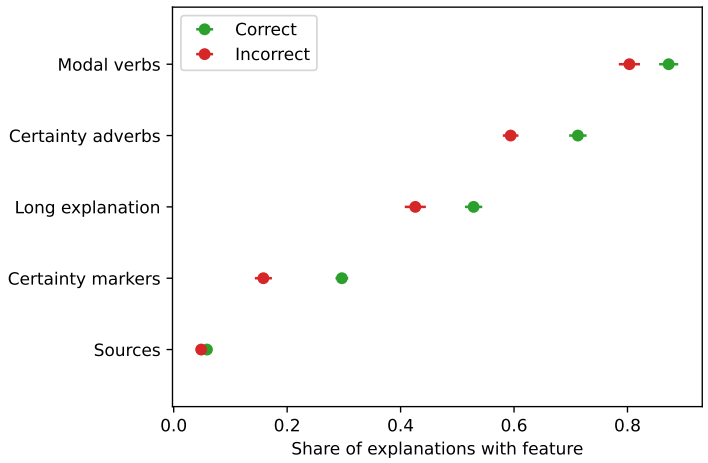


Figure 7: Unsigned features in explanations. *Notes: Modal verbs*, *Certainty adverbs*, *Certainty markers* and *Sources* are identified by annotating explanation transcripts using GPT-4 (see Section B.1). *Long explanation* equals 1 if an explanation has more words than the median and 0 otherwise. Unsigned features are markers that provide a direct way to infer the reliability of an explanation, irrespective of whether it is high or low. Sample is the Orator experiment subset to *Aligned Incentives* (106 Orators), resulting in 1589 observations. Whiskers show standard errors.

6.3 The Interpretation of Explanations

We now turn to our second approach, which concerns a complementary hypothesis: a source of the differential effect of explanations in learning opportunities is that Receivers in unlearning opportunities are relatively worse at identifying that an explanation is wrong than Receivers in learning opportunities are at discerning whether an explanation is right. The notion of *relative* is critical here because we are comparing different samples assessing different explanations for correct and incorrect answers: specifically, we mean the average improvement in perceived accuracy delivered by an explanation above and beyond just seeing another's answer without an explanation attached. Put differently, we benchmark the *Explanation* condition against the *Choice Only* treatment. We run regression analyses testing the relationship between features of explanations and Receivers' perceptions of accuracy. Unlike in the preceding section, this hypothesis is not limited to unspecific features. Instead, we intend to examine the responsiveness to *any* attributes of explanations that are indicative of accuracy as established in the preceding subsection. To accommodate the multicollinearity among all of our features, we first perform a principal components analysis to reduce the feature space. Details of the approach are provided in Appendix B.2. We observe a drop in eigenvalues between the sixth and seventh principal components and thus focus our analysis on the first six. Based on the feature loadings of the different principal components, we can relatively precisely identify which type of speech characteristics they capture. Using our full sample, we regress standardized perceptions of accuracy on the six principal components. Importantly, we control for the average perceived accuracy by configuration (right or wrong Receiver prior, right or wrong Orator) in the *Choice Only* treatment: this ensures that we can interpret the remaining variation in perceived accuracy as arising from the explanations only.

The left-hand panel of Figure 8 illustrates our results using a coefficient plot, showing coefficient estimates for the different labeled principal components. We find that, in the full sample, there is some responsiveness to the main margins of variation of speech features. Specifically, we document that perceptions of accuracy are strongly negatively related to short and abstract explanations, and somewhat positively related to more confident and complete explanations as well as those invoking external authorities and exhibiting more disfluencies.

The right-hand panel of Figure 8 tests our hypothesis directly, by looking at whether there are indications for differential responsiveness to speech features between Receivers in learning and unlearning opportunities. We find no significant differences in five out of our

six principal component features. The only significant difference emerges for the category of external authorities: respondents in learning opportunities are estimated to be somewhat more sensitive to the presence of this feature.

This set of results paints a clear picture: we do not find compelling evidence in favor of our second hypothesis. While confirming the importance of our identified speech features for perceptions of accuracy, we conclude that, at the very least, initially incorrect respondents do not seem better at responding to relevant features of explanations than initially correct respondents. This is reassuring to the extent that higher prior accuracy of respondents is correlated with cognitive skills, and this mechanism would have implied that those with higher cognitive skills are worse at interpreting explanations.

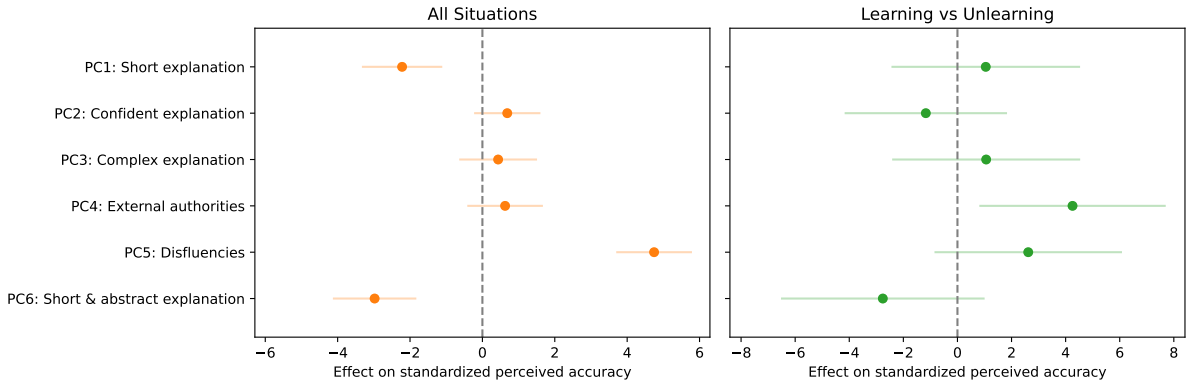


Figure 8: Effect of explanation features on perceived accuracy. *Notes:* The 28 explanation features are aggregated into 6 principal components (see Appendix B.1). Left panel shows effect of components on standardized accuracy, pooling all situations. Right panel shows effect of components in learning situations compared to unlearning situations, i.e. the interaction effect between learning situations and each principal component in a regression using only learning and unlearning situations. Both panels control for Orator correctness. To ensure that the variation arises from explanations only, both panels control for the average perceived accuracy by configuration (right or wrong Receiver prior, right or wrong Orator) in the *Choice Only* treatment. Sample for the left panel is the Receiver experiment (435 Receivers) subset to Orators with *Aligned Incentives* (100 Orators) and the *Explanation* treatment, resulting in 2353 observations; the right panel is additionally subset to learning and unlearning situations, resulting in 993 observations. Whiskers show 95% confidence intervals.

Result 4. *To account for the asymmetry in treatment effects for learning and unlearning opportunities, we establish the following facts:*

- *The differences for correct and incorrect answers therefore partly account for the asymmetry: Explanations associated with correct answers are more likely to include features that allow inference about their accuracy.*
- *We do not find strong evidence that the asymmetric effect is driven by the differential*

interpretation of features of explanations by Receivers in learning and unlearning opportunities.

