Explanations*

Thomas Graeber Christopher Roth Constantin Schesch

June 29, 2026

Abstract

When people exchange knowledge, both truths and falsehoods can proliferate. We study the role of explanations for the spread of truths and falsehoods in 15 financial decision tasks. Participants record the reasoning behind each of their answers with incentives for accuracy of their listeners' responses, providing over 6,900 unique verbal explanations. A separate group of participants either only observe one orator's choice or additionally listen to the corresponding explanation before making their own choice. While listening to explanations improves average accuracy, there is substantial heterogeneity: explanations enable the spread of truths, but do not curb the contagion of falsehoods. To study mechanisms, we extract every argument provided in the explanations alongside a large collection of speech features. The largest part of the asymmetry reflects how the reasoning is expressed: explanations for truths are richer—more detailed, structured, and articulated—than those for falsehoods, and receivers imitate richer explanations more regardless of whether the answer is correct. This richness gap accounts for about 60% of the asymmetric benefit. A mechanism experiment shows that varying linguistic richness exogenously increases imitation, even among explanations that independent human annotators judge to contain identical arguments.

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

*We thank the editor and three anonymous referees for excellent suggestions that significantly improved the paper. We are also grateful to Simon Cordes, Pietro Ducco, Maximilian Fell, Paul Grass, Jindi Huang, Milena Jessen, Julian König, Malte Kornemann, Maximilian Müller, Nicolas Röver, Gabriel Saliby, and Georg Schneider for outstanding research assistance. We thank our discussant at the Cognitive Economics Conference, Kirby Nielsen, and seminar audiences at the MPI Bonn, Duke, CESifo behavioral, the MiddExLab, Innsbruck, the University of Pittsburgh, the University of Zurich and Harvard Business School for useful feedback. The research described in this article was approved by the Institutional Review Board at Harvard Business School and the ethics committee of the University of Cologne. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC Starting Grant: 101160770 - VIRAL). Roth: Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/2-390838866. This work was partially supported by the RCN through its CoE Scheme, FAIR project No 262675. The data collections were pre-registered on AsPredicted (#155,323; #157,147; #159,277; #224,365; #234,443; #289,761; #291,528). Graeber: University of Zurich, thomas.graeber@econ.uzh.ch. Roth: University of Cologne, IZA, ECONtribute, and CEPR, roth@wiso.uni-koeln.de. Schesch: Sciences Po, constantin.schesch@sciencespo.fr.

1 Introduction

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

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

We run a series of pre-registered experiments using a two-stage design. We start with an Orator experiment that characterizes the *supply of explanations*. Respondents complete 15 canonical financial decision problems with an objectively correct answer, allowing us to characterize mistakes. These include questions on nominal illusion, the net returns of active and passive investing, the relationship between interest rates and bond prices, compound interest and other topics. In each task, respondents are shown the question and record a voice message with an explanation for their answer to randomly matched participants in a separate study. Orators' incentives induce aligned interests with the listeners: an orator's likelihood of receiving a bonus payment increases with the accuracy of the listener's subsequent response to the question.[1]

To study the *interpretation of explanations* and its consequences for social learning, we then conduct a Receiver experiment with 1,103 respondents, in which respondents face the same

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

15 tasks. They first make their own incentivized choice. For each task, we randomly assign respondents to the *Choice Only* or the *Explanation* condition. In *Choice Only*, they learn about the choice of a randomly chosen respondent in the Orator experiment. In the *Explanation* treatment, receivers additionally listen to the orator's explanation. In both conditions, respondents then again select their best choice, which may now differ from their initial choice, yielding a dataset with more than 30,000 incentivized choices. The *Choice Only* condition provides a natural benchmark that captures learning in the absence of an explanation. The comparison between *Explanation* and *Choice Only* allows us to identify the specific effect of listening to a verbal explanation on imitation, above and beyond the mere observation of another respondent's choice.

Do explanations matter? We begin by analyzing how explanations shape average optimality rates. The prior optimality rate across tasks is 55.3%.[2] In *Choice Only*, just observing someone else's answer increases the frequency of optimal choices to 59.5% ($p < 0.01$). Explanations somewhat increase social learning: across all tasks, the *Explanation* treatment raises the accuracy rate to 62.7% ($p < 0.01$). This modest average treatment effect of 3.2 p.p. ($p < 0.01$), however, conceals that a large fraction of receivers encounter confirmatory advice (58.5%), for whom we document very muted effects, as expected.

We next decompose the average effect of explanations by analyzing different combinations of prior receiver and orator accuracy. First, in "learning situations," a receiver with an incorrect prior receives information from a correct orator: in those cases, receivers imitate the orator's choice in 42.9% of cases in *Choice Only* but substantially more often at 55.8% in *Explanation*. This corresponds to a 12.9 p.p. ($p < 0.01$) increase in seizing learning opportunities. Second, in "unlearning situations," a receiver with a correct prior receives information from an incorrect orator: imitation is 23.1% in *Choice Only* and 22.9% in *Explanation*, an insignificant difference ($p = 0.87$) that means explanations do not make respondents more likely to switch to inaccurate answers. Third, in situations where orator and receiver are both correct, posterior optimality rates are 99.1% and 99.2% ($p = 0.77$). Fourth, when both begin with an incorrect answer, the *Explanation* condition yields a slightly higher optimality rate (3.8%) than *Choice Only* (2.0%; $p < 0.01$). In line with these muted effects in confirming situations, a decomposition shows that 88% of the average treatment effect of explanations is driven by conflicting advice.

Average effects of explanations also mask heterogeneity at the task level. Explanations have a positive effect in 14 out of 15 tasks, though it is only statistically significant for 5 tasks. At the

[2]Because twelve tasks have three options and three have two, the random-choice benchmark is 36.7%.

same time, asymmetric effects of explanations for learning and unlearning situations emerge in 14 out of 15 tasks, though only in a statistically significant manner for 7 of these.[3]

Why do explanations help? To understand the asymmetric effect, we ask what property of explanations could produce it. Any candidate must satisfy two conditions. On the supply side, it must be more common in explanations for correct than for incorrect answers. On the demand side, receivers must respond to it when deciding whether to imitate; otherwise even a large supply gap does not alter social learning. We check each candidate against both conditions in turn.

A natural first candidate is confidence. Correct orators are significantly more confident than incorrect orators, and measured orator confidence modestly predicts imitation in the *Explanation* treatment.[4] Yet, the implied contribution to explaining the asymmetric effect is small: combining the correctness gap in orator confidence with receivers' response to confidence accounts for only a small fraction of the 13.2 p.p. net-learning effect. Receiver confidence is important for heterogeneity—the effect of explanations is strongest for confident but wrong receivers—but its quantitative importance for the asymmetric effect is small.

To test the confidence channel more directly, we conduct an additional pre-registered experiment in which receivers observe the orator's choice and stated confidence rather than the explanation. Relative to *Choice Only*, additionally observing stated confidence does not significantly increase posterior optimality and does not significantly affect imitation in either learning or unlearning situations. Explanations are therefore not reducible to numerical confidence statements: they transmit information beyond the orator's stated certainty.