6.4 The Social Determinants of Accuracy Perceptions

Our preceding mechanisms analyses aimed to explain how explanations affect imitation rates by analyzing features of the explanation in scripts. In practice, however, our perceptions of accuracy are likely shaped by a variety of cues delivered through verbal explanations, including the characteristics of the Orator. We conclude our mechanism analyses by examining whether perceptions of accuracy respond to such *social cues*. This exercise is largely orthogonal to the above parts: by construction, such correlational relationships cannot explain our treatment effect.

Using the full sample in the *Explanation* treatment, we run regressions of perceived accuracy on different sets of respondent characteristics. To account for differences in accuracy among different demographic groups, which might drive some of the perceptions, we control for the actual accuracy in our regressions. Panel A of Figure 9 shows coefficients from a regression of accuracy perceptions on whether the Orator was accurate or not and a variety of socio-demographics of the Orator. We document that various Orator characteristics are strongly associated with perceived accuracy. Specifically, we find that recordings from male and educated Orators are more likely to be perceived as accurate ($p < 0.01$ and $p < 0.1$, respectively), whereas black and Republican Orators are less likely to be judged as correct ($p < 0.1$ and $p < 0.01$, respectively). Age and employment status of the Orator do not seem to play an important role in our data.

While panel A considers the effect of Orator characteristics independent of the Receiver characteristics, we next consider the similarity between the Orator and respondent: intuitively, the social distance or perceived similarity of identities plausibly plays a role in whether a Receiver perceives the Orator as accurate. Panel B of Figure 9 analyzes how the similarity between Orator and Receiver shapes accuracy perceptions, controlling for both Orator and Receiver characteristics. For every characteristic, we report the effect of an indicator for whether the Orator and Receiver share said feature. Strikingly, we find that perceived accuracy strongly increases for black Orators when the Receiver is also black ($p < 0.01$). Likewise, employed Orators are somewhat more likely to perceive other employed Orators as accurate ($p < 0.10$). Similarity in the other socio-demographic characteristics, on the other hand, is not predictive of perceived accuracy.

As a placebo check, Appendix Figure A2 confirms that Orator characteristics and the simi-

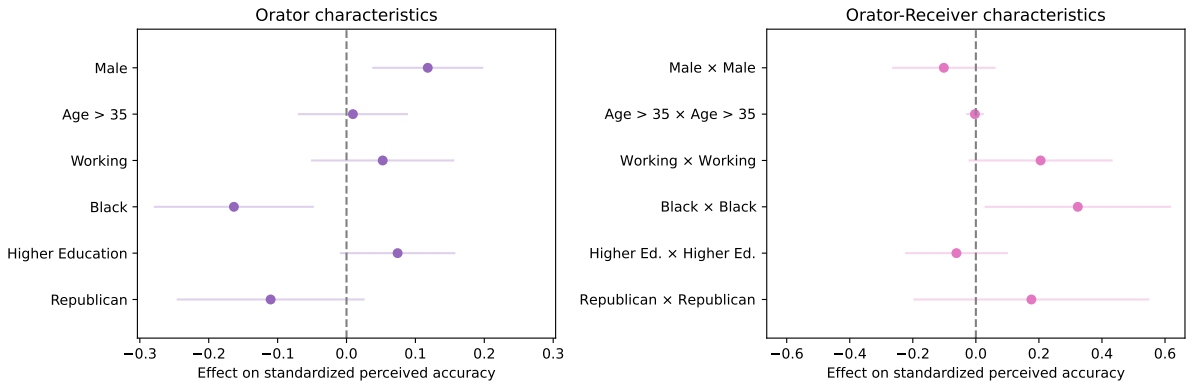


Figure 9: Effect of Orator and Orator-Receiver characteristics on perceived accuracy in *Explanation* treatment. *Notes:* Left panel shows effect of Orator characteristics on standardized perceived accuracy. Right panel shows the interaction term in a regression of standardized perceived accuracy on Orator characteristics, Receiver characteristics and Orator-Receiver interactions. All regressions control for the actual accuracy of the choice associated with a recording. *Higher Education* is equal to 1 if respondents have a Bachelor's or more. *Age > 35* is a binarization based on the approximate median in our dataset. Sample is the Receiver experiment (435 Receivers) subset to Orators with *Aligned Incentives* (100 Orators), resulting in 2967 observations. Whiskers show 95% confidence intervals.

larity of Orator and Receiver characteristics are not correlated with perceived accuracy in the *Choice Only* treatment. This suggests that mechanical factors cannot account for the above patterns, but instead that the voice recording is revealing of the Orator's background characteristics.

7 Strategic Incentives on the Marketplace of Ideas

Our main experiment emulates a marketplace of ideas on which different agents have aligned preferences to find out the ground truth. Many real-life settings such as conversations at social gatherings in which people learn from strangers have this feature, i.e. the person sharing their explanation has no immediate, tangible incentive other than helping the Receiver make better choices.[10] At the same time, many other situations involve interactions with strategic incentives, where the person delivering an explanation benefits from the Receiver taking a specific action. This includes politicians, social media influencers, marketers, and salespeople.

In practice, individuals face incentive schemes at various intermediate stages between the two extremes of perfectly aligned objectives and wanting to persuade of a specific, possibly wrong response. How do strategic incentives shape the nature of explanations and out-

[10]Note that this paper is concerned with "non-motivated" questions where people typically have no motivations regarding the true state, unlike, for example, politics or cultural topics.

comes on the marketplace of ideas? To examine this question systematically in our setup, we designed an additional experimental condition in our Orator study.

7.1 Design

This *Imitation Incentives* condition was run as an additional condition of our baseline study presented in Section 2. It is identical to the baseline treatment of the Orator study, referred to as *Aligned Incentives* here, except for incentives: in particular, the bonus of Orators in *Imitation Incentives* depends on whether the Receiver makes *the same* choice, rather than the accuracy of the Receiver's choice. Respondents in *Imitation Incentives* receive the following instructions:

> *Whether you receive the bonus payment of $10 only depends on whether the other participant gives the same answer as you. What this means is that for your bonus payment, it does not matter whether you know the right answer. It only matters that you make the other participant follow your answer, whether it's right or wrong. To maximize your chances of receiving the bonus payment, you should give advice that leads the other participant to select the same answer as you!*

After passing a comprehension check that includes a control question about the incentive scheme in both conditions, respondents in *Imitation Incentives* additionally see a screen that provides eight suggestions on what might make their explanations more likely to trigger imitation (see Appendix A.3).

The answers and explanations from the *Imitation Incentives* condition were included in the main collection of the Receiver experiment. Crucially, note that Receivers were never explicitly informed about the incentive condition under which the answer or explanation they saw was generated. Rather, they were simply informed in the instructions that "For some questions, you will listen to a voice message of another person once." This absence of an explicit statement about the Orator's incentive mirrors most real-life situations, where the incentives of people who advise us may be somewhat vague or unknown. An alternative is to explicitly inform Receivers about the incentive scheme underlying the next answer. This would allow us to study to which extent people can internalize the incentive schemes of others they observe, as in the research of Cain et al. (2005), for example. We believe this to be a promising extension of our setting.