We next separate the content of explanations from their oral delivery. In a further pre-registered experiment, receivers read a transcript of the orator's explanation rather than listen to a recording. The asymmetric effect persists: transcripts increase imitation in learning situations and leave unlearning essentially unchanged. The transcript treatment reproduces 62% of the effect of spoken explanations. Oral delivery matters, but most of the effect comes from *what* is said rather than *how* it is said. We thus turn to an analysis of explanation content.

[3] Our experiments were not designed with sufficient statistical power for task-level analyses. We cannot reject the null hypothesis that the variation across tasks reflects pure sampling variation ($p = 0.78$).

[4] Receivers never see this confidence directly—they only hear the explanation—so any such relationship reflects confidence that receivers infer from what they hear.

Mechanisms: The content of explanations. We next ask which content features of explanations can account for the asymmetric treatment effect. We begin with the most concrete aspect of content: arguments. We identify and code every *argument*—substantive, domain-specific claims linking question and answer—across the more than 6,900 explanations. This yields, for each of the 15 tasks, the full set of arguments, their frequency in correct versus incorrect explanations, and estimates of their effects on imitation. The argument-level data show that correct and incorrect orators often reason in systematically different ways, and that some arguments are substantially more persuasive than others. This makes arguments a natural candidate mechanism.

At the same time, task-specific arguments are difficult to compare across questions. To obtain a common measure of argument quality, we classify explanations into four categories: (i) no argument, (ii) irrelevant arguments, (iii) fallacious arguments with false premises or invalid reasoning, and (iv) sound arguments. Argument composition differs sharply between learning and unlearning situations. In unlearning situations, respondents more often face no argument (21.0% vs. 13.0% in learning), irrelevant arguments (21.1% vs. 15.1%), or fallacious arguments (53.6% vs. 11.1%). In learning situations, by contrast, they encounter sound arguments far more often (60.8% vs. 4.3%). Thus, the supply side points in the expected direction: explanations for correct answers contain substantially higher-quality arguments.

The demand side is more nuanced. Argument quality matters for imitation, but not in a way that fully explains the asymmetry. Sound arguments, which are far more prevalent in correct explanations, increase imitation, but fallacious arguments, which in turn are more frequent in incorrect answers, also increase it. The two partially offset each other, so that differences in argument categories account for only around one quarter of the asymmetric treatment effect.

This residual asymmetry suggests that receivers are not responding only to the presence or quality of arguments. Explanations for correct and incorrect answers may also differ in their form: how clearly, elaborately, and coherently the orator articulates the reasoning behind a choice. We therefore turn from the classification of arguments to domain-general features of explanations.

Correct explanations are not set apart by particular textual features but score higher on nearly all of them: 36 of 50 domain-general speech and text features occur more often in correct explanations. Because this broad elevation is not captured by any single feature, we summarize them in a pre-registered, LLM-based measure of explanation *richness*, rating how detailed, comprehensive, logically structured, nuanced, and context-tailored each explanation is.

Richness is the property whose supply gap and demand response are both large. On the

supply side, correct explanations are 0.77 SD richer than incorrect ones, a gap that holds within each argument category and even within the same orator answering different questions. On the demand side, richness is by far the strongest predictor of imitation: a one-standard-deviation increase raises imitation by 10.6 p.p. ($p < 0.01$), the effect survives controls for orator confidence and argument content, and it is similar in learning and unlearning situations. The richness gap accounts for about 60% of the asymmetric treatment effect, and orator and receiver characteristics do not account for the asymmetry once richness is taken into account. Because receivers reward richness whether or not the answer is right, and correct answers come with richer explanations, this property spreads truths without equally spreading falsehoods.

Does richness matter when arguments and accuracy are held constant? To test whether richness causally affects imitation, we conduct an additional pre-registered experiment that exogenously varies linguistic richness while holding the underlying explanation fixed. A large language model generates a rich and a sparse version of each transcript, instructed to preserve the same arguments, factual claims, and references while varying the linguistic form. Participants are randomly assigned to one version at the task level.

The richness manipulation substantially increases imitation. In learning situations, imitation rises from 39.7% in the sparse version to 58.4% in the rich version. In unlearning situations, it rises from 15.0% to 26.3%. Since the manipulation is applied symmetrically to explanations for correct and incorrect answers, induced richness is orthogonal to accuracy; the experiment therefore isolates receivers' response to richness itself.

We further test whether this effect survives when *perceived* argument content is held fixed. In a separate pre-registered annotation experiment, human annotators identify the arguments contained in the rich and sparse versions. For the subset in which annotators identify exactly the same set of arguments in both versions, the effect of richness remains large: the rich version raises imitation by 16.9 p.p. in learning situations and by 8.9 p.p. in unlearning situations. The manipulation also modestly changes perceived argument content, indicating two channels: richness can make arguments appear more numerous or legible, and it can raise imitation even conditional on perceived arguments, with the latter accounting for most of the effect.

Do the findings extend beyond financial reasoning? We examine external validity in a canonical balls-and-urns belief-updating task (Edwards, 1968). In a pre-registered binary version of the task, explanations somewhat improve posterior optimality by 3.7 p.p. ($p < 0.01$), 1.8 p.p.

more than *Choice Only* ($p = 0.09$). The same asymmetry appears: imitation increases strongly in learning but not in unlearning. The richness mechanism also replicates. Correct explanations are 0.54 SD ($p < 0.01$) richer, and richer explanations are more likely to be imitated ($p < 0.01$).

In a continuous balls-and-urns task, by contrast, prior optimality is only around 16%, and the additional effect of explanations relative to *Choice Only* is muted, even though correct explanations remain richer. The key difference is that receivers respond much less to richness in this more difficult continuous task. Social learning in that environment appears to operate through salient numerical modes rather than through fine-grained use of the reasoning supplied by orators. Explanations appear to facilitate social learning when the task is tractable enough for receivers to act on the content and richness of the reasoning they receive.

Literature. Our paper contributes to an emerging literature on learning from qualitative information, e.g., in the form of stories and narratives (e.g., Graeber et al., 2024; Andre et al., 2026; Kendall and Charles, 2026; Hüning et al., 2022; Schwartzstein and Sunderam, 2021; Eliaz and Spiegler, 2020; Bursztyn et al., 2023; Barron and Fries, 2024). Thaler et al. (2026) provide evidence that senders with incentives to directionally persuade are more likely to communicate using language rather than numbers. Graeber et al. (2026) examine how verbal transmission distorts the supply of qualitative economic information and show that information about signal reliability gets lost in transmission more than information about signal values because it fails to come to mind. We differ from existing work in our focus on characterizing the supply and interpretation of explanations for people's choices in canonical financial decision problems.[5]

We further contribute to a literature that studies how social learning among non-experts, labeled "naive advisors" by Schotter (2003), affects the prevalence of biases and misinformation.[6] In a setting that, unlike ours, features incentives for deception, Serra-Garcia and Gneezy (2021) show that individuals fail to detect others' lies when they are shown a video of another respondent paid to invent a news story. A series of papers has examined social learning in the context of motivated beliefs (Oprea and Yuksel, 2022; Thaler, 2025; Grunewald et al., 2024).[7] We differ from this prior literature in attributing social learning not to senders' incentives, but

[5]We also relate to work on explanations (Lombrozo, 2006) and arguments (Sloman et al., 1998).