7.2 How do Strategic Incentives Affect Optimality Rates, Learning and Unlearning?

Figure 10 shows results from the *Imitation Only* treatment on optimality rates among Receivers as well as learning and unlearning rates in comparison to the *Aligned Incentives* condition. The left-hand panel displays average optimality rates across all tasks. We find that average posterior optimality does not differ between the two incentive treatments ($p = 0.34$). Put differently, paying people to be imitated rather than to spread truth neither worsens nor improves Receivers' choices, on average. This might have several reasons, including that the incentive manipulation did not change Orator behavior. The right panel zooms in by showing the likelihood of imitation, separately for learning and unlearning opportunities. We find that *both* learning and unlearning rates significantly increase under *Imitation Incentives*. Learning rates increase from 47% to 54% ($p = 0.02$), while unlearning rates increase from 24% to 27% ($p = 0.23$). Therefore, imitation incentives increase the dispersion in net learning rates.

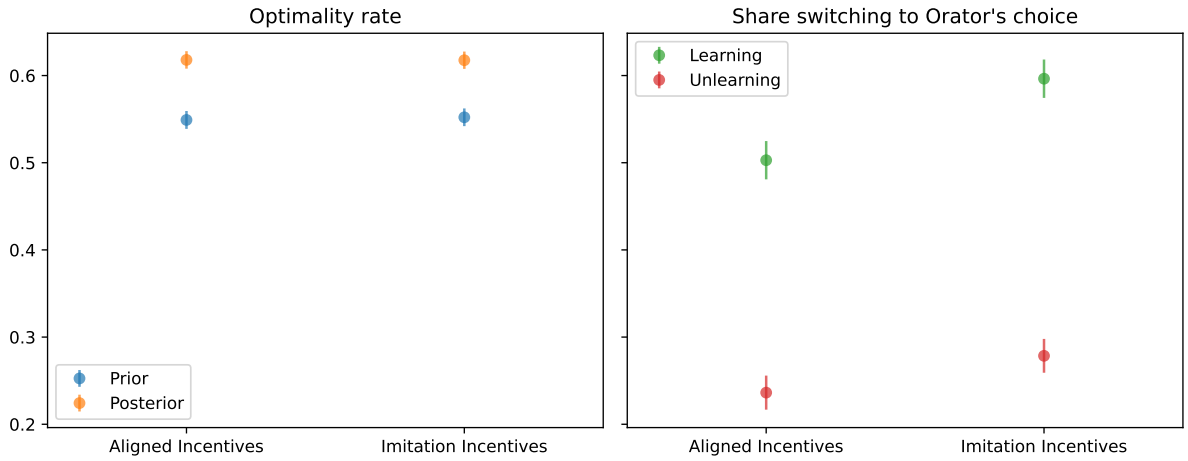


Figure 10: Effect of strategic incentives. *Notes:* Left panel is equivalent to Figure 3 and right panel to Figure 4, now with both *Aligned Incentives* and *Imitation Incentives*. Sample in the left panel is the Receiver experiment (435 Receivers) subset to Orators with *Aligned Incentives* (100 Orators) and *Imitation Incentives* (100 Orators), resulting in 5933 observations; right panel is additionally subset to learning and unlearning situations, resulting in 2535 observations. Whiskers show standard errors.

7.3 How Do Strategic Incentives Shape the Supply of Explanations?

To shed light on how exactly imitation incentives change the nature of explanations, we examine the prevalence of different features in the transcripts along the lines of Section 6. Figure 11 displays the frequency of different speech elements in *Imitation Incentives* and

benchmarks them against the *Aligned Incentives* condition. We show categories of speech features created through a principal components analysis as described above (Section 6.3). We find that incentives for imitation decrease the likelihood of giving short explanations, increase the presence of expressions indicating confidence, decrease complexity, increase references to external authorities, increase disfluencies, and decrease abstract explanations. This highlights that strategic incentives for imitation substantially distort the nature of explanations.

Result 5. *Receivers exposed to recordings from Orators with imitation rather than aligned incentives have similar accuracy rates. Yet, this average effect shrouds substantial heterogeneity: Imitation incentives lead to an increase in the likelihood of being imitated for both learning and unlearning opportunities. This in turn suggests that strategic incentives increase heterogeneity in net learning rates from others' explanations.*

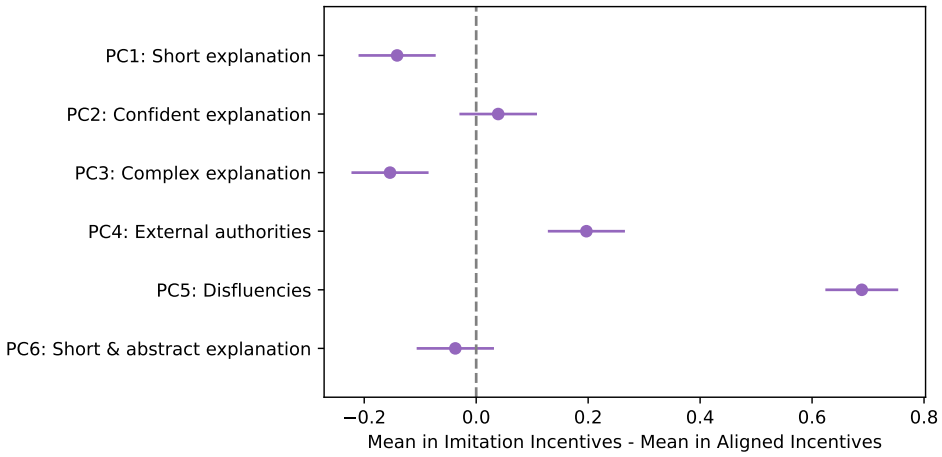


Figure 11: Difference in principal components by incentives. *Notes:* Difference between mean value of principal components in the *Imitation incentives* treatment and in the Aligned incentives treatment. Sample is the Orator experiment (215 Orators), resulting in 3215 observations. Whiskers show 95% confidence intervals.

8 Conclusion

We study how truths and falsehoods in fifteen canonical financial reasoning tasks propagate when people interact on the *marketplace of ideas*. The paper embraces that people typically communicate in spoken language, a central feature of most human learning in practice. Our experimental design allows us to pinpoint the unique effect of verbal explanations in natural language above and beyond the mere observation of another person's action, as explored in most social learning models. Our main finding on the effect of verbal explanations on optimality rates in the aggregate is an optimistic one: when people talk to each other instead of just observing each other's choices, they tend to behave more optimally. However, we document a striking pattern underlying this main effect: when people communicate, truths spread more quickly, but the contagion of falsehoods is not significantly curbed. Using a series of model-guided experiments, we attempt to pin down the mechanisms and conclude that explanations associated with correct answers are structurally different from those associated with incorrect ones. This difference makes it easier to tell that a correct explanation is right than to tell that an incorrect explanation is wrong. We believe that the empirical excercise in this paper provides a blueprint for studying the contagion of beliefs and behavior in an ecological and scalable way.

Limitations. The evidence presented in this paper may be extended in various directions. In many practical contexts, the exposure to others may not be determined by random matching. This implies that learning and unlearning opportunities will not be equally frequent, and hence the sign of the difference between learning and unlearning rates does not serve anymore as a sufficient statistic for whether there is improvement. Morever, in many situations, people not only listen to but also see see each other. This broadens the scope of potential cues that can be used to infer the accuracy of the other's choice. Furthermore, interactions are often repeated rather than one-shot, both in the dyadic back-and-forth within a conversation, and across different contexts. All of these considerations are specifically associated with interactions that occur with people that are not strangers, unlike in our experiments. This suggests that a productive extension of our work is to study the contagion of truths and falsehoods in real social networks.

References

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," 2023.

Andre, Peter, Ingar Haaland, Christopher Roth, and Johannes Wohlfart, "Narratives about the Macroeconomy," 2022.

Atkinson, Adele and Flore-Anne Messy, "Measuring financial literacy: Results of the OECD/International Network on Financial Education (INFE) pilot study," 2012.

Banerjee, Abhijit V, "A simple model of herd behavior," *The quarterly journal of economics*, 1992, *107* (3), 797–817.

Barron, Kai and Tilman Fries, "Narrative persuasion," Technical Report, WZB Discussion Paper 2023.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of political Economy*, 1992, *100* (5), 992–1026.

Brown, Jeffrey R, Zoran Ivković, Paul A Smith, and Scott Weisbenner, "Neighbors matter: Causal community effects and stock market participation," *The Journal of Finance*, 2008, *63* (3), 1509–1531.

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.

Cain, Daylian M, George Loewenstein, and Don A Moore, "The dirt on coming clean: Perverse effects of disclosing conflicts of interest," *The Journal of Legal Studies*, 2005, *34* (1), 1–25.

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

Conlon, John J, Malavika Mani, Gautam Rao, Matthew W Ridley, and Frank Schilbach, "Learning in the Household," Technical Report, National Bureau of Economic Research 2021.

_ , _ , _ , _ , and _ , "Not Learning from Others," Technical Report, National Bureau of Economic Research 2022.

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

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

Eyal, Peer, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina, "Data quality of platforms and panels for online behavioral research," *Behavior Research Methods*, 2021, pp. 1–20.

Eyster, Erik and Matthew Rabin, "Extensive imitation is irrational and harmful," *The Quarterly Journal of Economics*, 2014, *129* (4), 1861–1898.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli, "Chatgpt outperforms crowd-workers for text-annotation tasks," *arXiv preprint arXiv:2303.15056*, 2023.

Graeber, Thomas, Christopher Roth, and Florian Zimmermann, "Stories, Statistics, and Memory," 2023.

_ , Shakked Noy, and Christopher Roth, "Lost in Transmission," 2023.

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

Haliassos, Michael, Thomas Jansson, and Yigitcan Karabulut, "Financial literacy externalities," *The Review of Financial Studies*, 2020, *33* (2), 950–989.

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

Hvide, Hans K and Per Östberg, "Social interaction at work," *Journal of Financial Economics*, 2015, *117* (3), 628–652.

Kendall, Chad W and Constantin Charles, "Causal narratives," Technical Report, National Bureau of Economic Research 2022.

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.

List, John A, "Does market experience eliminate market anomalies?," *The Quarterly Journal of Economics*, 2003, *118* (1), 41–71.

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

_ , "Explanatory preferences shape learning and inference," *Trends in cognitive sciences*, 2016, *20* (10), 748–759.

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 No. WP*, 2007, *157*.

Mobius, Markus and Tanya Rosenblat, "Social learning in economics," *Annu. Rev. Econ.*, 2014, *6* (1), 827–847.

_ **, Tuan Phan, and Adam Szeidl**, "Treasure hunt: Social learning in the field," Technical Report, National Bureau of Economic Research 2015.

Pennycook, Gordon and David G Rand, "The psychology of fake news," *Trends in cognitive sciences*, 2021, *25* (5), 388–402.

Russell, Thomas and Richard Thaler, "The relevance of quasi rationality in competitive markets," *The American Economic Review*, 1985, *75* (5), 1071–1082.

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

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

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.

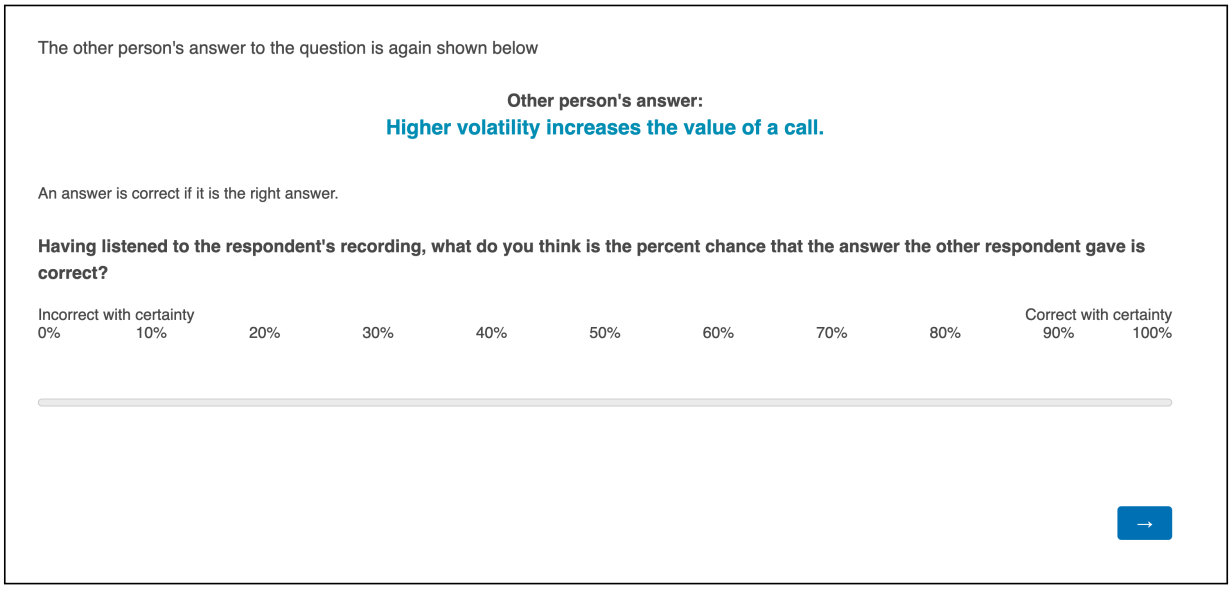
_ , *Narrative economics: How stories go viral and drive major economic events*, Princeton University Press, 2020.

Sonnemann, Ulrich, Colin F Camerer, Craig R Fox, and Thomas Langer, "How psychological framing affects economic market prices in the lab and field," *Proceedings of the National academy of Sciences*, 2013, *110* (29), 11779–11784.

Weizsäcker, Georg, "Do we follow others when we should? A simple test of rational expectations," *American Economic Review*, 2010, *100* (5), 2340–2360.