[6]The literature on social learning is vast (Jackson and Yariv, 2007; Galeotti et al., 2010; Conlon et al., 2025), especially in the context of financial decisions (Ambuehl et al., 2022; Hirshleifer et al., 2023). The present paper builds on an earlier literature on advice-giving (Schotter and Sopher, 2003; Çelen et al., 2010; Schotter, 2023), which focused on pre-structured messages.

[7]We also contribute to a literature on whether individual-level biases matter for aggregate market-level outcomes (e.g., Fehr and Tyran, 2005; Enke et al., 2023; Amelio, 2024).

to the content and delivery of the explanations they supply.

2 Experimental Design

2.1 Overview

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

2.2 Financial Decision Problems

A central feature of our design is that we study many financial decision tasks rather than a single one: the variation across tasks elicits structurally different explanations, which are the raw material for our mechanism analysis, and lets us show that our findings are not specific to any one problem. We select 15 financial decision tasks based on three criteria. First, we aim for a collection that is broadly representative of the reasoning and decision biases studied in the finance literature. This spans behavioral phenomena like exponential growth bias and nominal illusion, but also more specific knowledge about different asset classes and investment decisions, for example expected returns under active versus passive investing. Many of the problems we study are tightly linked to common high-stakes financial decisions. Second, we restrict our attention to questions with an objectively correct answer or ones where a broad consensus exists in the financial economics literature, excluding those that rely on tastes. Moreover, participants in our studies are made aware that a "correct" solution exists. Third, the questions should be reasonably short. Some tasks have a logically correct answer, for example the following question about the concept of inflation, with the correct answer underlined:

> *Imagine that the interest rate on your savings account was 2.5% per year and inflation was 3% per year. After 1 year, how much would you be able to buy with the money in this account?*
>
> 1. *More than today*
> 2. *Exactly the same as today*
> 3. *Less than today*

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

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

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

We embrace that differences across these tasks will likely evoke structurally different explanations. A participant might give a wrong answer because they have not heard of the concept of a call option—in a sense, they may not really know what the question is about—or they fully understand the question but still do not know its answer. This difference captures two separate important features of bias in practice and may be reflected in explanations as we discuss in the following sections. Appendix Table E4 outlines the motivation and origin of the tasks, as well as the exact wording of all questions. Three tasks have two options, while the others all have three. In most analyses, we are only interested in whether the correct option is chosen.[8]

2.3 Part 1: Orator Experiment

The main objective of the Orator experiment is to obtain recordings of people's verbal explanations for each of the tasks.[9] Participants are told we are interested in how they would give advice in an informal conversation. They are informed that they should share an explanation behind their response and that their recording will be played to other participants who will answer the same question. We ask respondents not to search for answers on the internet.[10]

People typically have some time to think about the explanations they give. Correspondingly, rather than forcing respondents to talk immediately upon reading the question, we show them the question first and they decide when to start their recording.[11] An example screen is shown in Online Appendix Figure M25.[12] After recording their explanation, respondents first select their

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

[9]The full set of instructions for all surveys in this paper is reproduced in an Online Supplement.

[10]At the end of the study, we ask participants whether they searched for answers (stressing no penalty for admitting it) and, as pre-registered, exclude the 3.9% who say they did.

[11]Orators hence gave their answers while explaining them. Giving answers without knowing one will have to explain them could plausibly induce less deliberate or less conformist answers, connecting with the literature on accountability (Lerner and Tetlock, 1999). Our results should be read with this in mind.

[12]The Online Appendix, which also contains numerous additional analyses, is available at https://cschesch.github.io/documents/Explanations/Explanations%20Online%20Appendix.pdf.

preferred answer and then state their confidence in its accuracy by answering "*How certain are you that your above answer is correct?*" on a scale from 0 "*Not at all certain*" to 100 "*Fully certain*".

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

Speech recordings. The Orator experiment relies on speech recordings, which offer several advantages over written text. A literature outside economics has documented systematic differences between written and spoken production (e.g., Akinnaso, 1982): written text tends to be more formal, structured, and cognitively taxing (e.g., Bourdin and Fayol, 2002), and, as speaking is less exhausting than writing, transcripts of oral explanations are often substantially longer. Speech is also the more natural medium of social learning, making recordings an ideal testing ground to study explanations. Moreover, speech captures features plausibly relevant for social learning but largely absent from written text: tone, emphasis, and disfluencies such as pauses, repetitions, revisions, hesitations, and filler words. Finally, writing imposes an additional filter that may distort explanations relative to those people give spontaneously in the real world.

2.4 Part 2: Receiver Experiment

To characterize the effect of explanations on social learning, we conduct a Receiver experiment. As in the Orator experiment, respondents complete the 15 decision tasks. To measure imitation rates at the individual level, we use a within design with five steps in each round. First, respondents read the financial decision task and are incentivized to indicate their preferred choice, which provides our measure of their prior belief. Second, they indicate their confidence in the accuracy of their response in the same format as respondents in the Orator experiment. Third, they either only learn about the choice of another randomly selected respondent in the Orator experiment (*Choice Only* condition) or additionally listen to the recording of their explanation

(*Explanation* condition). Fourth, the receiver again has an opportunity to select their preferred choice with incentives for accuracy. Fifth, they indicate their confidence in their posterior answer.

Treatments. In the *Choice Only* treatment, receivers may infer and adjust their belief about the optimal answer from learning what someone else chose, even absent an explanation. This same source of learning is present in the *Explanation* treatment, but the explanation provides an additional source of information. We randomize treatments between participants, at the task level. For each task, 80% of receivers are sampled into the *Explanation* condition, while the remaining 20% are assigned to *Choice Only*.[13] The comparison between *Explanation* and *Choice Only* allows us to identify the specific effect of listening to a recorded explanation on learning and unlearning, above and beyond the mere observation of another respondent's choice. The *Choice Only* condition is critical to control for (i) the effects of prior confidence, (ii) measurement error in priors, and (iii) other confounders, such as experimenter demand effects.

Incentives. Receivers had a 10% chance of being eligible for an additional $10 bonus payment. Whether they receive the bonus is determined by the accuracy of their answer in a randomly selected 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

Both studies were conducted on Prolific in December 2023 with U.S. respondents, who received a $6 base reward; median completion times were 25 minutes (Orator) and 22 minutes (Receiver). The Orator experiment recruited 505 respondents, of whom 466 provided valid responses, and required a working microphone to record voice messages.[14] Integrating speech recordings into surveys yielded 6,910 valid recordings, which we stratified and distributed to the Receiver experiment via AWS. The Receiver experiment recruited 1,236 respondents, of whom 1,103 provided valid responses. Appendix Table E5 gives an overview of all collections and pre-registrations.

3 Explanations and Social Learning

We begin with basic descriptives of respondents' explanations, then examine their effects on imitation and optimality, and close with heterogeneity and robustness.

[13] We oversample *Explanation* to have more power to examine its heterogeneous effects.

[14] As pre-specified, orators are valid if at least one recording was non-blank and they did not report looking up answers online; we likewise drop receivers who self-reported doing so.

3.1 Basic Characteristics of Explanations

Our Orator experiment generated 6,910 audio recordings with a median duration of 26 seconds (10th percentile: 11 seconds; 90th percentile: 55 seconds). Audio quality is high: only 1.1% of recordings are unusable, typically due to microphone problems or premature submission.[15] Transcripts preserve the details and nuances of spoken language, including filler words such as "um" and "eh," with a median length of 54 words.

Human coders classify the transcripts using a simple coding scheme (see Appendix B). A negligible 2.6% are nonsensical, and 13.1% are pure restatements of the question or answer without added content, plausibly reflecting low effort. A further 8.6% are coherent but contain *no substantive arguments*, while the remaining 74.3% contain *substantive arguments*. Separately, 13.7% include an explicit expression of the speaker's confidence. These descriptives suggest a rich and varied database of explanations, which we examine in depth in Section 4.

3.2 Explanations and Optimality

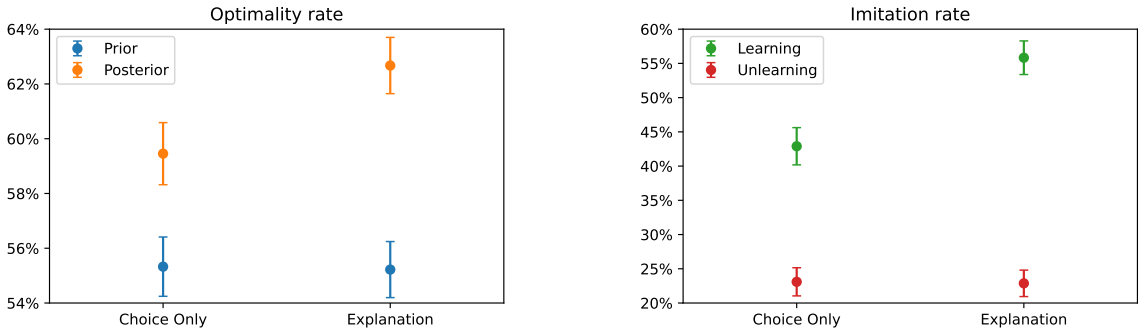
We start by analyzing treatment effects on the frequency of correct choices, which we refer to as the *optimality rate*. To maximize statistical power, we pool observations for *Choice Only* obtained from different between-subject collections in which it serves as the control condition (see Sections 4.1 and 4.2). The prior optimality rate reflects receivers' knowledge about a task before learning from another respondent. The posterior optimality rate captures accuracy after receivers observe another respondent's answer only (*Choice Only*) or additionally listen to their verbal explanation (*Explanation*).

Figure 1a shows these optimality rates pooled across all 15 tasks. Prior to exposure, 55.3% and 55.2% of respondents provided correct answers in the *Choice Only* and *Explanation* conditions, respectively ($p = 0.88$).[16] We document two main findings on posterior optimality. First, just observing another's choice increases optimality rates by 4.1 p.p. ($p < 0.01$), creating an aggregate improvement. Second, additionally listening to another person's explanation raises the size of the improvement to 7.5 p.p. (0.15 SD, $p < 0.01$). The difference in improvement rates between the *Choice Only* and *Explanation* conditions is statistically significant ($p < 0.01$). To interpret the overall effect size, note the limited scope for switching: a large fraction of receivers

[15] Incomprehensible messages and high background noise are very rare and not relevant concerns.

[16] Throughout, all confidence intervals and p-values use clustered standard errors. Receiver-level analyses cluster at the orator and receiver levels, and explanation-level analyses cluster at the orator level.

(58.5%) encounter confirmatory advice, where they have little reason to revise their choice (and overwhelmingly do not, see below). The treatment effect is therefore concentrated among the receivers who face disconfirming advice, implying a substantially larger effect within that group.



(a) Prior and posterior optimality by treatment

(b) Imitation in learning and unlearning

Figure 1: Effect of *Explanation* on optimality and imitation. *Notes:* Left panel shows share of correct receivers before and after exposure to the orator's choice or explanation. Right panel shows share of receivers picking the same answer as the orator in learning and unlearning situations. *Explanation* sample is the main Receiver experiment (1,103 receivers, 13,111 obs.), *Choice Only* is pooled from all collections (2,642 receivers, 8,240 obs.). Whiskers show 95% CIs with clustering at the orator and receiver levels.

3.3 Explanations and Imitation Rates

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

Imitation in response to conflicting signals. Among respondents who receive conflicting signals, we distinguish between two distinct situations: those with initially incorrect choices who are exposed to correct ones, and those with initially correct choices who are exposed to incorrect ones. We refer to the former as *learning opportunities* and the latter as *unlearning opportunities*.

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

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

Second, we do find a quantitatively large treatment effect on the learning rate. Learning opportunities are far more likely to be seized in *Explanation*, where people imitate in 55.8% of cases, than in *Choice Only*, with 42.9%. This 12.9 p.p. ($p < 0.01$) increase in the learning rate shows that, on average, explanations help listeners identify the correct answer.

Imitation in response to confirmatory signals. To benchmark effect sizes in learning and unlearning contexts, we next examine imitation when receivers are exposed to confirming signals. Appendix Figure D3 displays optimality and imitation rates in these situations. When both the orator and the receiver are initially correct, imitation rates are extremely high, at 99.1% in *Choice Only* and 99.2% in *Explanation* ($p = 0.77$). By construction, posterior optimality rates are identical. This indicates that "destructive" arguments, agreeing with a listener's correct choice but nevertheless prompting them to switch to an incorrect one, are essentially absent in our data.

On the other hand, when both the orator and the receiver are initially incorrect, imitation rates stand at 82.6% and 82.2%, respectively ($p = 0.75$). Corresponding aggregate posterior optimality rates are quite small at 2.0% in *Choice Only* and 3.8% in *Explanation*, but do display a small 1.8 p.p. ($p < 0.01$) treatment effect. For this specific analysis, the distinction between two- and three-option tasks plays an important role and provides additional insights. Indeed, Online Appendix C.2 shows that the imitation rate below 100% when both are wrong is driven by three-option tasks in which the orator and receiver chose different wrong options: when the receiver sticks with their answer, this counts as non-imitation. Interestingly, explanations make receivers more likely to select the correct answer in this case. Although these effects are small, arguments can therefore play an "enabling" role: by supporting a wrong choice different from the receiver's, they may still prompt the receiver to think again and switch to the correct answer.[17]

Decomposing the aggregate effect. If matching is random, and the probability of a correct answer is p for receivers and orators, learning and unlearning situations both have population frequency $p(1-p)$. Indeed, in our sample, there are 21.2% learning and 20.3% unlearning