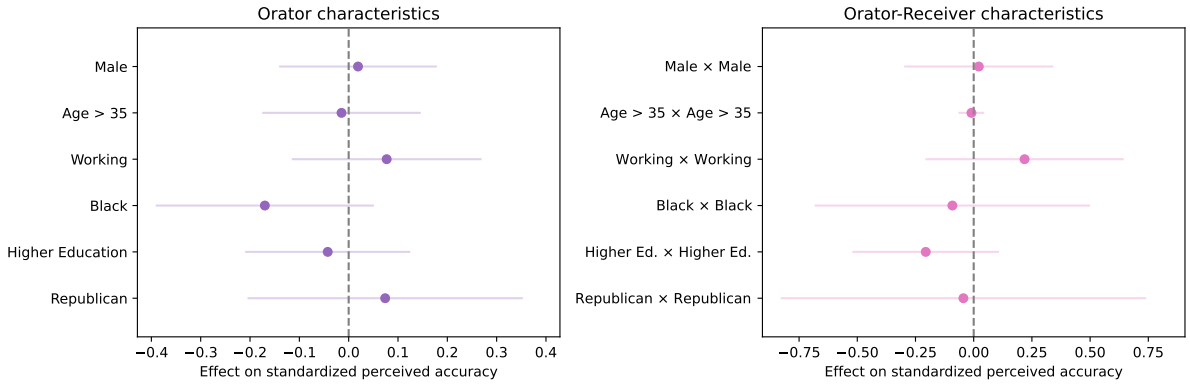
A Additional Figures

A.1 Additional survey illustrations

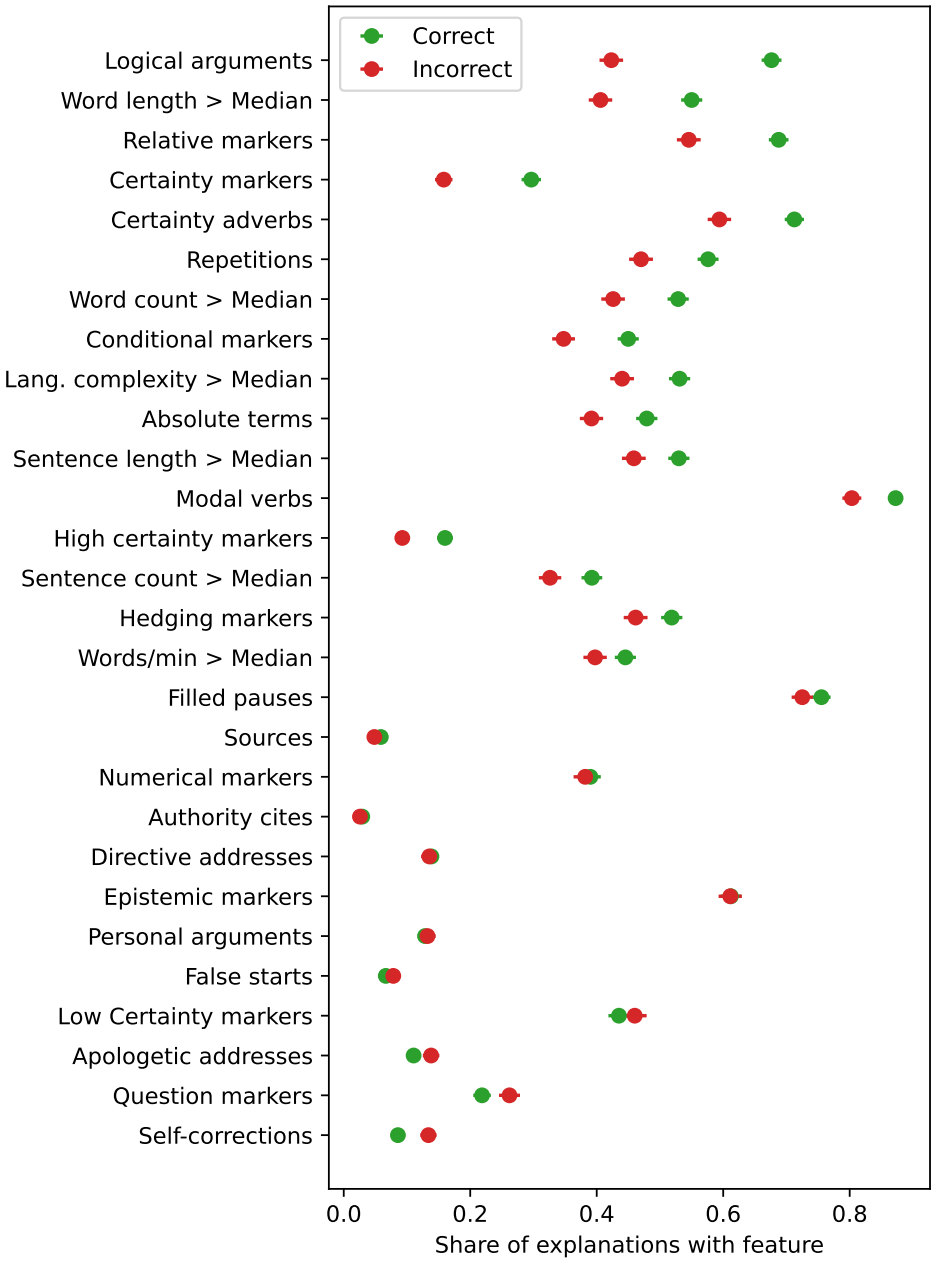


Appendix Figure A1: Elicitation screen for the perception of Orator's accuracy in the Receiver experiment.