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

opportunities.[18] Let $C = 1$ if the orator is correct, $R = 1$ if the receiver is initially correct, and $Y = 1$ if the receiver is correct after hearing from the orator, with these variables equal to 0 otherwise. Further define E that equals 1 in *Explanation* and 0 under *Choice Only*. The effect of explanations on optimality $\Delta_Y = \mathbb{E}[Y \mid E = 1] - \mathbb{E}[Y \mid E = 0]$ then decomposes as:

$$\underbrace{\Delta_Y}_{3.2\text{pp}} = \underbrace{p(1-p)\Delta_Y^{\text{Learning}}}_{21.2\%\times 12.9\text{pp}=2.7\text{pp}} + \underbrace{p(1-p)\Delta_Y^{\text{Unlearning}}}_{20.3\%\times 1.3\text{pp}=0.3\text{pp}} + \underbrace{(1-p)^2\Delta_Y^{\text{Both Wrong}}}_{23.5\%\times 1.8\text{pp}=0.4\text{pp}} + \underbrace{p^2\Delta_Y^{\text{Both Right}}}_{35.0\%\times 0.1\text{pp}=0.0\text{pp}},$$

where $\Delta_Y^{\text{Learning}} = \mathbb{E}[Y \mid E = 1, R = 0, C = 1] - \mathbb{E}[Y \mid E = 0, R = 0, C = 1]$ etc. are effects on optimality in each situation. This suggests 12% of the treatment effect of explanations is driven by situations where both are wrong, 8% by unlearning and 80% by learning.[19]

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

Robustness. In Online Appendix C, we document that the asymmetric effects of explanations are robust to excluding the shortest and longest recordings, to distinguishing only by orator optimality, and to separating tasks with two or three options. Online Appendix G.6 further analyzes the effects of explanations on changes in receiver confidence between prior and posterior.

3.4 Cross-Task Heterogeneity and Robustness

Appendix Figure D5 shows optimality and imitation rates across tasks. Optimality rates vary from more than 90% for "Nominal Illusion" to less than 20% for "Disposition effect", with the difference in optimality after receiving advice and between treatments being almost an order of magnitude smaller than cross-task heterogeneity. We see this variability as a useful feature of our selection of tasks from the literature, since both harder and easier questions naturally occur in economic decision-making. In turn, imitation rates range from around 60% in learning situations to around 10% in unlearning situations, both for "Nominal Illusion". Tasks with higher optimality rates exhibit a larger gap between learning and unlearning, showing that, in both treatments, easier tasks make it more likely to adjust incorrect beliefs and less likely to adjust correct ones.

[18] These are slightly below the frequency implied by the average optimality $0.553 \times (1-0.553) = 24.7\%$ because $p(1-p)$ is concave in p while we average across tasks with heterogeneous optimality rates.

[19] The decomposition uses the sample frequency of situations. Effects sum to 3.4 p.p. instead of 3.2 p.p. due to finite-sample imbalance of situations across treatments. Percentages use this actual total.

Figure 2 illustrates cross-task variation in the effect of explanations. The left panel shows a positive treatment effect of *Explanation* on optimality in 14 of the 15 tasks. All 95% confidence intervals contain the aggregate effect of 3.2 p.p. However, the treatment effect is statistically significant in only 5 out of 15 tasks, reflecting lower power at the individual task level, since each task only has approximately 1/15th of the overall sample size. In Figure 2, this can be seen in the widths of confidence intervals for specific tasks and the aggregate effect.

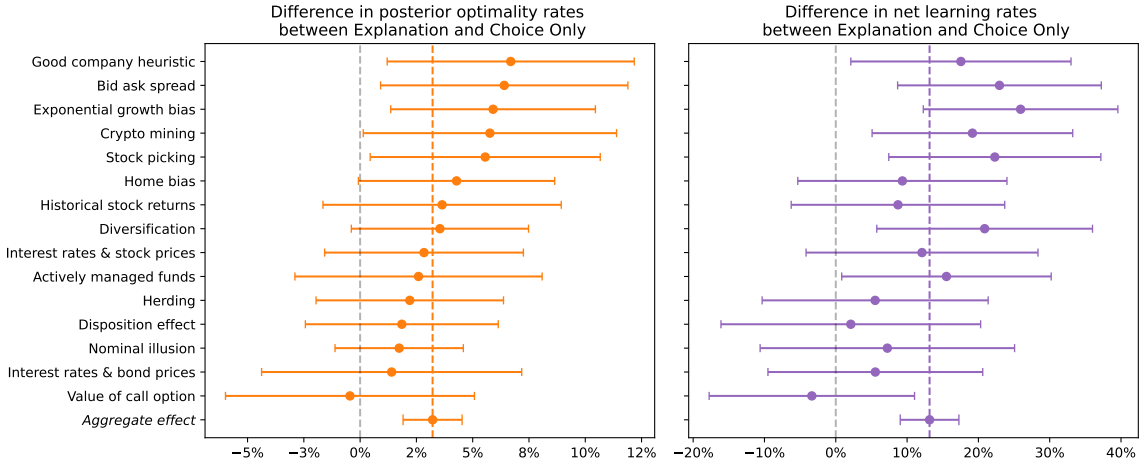


Figure 2: Difference in posterior optimality rates between *Explanation* and *Choice Only* by task, and difference in net learning rates between *Explanation* and *Choice Only* by task. *Notes:* The net learning rate is defined as the difference in imitation rates between learning and unlearning situations. Dashed vertical lines show aggregate effects across tasks, of 3.2 p.p. and 13.2 p.p., respectively. *Explanation* sample is the main Receiver experiment, *Choice Only* is pooled from all collections. Whiskers show 95% CIs with clustering at the orator and receiver levels.

To separate the share of variation reflecting true heterogeneity rather than sampling variation, we perform Cochran's Q-test which yields a statistic of 9.81 with 14 degrees of freedom, failing to reject the null of homogeneity ($p = 0.78$). The corresponding I^2 index is 0%, suggesting all variability should be attributed to sampling error. However, with only 15 tasks, power to detect modest variation is limited. Despite our large data collection of nearly 50,000 individual decisions, the noisy nature of social learning makes precise estimation at the task level difficult: though our experiments were not designed to be highly powered to measure it, we find no evidence to suggest strong heterogeneity in the effects of *Explanation* on optimality.

This conclusion is more nuanced for the asymmetric benefit of explanations. The right panel of Figure 2 shows that the net learning effect of explanations, defined as the difference in learning rates between *Explanation* and *Choice Only* minus the corresponding difference in unlearning

rates, is similarly pervasive. It is positive in 14 tasks, with confidence intervals also containing the 13.2 p.p. aggregate effect in 14 tasks, though it is statistically significant in only 7 tasks. The figure displays substantial heterogeneity, with estimates ranging from a maximum of +25.9 p.p. in "Exponential growth bias" to a median of +12.1 p.p. in "Interest rates and stock prices" to a minimum of -3.4 p.p. in "Value of call option." For net learning, Cochran's Q stands at 18.09 ($p = 0.20$) and the I^2 index is 22.0%, suggesting that roughly one fifth of observed variation likely reflects genuine task-level heterogeneity.

4 Why Do Explanations Spread Truths but Not Falsehoods?

Explanations raise imitation when the orator is right but do not reduce it when the orator is wrong. What property of explanations produces this asymmetry? For any candidate property, two conditions must hold. First, does it differ between explanations for correct and incorrect answers? If not, it cannot generate an asymmetry, however persuasive it is. Second, do receivers respond to it when deciding whether to imitate? If not, even a large supply gap does not alter social learning. A property contributes to the asymmetry only to the extent that both conditions hold: how much it separates truths from falsehoods on the supply side, and how strongly receivers reward it on the demand side. We measure each candidate against both conditions.

We begin with the orator's confidence, the natural candidate in a standard model of social learning, and find that it is too weak on the demand side. We then turn to the explanation itself: first locating the effect in the words rather than the voice, then asking what the words convey: the specific arguments they make, and the richness with which the reasoning is expressed.

A supply-and-demand framework. To compare candidate features on a common scale, we express each one's contribution as a share of the same quantity: the net effect of explanations on imitation of correct versus incorrect orators. Abstracting from confirming advice (cf. Section 3.3), the effect of *Explanation* on optimality is $\Delta_Y = p(1-p)\Delta_I$, with the net-learning effect

$$\Delta_I := \Delta_I^{\text{Learning}} - \Delta_I^{\text{Unlearning}}$$
$$= (\mathbb{E}[I \mid E=1, C=1] - \mathbb{E}[I \mid E=1, C=0]) - (\mathbb{E}[I \mid E=0, C=1] - \mathbb{E}[I \mid E=0, C=0]),$$

where I is an indicator for imitation. Column (1) of Table 1 estimates it from the regression

$$I_{ij} = \alpha + \gamma C_j + \lambda E_{ij} + \Delta_I (E_{ij} \times C_j) + \varepsilon_{ij}, \tag{1}$$

where j indexes orators and i receivers, which puts Δ_I at 13.2 p.p. ($p < 0.01$).

Now consider a feature f of the orator's explanation. Its role depends on two distinct questions: whether it is actually diagnostic of correctness, and whether receivers condition on it. On the supply side, what matters is the gap in its frequency between correct and incorrect,

$$\Delta_f = \mathbb{E}[f \mid C = 1] - \mathbb{E}[f \mid C = 0],$$

and on the demand side, the effect of f on imitation when the explanation is observed, which we denote β_f. The contribution of f to net learning is the product $\beta_f \Delta_f$. We estimate β_f, along with the residual net-learning effect $\Delta_{I,-f}$ that remains after accounting for f, from

$$I_{ij} = \alpha + \gamma C_j + \lambda E_{ij} + \Delta_{I,-f}(E_{ij} \times C_j) + \beta_f(E_{ij} \times f_j) + \varepsilon_{ij}, \tag{2}$$

so that the share of the net-learning effect explained by f is

$$\frac{\beta_f \Delta_f}{\Delta_I} = 1 - \frac{\Delta_{I,-f}}{\Delta_I}. \tag{3}$$

Equation (3) is the common scale we use throughout: for each candidate feature, we measure its supply gap Δ_f and its demand slope β_f and report the share $\beta_f \Delta_f / \Delta_I$ of the asymmetry it accounts for. Because these shares are each measured against the same Δ_I but features may overlap, we read them as the contribution of each feature on its own, not as a partition that sums to one. Appendix A connects the decomposition to a standard Bayesian belief-updating model, in which a more diagnostic feature implies a larger gap Δ_f and, if receivers read the signal correctly, a larger β_f. Because the two coefficients are linked through the unobserved distribution of beliefs and signals, we draw no normative comparison to the Bayesian benchmark; throughout, we compare more versus less inference *across treatments* and ask how much of the treatment effect each feature's coefficients explain.

4.1 Are Explanations Equivalent to Confidence Statements?

In the canonical model of social learning, a receiver observes another's choice and treats it as a signal of unknown reliability; what an explanation could add, on this view, is information about that reliability. This makes the orator's confidence the natural first candidate. Confidence could produce the asymmetry only if both conditions hold: on the supply side, explanations for correct

answers must convey higher confidence than those for incorrect ones; and on the demand side, receivers must be more willing to imitate orators with higher stated confidence. We first assess both conditions using our elicited confidence measures, then isolate the channel experimentally with a treatment in which receivers directly observe the orator's stated confidence.

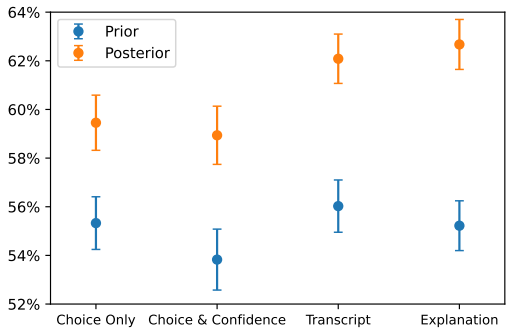
Reduced-form evidence. The supply-side condition for the confidence mechanism is clearly fulfilled. In our experiments, we asked both orators and receivers to report their confidence in their answer on a scale from 0% to 100%. Mean confidence is 75% for correct orators and 63% for incorrect orators ($p < 0.01$), a sizable 0.42 SD difference.[20] Thus, learning explanations are supplied by substantially more confident orators than unlearning explanations.

The demand-side response to confidence is more limited. Column (3) of Table 1 implements regression (2) for orator confidence and shows that a 1 SD increase in orator confidence increases imitation by 5.5 p.p. ($p < 0.01$) in the *Explanation* treatment. Combining this response with the 0.42 SD confidence gap implies an effect of only 0.42×5.5p.p. $= 2.3$ p.p. This falls well short of the total 13.2 p.p. net-learning effect. Hence, while confidence contributes to social learning from explanations, it accounts for only a limited share of the asymmetric treatment effect. This correlational evidence suggests that explanations may not be equivalent to statements of confidence: correct explanations transmit information beyond the orator's stated certainty.[21]

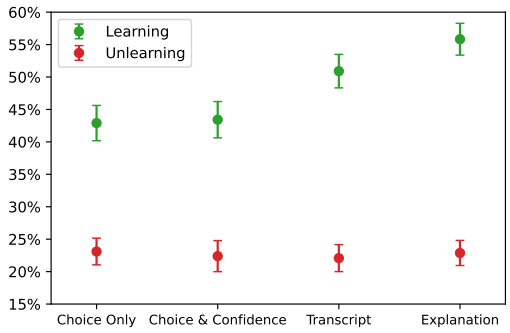
Experimental design. We test this conclusion more directly in a mechanism experiment that isolates the confidence channel. The design closely follows the baseline Receiver experiment and uses the same orator data collected in the baseline Orator experiment. The *Choice Only* condition is identical to the baseline version. The new condition, *Choice & Confidence*, resembles *Choice Only* except that receivers also observe the orator's stated confidence, reported as a number between 0% and 100%. Decision screens from this experiment are shown in Online Appendix Figure M26. In each task, we randomly assign respondents to the *Choice Only* treatment (20%) or the *Choice & Confidence* treatment (80%). This pre-registered experiment was run with 768 U.S. respondents in January 2024, of whom 714 provided valid responses.

[20] Orator confidence reflects both individual differences and task-level effects. More precisely, a random effects decomposition attributes 35.3% of the variation in orator confidence to persistent between-orator differences, with the remainder reflecting within-orator variation across tasks.