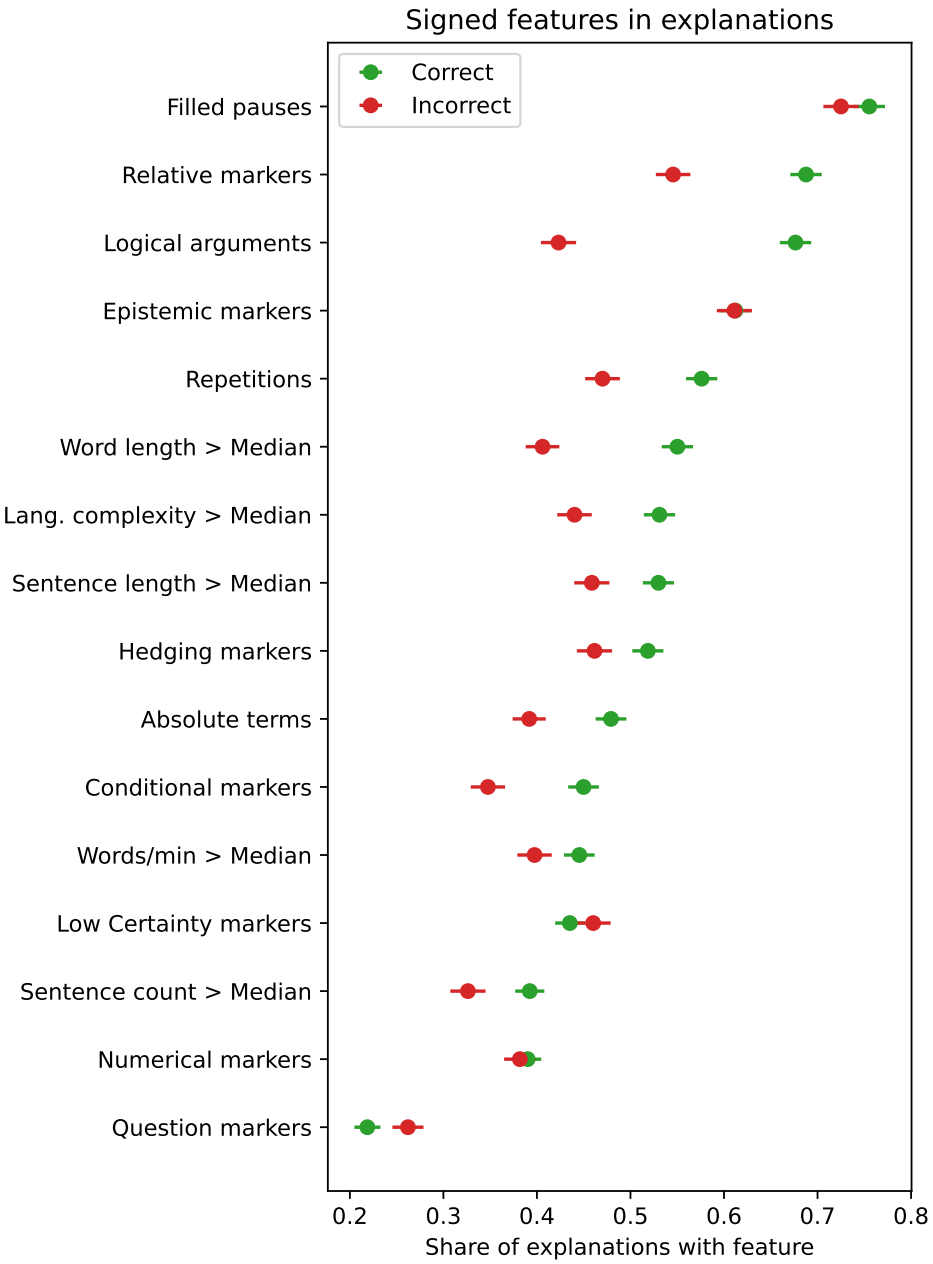
A.2 Additional results



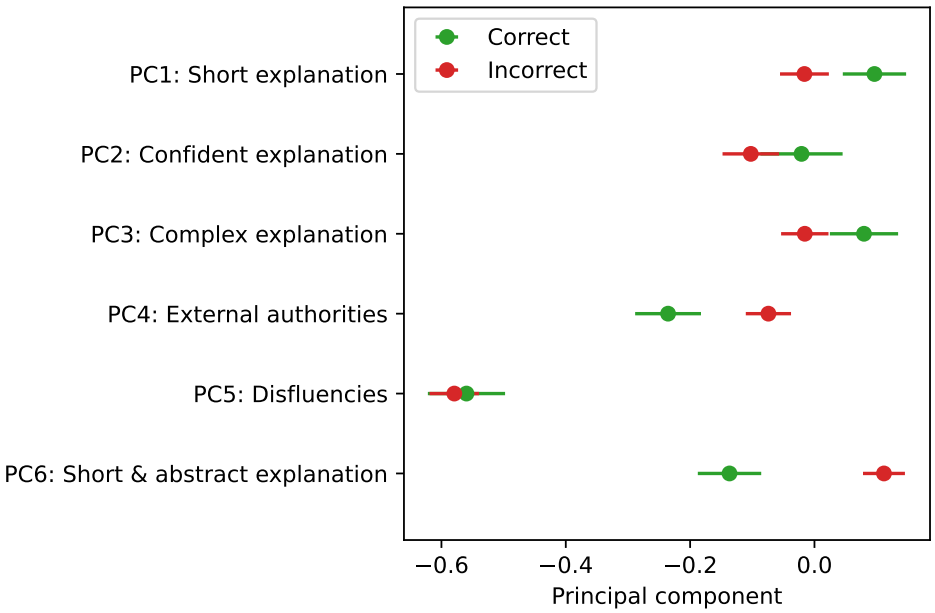
Appendix Figure A2: Effect of Orator and Orator-Receiver characteristics on perceived accuracy in *Choice Only* treatment. *Notes:* This plot reproduces Figure 9 for *Choice Only* as a placebo test: the fact that none of the Orator or Orator-Receiver characteristics have a significant effect shows that selection mechanisms are not at play. Sample is the Receiver experiment (435 Receivers) subset to Orators with *Aligned Incentives* (100 Orators), resulting in 2967 observations. Whiskers show 95% confidence intervals.



Appendix Figure A3: Explanation features by optimality. *Notes:* Share of explanations with each feature, split by correct and incorrect explanations. Sample is the Orator experiment subset to *Aligned Incentives* (106 Orators), resulting in 1589 observations. Whiskers show standard errors.



Appendix Figure A4: Signed explanation features in explanations by optimality. *Notes:* Share of explanations with each feature, split by correct and incorrect explanations. Signed features are markers that provide a way to infer the reliability of an explanation exclusively by indicating whether it is high or low. Sample is the Orator experiment subset to *Aligned Incentives* (106 Orators), resulting in 1589 observations. Whiskers show standard errors.



Appendix Figure A5: Explanation principal components by optimality. *Notes:* Average value of each principal component of explanation feature, split by correct and incorrect explanations. Sample is the Orator experiment subset to *Aligned Incentives* (106 Orators), resulting in 1589 observations. Whiskers show standard errors.

A.3 Orator Experiment

General instructions Thanks for recording your first voice message! This study will take approximately 30 minutes to complete. You will earn a reward of $6.00 for completing the survey.
In the rest of this study, there will be 15 different questions. For each question, you will be asked to record yourself once to give advice on the question and explain your reasoning.
We are interested in how you would give advice in an informal conversation:

- You should share an explanation behind your response.
- Your recording will be played to a few other participants who will have to respond to the same question.
- Other participants can win a bonus for selecting the correct answer.

Importantly:

- You should first read the question, think about your response and then record your answer.
- The recording begins once you click "Start Recording".

We ask you not to search the answers on the internet:

- We are interested in how you form your thoughts about a problem. The type of questions you will be asked makes searching for answers online futile.
- To confirm that you do not search for answers, the survey will monitor whether the survey window remains active.
- If you leave the browser tab of this survey, you will not be eligible for the $6.00 reward.
- You should remain focused on the survey window and answer questions as best you can using your previous knowledge.

After you click to submit a recording, it can take a little while to upload. We kindly ask you to be patient. The upload typically takes no more than 1 minute. To complete the study, you will need to read all instructions carefully and correctly answer the comprehension questions.

Bonus payment

You may receive a bonus payment of $10 for one randomly selected recording. For the bonus, you will be matched to another randomly selected participant who will listen to your recording.

[Description for accuracy treatment]

Whether you receive the bonus payment of $10 depends on how accurate the guess of the other participant is after listening to your voice message. What this means is that if you think you don't know the correct answer, you should make sure that your advice does not induce the other participant to select an incorrect answer. To maximize your chances of receiving the bonus payment, you should give advice in a way that makes the other respondent most likely to select the correct answer!

[Description for imitation treatment]

Whether you receive the bonus payment of $10 only depends on whether the other participant gives the same answer as you. What this means is that for your bonus payment it does not matter whether you know the right answer. It only matters that you make the other participant follow your answer, whether it's right or wrong. To maximize your chances of receiving the bonus payment, you should give advice that leads the other participant to select the same answer as you!

[Comprehension questions]

PAGEBREAK

[Accuracy treatment]

Remember!

Your chances of receiving the bonus payment are highest if the other participant chooses the correct answer.

[Imitation treatment]

Remember!

Your chances of receiving the bonus payment are highest if the other participant chooses the same answer as you.

PAGEBREAK

[Only for imitation treatment]

Creating a Recording that Convinces the Other Participant To Follow You: A Hands-On Tutorial

Here's a step-by-step guide to create a recording that maximizes the likelihood of the advice being followed:

(i) Explicitly state that you know the right answer for sure: Make it very clear that you know the right answer to the question.

(ii) Explain why you know the right answer: Make it clear how you know that the answer is right, e.g. by mentioning that you just read about this in the news or that a friend told you about this, for example.

(iii) Speak confidently: Make sure that you have a confident tone in your voice when speaking.

(iv) Use phrases expressing certainty: Emphasize that you are very confident that you know the right answer.

(v) Avoid phrases expressing uncertainty: Do not use uncertainty words, like probably, maybe, possibly in your recording.

(vi) Plan the Content: Before you start recording, outline the key points you want to make. If you want to, you can take notes to guide what you will say.

(vii) Be Clear and Concise: Communicate your advice clearly. Avoid jargon or tangents, and get straight to the point.

(viii) Quality Matters: Ensure that the recording is of good quality. Poor audio can distract from the message. Speak clearly, in a pleasant tone, and consider using a good microphone.

Inflation

On the next page, a question will be displayed. You should first read the question, think about your response and then record your answer. The recording begins once you click "Start Recording". After recording your advice, you will select your own answer to the question.

PAGEBREAK

Read the question, then record your advice!

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
ii) Less than today

Record your advice that convinces the other participant to select the same answer as you.

[recording]

PAGEBREAK

[Question text with multiple answer response]

Your answer is considered "optimal" if you selected the correct statement.
How certain are you that your above answer is optimal?

[Slider from 0% (Not at all certain) to 100% (Fully certain)]

PAGEBREAK

Suppose you can now choose how many additional participants your recording on this question will be shared with. Which number of other respondents would you like your recording to be shared with?

[Radio buttons from 1 to 10]

Assume that this decision affects your payment as follows: If you win the bonus of $10 and your advice is shared with additional participants:
$1 is added to your bonus for each additional respondent who selects the correct response
$1 is subtracted from your bonus for each additional respondent who selects an incorrect response

A.4 Receiver experiment

General instructions This study will take approximately 35 minutes to complete. You will earn a reward of $6.00 for completing the survey. To complete the study, you will need to

read all instructions carefully and correctly answer the comprehension questions.

In this study, you will be asked to answer 15 questions on various topics.

- In each round, there are four steps: (1) you provide your best answer to the question, (2) you get information about another participant's answer, (3) you answer the same question yourself again, and (4) you assess the accuracy of the other participant's answer. Your answer in (3) may or may not be different from your answer in (1) based on what you learn in (2).
- For some questions, you will listen to a voice message of another person once.
- For other questions, you will see the answer of another participant to the question.

Importantly, on pages where there is a voice recording, each recording starts playing automatically.

Bonus payment

At the end of the survey, one out of every ten participants is randomly selected to be eligible for an additional bonus of $10.00 for choosing the correct answer to a randomly drawn question-that-counts. For the question-that-counts, it will be randomly selected whether your first answer (step (1)) or your second answer (step (3)) determines the bonus payment.

Inflation

Provide your best answer

Please answer what you think is the correct answer to the question.

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
ii) Less than today

Your answer is correct if you selected the right answer.

How certain are you that your above answer is correct?

[Slider from Not at all certain to Extremely certain]

PAGEBREAK

[Speech treatment]

Now, you will listen to a recording of a voice message from a previous respondent who shares their thoughts on the exact same question that you just answered. The voice message will automatically start playing.

Please listen closely to the recording.

You will be able to proceed to the next page once the recording has finished playing.

[Action treatment]

Now, you will observe the answer from a previous respondent.

Please pay close attention to the other person's answer.

PAGEBREAK

[Speech treatment]

Listen to the other respondent's answer

[Box with question text]

[Listen to recording]

[Action treatment]

Read the other respondent's answer

[Box with question text]

[Read answer of other respondent]

PAGEBREAK

Provide your best answer

Please answer what you think is the correct answer to the question.

Your answer may or may not be different from your previous response, given what you learned about the other participant's answer.

[Question with multiple choice answer]

Your answer is correct if you selected the right answer.

How certain are you that your above answer is correct? [Slider from 0% (Not at all certain) to 100% (Fully certain)]

PAGEBREAK

The other person's answer to the question is again shown below

Other person's answer: [Show other respondent's answer to multiple choice question]

An answer is correct if it is the right answer.

Having listened to the respondent's recording, what do you think is the percent chance that the answer the other respondent gave is correct?

[Slider from Incorrect with certainty to Correct with certainty]

A.5 Financial Questions

Note: Correct answers are denoted with (*).

Inflation: 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
ii) Less than today (*)

Exponential growth bias: 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

Interest rates and bond prices: If the interest rate falls, what should generally happen to bond prices?
i) Rise (*)
ii) Fall
iii) Bond prices are not affected

Stock picking: 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 (*)

Actively managed funds: 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 systematically outperform passively managed ones.
(ii) Actively managed funds do not systematically outperform passively managed ones. (*)

Value of a call option: 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.

Interest rates and stock prices: 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

Bid ask spread: 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 (*)

Historical stock returns: What is the average annual return (in %) of the SP 500 stock market index over the past 20 years? [state beliefs about return in %]
9% (*)

Crypto mining: Since the blockchain is decentralized, most Bitcoin mining is done by many small miners.

i) True
ii) False (*)

Diversification: When an investor spreads his money among different assets, does the risk of losing money:
(i) Increase
(ii) Decrease (*)
(iii) Stay the same

Disposition effect: 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
b) No
c) This does not make a difference (*)

Herding: 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 not affect my decision (*)
iii) Should invest less

Good company heuristic: 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 (*)

Home Bias: 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. (*)

B Appendix: Methodological Details

B.1 Details on annotation of explanations via GPT-4

We use transcripts generated by Phonic using Amazon Transcribe. In light of recent evidence suggesting that large language models can annotate text data in a reliable and reproducible manner (Gilardi et al., 2023), we use OpenAI's GPT-4 to annotate the transcripts. We instruct it to identify all instances of a list of features, returning them as a JSON dictionary of lists. The annotation output of the model can then easily be audited, and indeed appears sensible upon inspection. For reproducibility, the temperature of the model is set to 0. Instances are then counted, and counts are then standardized (intensive margin) or turned into indicators equal to 1 if any instance has been detected (extensive margin).