[21] Note that receiver prior confidence is relevant for heterogeneous treatment effects but cannot itself be a mechanism behind the main effect of explanations. We therefore report the corresponding analysis in Online Appendix G, which shows this effect is strongest for receivers who are confident but wrong.



(a) Prior and posterior optimality by treatment



(b) Imitation in learning and unlearning

Figure 3: Effect of *Choice & Confidence* and *Transcript* on optimality and imitation rates. *Notes:* Left panel shows share of correct receivers before and after exposure to the orator's choice, confidence & choice, explanation transcript or explanation speech. Right panel shows share of receivers picking the same answer as the orator in learning and unlearning. Samples are the corresponding Receiver experiments for *Explanation* (1,103 receivers, 13,111 obs.), *Choice & Confidence* (714 receivers, 8,531 obs.) and *Transcript* (918 receivers, 10,977 obs.), while *Choice Only* is pooled from all collections (2,642 receivers, 8,240 obs.). Whiskers show 95% CIs with clustering at the orator and receiver levels.

Results. The results are shown in Figure 3. Panel (a) shows that respondents in *Choice & Confidence* improve their optimality rate by 5.1 p.p. ($p < 0.01$). However, adding stated confidence to the orator's choice does not significantly increase posterior optimality relative to *Choice Only*: posterior optimality is 58.9% in *Choice & Confidence* and 59.5% in *Choice Only* ($p = 0.51$). The difference in prior-to-posterior improvements is also not statistically significant ($p = 0.09$). Results for optimality across treatments are summarized in Appendix Table E6.

Panel (b) shows that stated confidence also has essentially no effect on imitation in either learning or unlearning opportunities. In learning opportunities, imitation is 43.4% in *Choice & Confidence* and 42.9% in *Choice Only* ($p = 0.78$). This is far below the 55.8% learning rate in *Explanation*. In unlearning opportunities, imitation is 22.4% in *Choice & Confidence* and 23.1% in *Choice Only* ($p = 0.66$). Appendix Table E7 reports the corresponding regressions.

Result 2. *The effect of explanations on social learning is not reducible to the transmission of stated confidence. Correct orators are more confident, and confidence modestly predicts imitation in the Explanation treatment, but directly revealing the orator's stated confidence does not significantly increase optimality or affect learning and unlearning rates relative to Choice Only.*

This evidence does not rule out that confidence embedded in speech matters in richer and less easily codified ways. It does, however, show that explanations are not equivalent to numerical

confidence statements. The remaining effect must therefore reflect other aspects of the explanation: its content or its delivery. We next separate these two channels by comparing spoken explanations to transcripts.

4.2 What Is Said Versus How It Is Said

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

Experimental design. We transcribe the explanations from our Orator experiment in a way that preserves the nuances of spoken text, including filler words such as "um" and "eh". The design is identical to our baseline Receiver experiment except that *Explanation* is replaced with a *Transcript* treatment, in which participants read the transcript of a recording. In each task, we randomly assign respondents to *Choice Only* (20%) or *Transcript* (80%). To keep the *Explanation* and *Transcript* treatments as comparable as possible, the text is displayed progressively. Decision screens are provided in Online Appendix Figure M27. This pre-registered experiment was run with 1,010 U.S. respondents in January 2024, of whom 918 provided valid responses.

Results. Figure 3a shows that explanation transcripts also strongly increase optimality rates relative to *Choice Only*. *Transcript* induces a posterior optimality rate of 62.1%, significantly above *Choice Only* (59.5%, $p < 0.01$) but not significantly different from *Explanation* (62.7%, $p = 0.31$). Looking at improvements, *Transcript* produces a 2.0 p.p. ($p < 0.01$) larger increase from prior to posterior than *Choice Only*. This corresponds to approximately 58% of the 3.4 p.p. additional improvement in *Explanation*, a significant difference ($p < 0.01$, cf. Table E6).

In turn, Figure 3b shows that *Transcript* also induces a strong asymmetric effect between learning and unlearning. While transcripts have a strong effect of 8.1 p.p. on the learning rate (at 50.9%, $p < 0.01$), the unlearning rate is virtually unaffected relative to *Choice Only*, at 22.0% ($p = 0.46$). The size of the treatment effect of *Transcript* on the learning rate corresponds to 62% of the treatment effect in *Explanation*, also a significant gap ($p = 0.02$, cf. Table E7).

Table E8 replicates the analyses of Table 1 for the *Transcript* treatment. In particular, its

column (1) shows that in this new experiment, imitation is 9.0 p.p. ($p < 0.01$) higher in learning than unlearning situations. The net learning effect of explanations therefore still stands, even when they are only reproduced in the form of a written transcript. Interpreting this coefficient as the effect of explanations after controlling for all features of oral delivery, we conclude that these account for at most $1 - 9.0/13.2 = 31\%$ of the explanation treatment.

This evidence shows that listening to a spoken explanation leads to somewhat more imitation than just reading the same explanation in learning opportunities, while the asymmetric treatment effects emerge in both *Transcript* and *Explanation*. The gap between these treatments highlights that naturalistic oral communication is an important determinant of social learning. At the same time, it suggests that most of the effect of explanations is driven by their content.[22]

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

Does the voice reveal who is speaking? The residual carried by oral delivery raises the question of what it contains. One natural possibility, and the one much of the social-learning literature has emphasized, is the identity of the speaker: hearing a voice lets receivers infer characteristics of the orator, and if they believe those characteristics correlate with correctness—even on biased or unjustified grounds—this could drive imitation on its own (e.g., Cialdini, 2001). The asymmetry would then partly reflect who is speaking rather than what is said.

The evidence does not support this channel. Online Appendix H shows that demographic features of the orator only modestly predict imitation. More tellingly, they predict imitation very similarly in our *Transcript* treatment, where explanations are shown only as text so that any cue to the orator's identity is hidden. To the extent orator characteristics affect imitation at all, they therefore operate through the content of explanations rather than through vocal identification. Column (2) of Table 1 confirms this: controlling for a range of orator characteristics explains only 17% of the learning effect.

Summing up, explanations do not simply transmit the orator's confidence, and the residual carried by their oral delivery is not receivers inferring who is speaking. Our setting thus lets us isolate the role of *content* over the *identity* of the speaker that much of the prior literature has emphasized—and it is content that primarily drives the effect, which we analyze next.

[22]Note that the *Transcript* treatment still relies on text that was originally produced in *spoken* format.

4.3 What is Argued: the Role of Arguments

Having located the primary effect in content rather than confidence or voice, we ask which aspect of content is doing the work. Content has two aspects: the substance of the reasoning, i.e. the arguments an explanation makes, and the manner in which that reasoning is expressed. We begin with substance. If correct and incorrect orators systematically invoke different arguments, and receivers are differentially persuaded by them, argument composition could generate the asymmetry documented above: explanations would help receivers identify truths without equally helping them reject falsehoods.