We extract 22 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. Additionally, we generate 6 textual & speech features via direct computation. Table A1 provides an overview of features. Table A2 provides summary statistics on their occurrence frequencies, means and standard deviations.

B.2 Details on Principal Component Analysis of features

To identify key directions of variation in the features 28 identified above, we implement a PCA. Figure A6 shows that the plot of eigenvalues displays an inflexion around 6 components, leading us to select 6 components for Figure 8. Some of the lower components display an effect on perceived accuracy in all configurations, but virtually none of them display an effect in learning situations compared to unlearning situation, confirming our conclusion from the 6 principal components.

Figure A7 shows each principal component's loading on each feature. Principal compo-

Appendix Table A1: Explanation features annotated via GPT-4

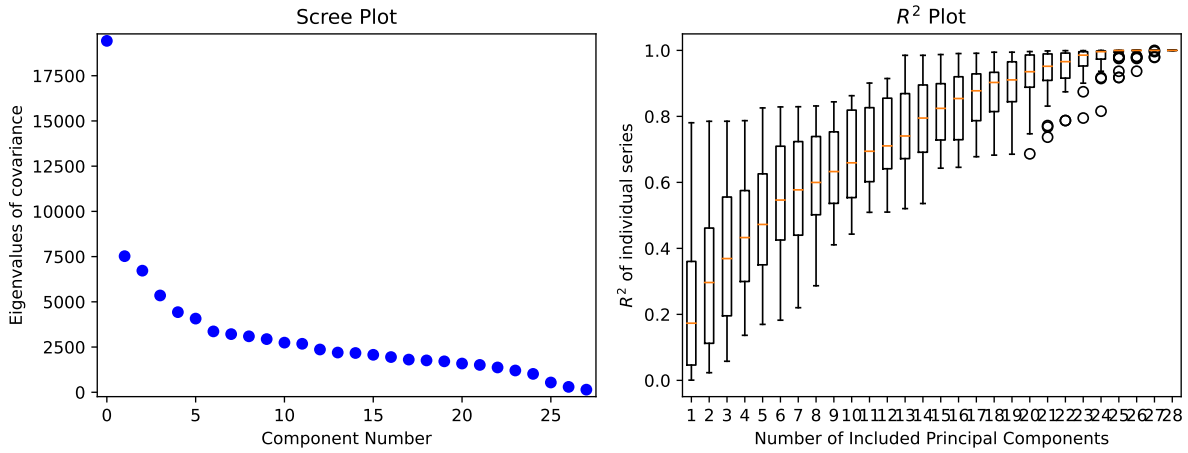
Category	Feature	Prompt
Language markers		
	Modal verbs	Verbs indicating possibility, probability, or necessity. Example: “might”, “could”, “would”
	Certainty adverbs	Adverbs indicating certainty or doubt. Example: “possibly”, “probably”, “likely”
	Hedging language	Phrases indicating hedged claims. Example: “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 Orator 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”
	Logical reasoning args.	Arguments based on logical reasoning. Example: “Since active managers put in more research”
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”
Computed features[a]		
	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 score, flipped so higher values indicate higher complexity.

[a]Text & speech features were obtained via direct computation.

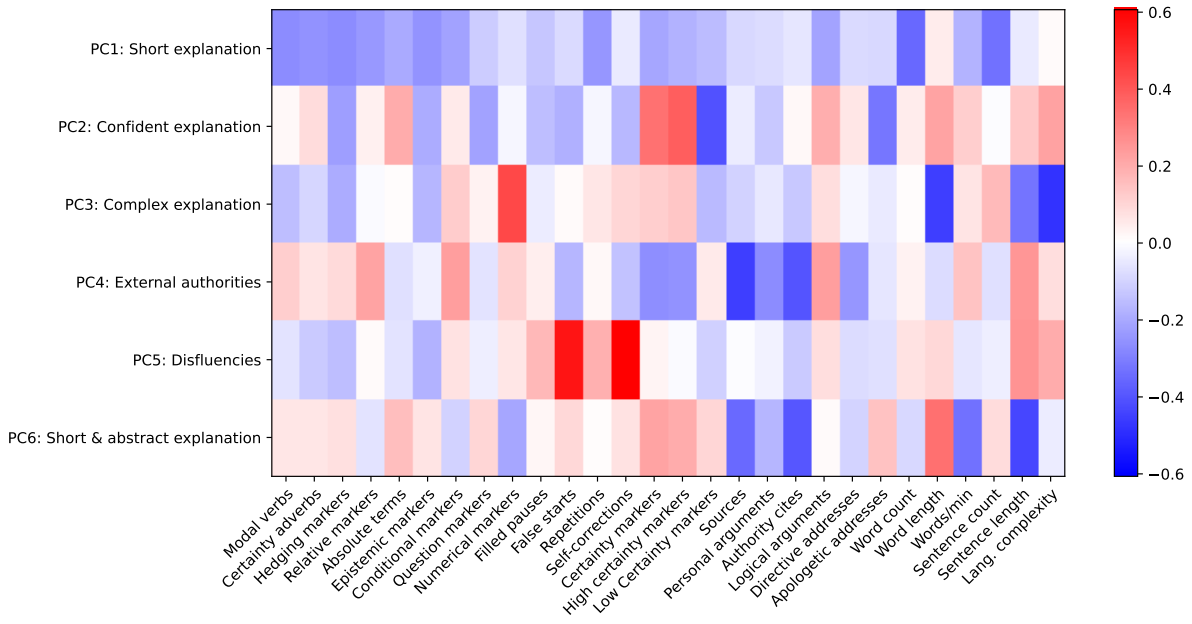
Appendix Table A2: Summary statistics on explanation features

	Frequency	Mean	Std. dev.
Modal verbs	0.84	1.28	0.94
Certainty adverbs	0.66	0.74	0.65
Hedging markers	0.43	0.74	1.08
Relative markers	0.64	0.97	1.02
Absolute terms	0.51	0.60	0.71
Epistemic markers	0.56	0.95	1.11
Conditional markers	0.41	0.60	0.90
Question markers	0.19	0.23	0.56
Numerical markers	0.40	1.10	1.95
Filled pauses	0.67	1.21	1.40
False starts	0.06	0.07	0.27
Repetitions	0.51	0.92	1.14
Self-corrections	0.10	0.11	0.35
Certainty markers	0.33	0.41	0.71
High certainty markers	0.23	0.29	0.66
Low Certainty markers	0.32	0.46	0.82
Sources	0.06	0.08	0.35
Personal arguments	0.13	0.16	0.47
Authority cites	0.04	0.04	0.23
Logical arguments	0.61	0.93	1.16
Directive addresses	0.15	0.21	0.59
Apologetic addresses	0.08	0.09	0.37
Word count		78.01	54.05
Word length		5.39	0.47
Words/min		129.62	34.88
Sentence count		4.38	2.97
Sentence length		18.84	8.07
Lang. complexity		-73.63	15.29

nent labels were obtained using OpenAI's GPT.



Appendix Figure A6: Diagnostics on principal components of explanation features. *Notes:* Left panel shows scree plot of covariance eigenvalues. Right panel shows box plots of individual series' R^2's.



Appendix Figure A7: Loadings of principal components of explanation features.