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

Annotating explanations. The team of research assistants annotates the presence of all arguments in the scheme across all 6,910 explanations. To assess the quality of the main annotation, the manual annotation is then performed again, with each task allocated to a new research assistant who is blind to the previous results. Inter-rater reliability is high. If one coder identifies a specific argument, there is a 72% chance the other coder does so as well. If one coder does not identify an argument, there is a 95% chance the other coder does not either. Cohen's κ is 0.67, indicating "substantial agreement" (Cohen, 1960; Landis and Koch, 1977). Inter-rater reliability is even higher for argument categories. When one coder identifies "any argument", there is a 100% chance that the other coder identifies an argument as well; when one coder does not identify "any argument", there is a 0% chance the other coder identifies an argument. These chances are 82% and 89% for "any fallacious argument" and 83% and 89% for "any sound argument". As a second test, we performed the annotation again using GPT-4. When a human coder identifies an argument, there is a 79% chance GPT-4 does so as well. If a human coder does not identify an argument, there is a 90% chance GPT-4 does not either. Cohen's κ is 0.61,

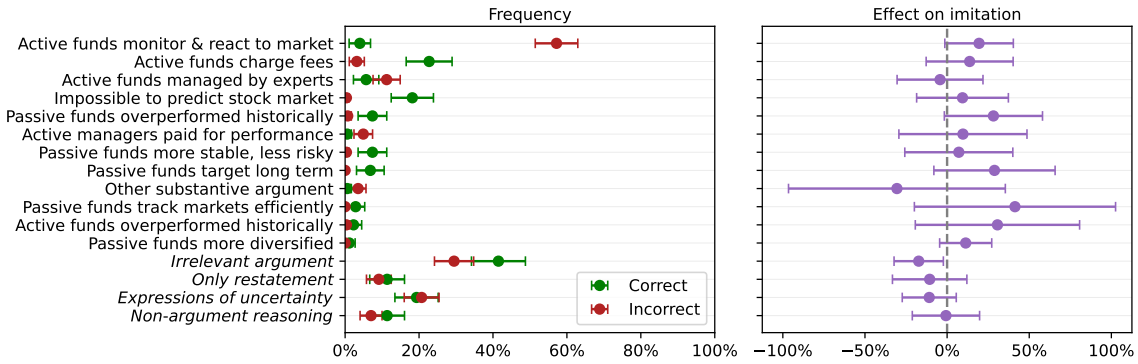


Figure 4: Argument frequencies and effects for the *Actively managed funds* question. *Notes:* Left panel shows frequency in correct and incorrect explanations. Sample is the Orator experiment. Right panel shows effect on imitation, estimated via (2), separately for each argument. *Explanation* sample is the main Receiver experiment, *Choice Only* is pooled from all collections. Whiskers show 95% CIs with clustering at the orator and receiver levels.

again indicating "substantial agreement". Agreement on categories is similarly high.

Illustration: Actively managed funds. To illustrate the argument annotation, consider the results for *Actively managed funds* (cf. Section 2.2), shown in Figure 4. The left panel displays the frequency of a given argument separately for the sample of explanations for correct and incorrect answers, ordered by total frequency in our data. The bottom of the panel also shows the frequency of irrelevant arguments, pure restatements of the answer, any expressions of the speaker's certainty, and any non-argument reasoning. The right panel displays the estimated effect of a given argument on the likelihood of imitation across learning and unlearning situations.

The dominant argument is that active funds can adapt quickly to market changes—common in incorrect explanations (57.2%) but rare in correct ones (4.0%). By contrast, the claim that active funds charge higher fees is frequent in correct explanations (22.7%) and nearly absent in incorrect ones (3.2%). A third theme highlights expert management (5.7% in correct and 11.2% in incorrect), while a fourth stresses market unpredictability (18.2% of correct, absent from incorrect). Irrelevant arguments are frequent, more so in correct (41.5%) than incorrect (29.5%) answers; restatements occur in about 10%; and certainty markers in about 20%. This suggests that the nature and number of arguments vary strongly on the supply side of explanations.

Turning to demand, the right panel shows that adaptability arguments raise imitation by 19.4 p.p., and fee arguments raise imitation by 13.7 p.p. Irrelevant arguments reduce imitation, and so do restatements. Uncertainty has no systematic effects. Individual arguments are noisy,

however, and most effects are imprecisely estimated, so we turn from arguments one at a time to a systematic taxonomy of their quality.

Arguments across tasks. Analogous results for all 15 tasks are displayed in Appendix Figure D6. Explanations vary widely across tasks. Some tasks generate many distinct arguments, others only a few; in some, consensus is elusive, in others a single line of reasoning dominates. Arguments also differ sharply in their persuasive power, sometimes encouraging imitation even when wrong and discouraging it even when correct.

Column (4) summarizes this very simply by counting the number of distinct arguments in each explanation. Overall, correct explanations have 0.32 SD more arguments ($p < 0.01$), but 1 SD more arguments only increases imitation by 6.0 p.p. ($p < 0.01$). This means the *argument quantity gap* accounts for only 11% of the effect of explanations.

A raw count thus reveals a supply-side gap but explains little, because it ignores argument *quality*. These patterns also highlight the difficulty of comparing arguments across tasks. Both considerations motivate a systematic, quality-based taxonomy, which we turn to now.

Categorizing arguments. To draw more general conclusions about differences between explanations for correct and incorrect choices, we define four domain-general argument categories. First, we code the absence of any argument. Second, we define an argument as irrelevant if the premises are unrelated to the question or its answer, i.e., an argument might be entirely off-topic. Third, an argument might be relevant but fallacious: one or more of the premises are false, or the conclusion is not valid given the premises. Finally, we classify a sound argument as one with true premises and a valid conclusion. In classifying explanations, we embrace the fact that there are often several explanations for the correct answer that are sound.[23] Online Appendix Table F1 lists all arguments and their categorization in each task.

Analyzing the prevalence of different argument classes is important as it is *ex ante* unclear whether right or wrong answers are supported by different kinds of arguments. For example, it is in principle possible to give the right answer based on fallacious or irrelevant arguments or to give a wrong answer even if some part of the explanation relies on a weakly sound argument. To classify each explanation, we treat these categories as hierarchical: an explanation will be coded based on the category of the "highest-quality" argument it contains. For example, if an explanation contains both a sound and a fallacious argument, we assign it to the sound bucket.

[23]We also include "weakly sound" arguments in the "sound" category where, from a strict logician's perspective, the premises might not quite be sufficient for the conclusion.

In practice, 91.8% of explanations contain at most one argument type.

The argument quality gap. The left panel of Figure 5 shows the frequency of different classes of arguments encountered in learning versus unlearning opportunities. A large fraction of explanations in unlearning opportunities contain no (21.0%), irrelevant (21.1%) or, most often, fallacious arguments (53.6%). All three types of arguments are significantly less common in learning opportunities, at 13.0% for none, 15.1% for irrelevant and 11.1% for fallacious arguments, respectively. This means that the three categories of "lower-quality" explanations are more frequent in unlearning situations, with the most pronounced gap in the case of fallacious arguments. Sound arguments, by contrast, are practically absent in unlearning explanations (4.3%),[24] yet they constitute the dominant category in learning (60.8%). We refer to this stark imbalance in the distribution of argument types as the *argument quality gap*: learning explanations contain "better" types of arguments than unlearning explanations according to this taxonomy.

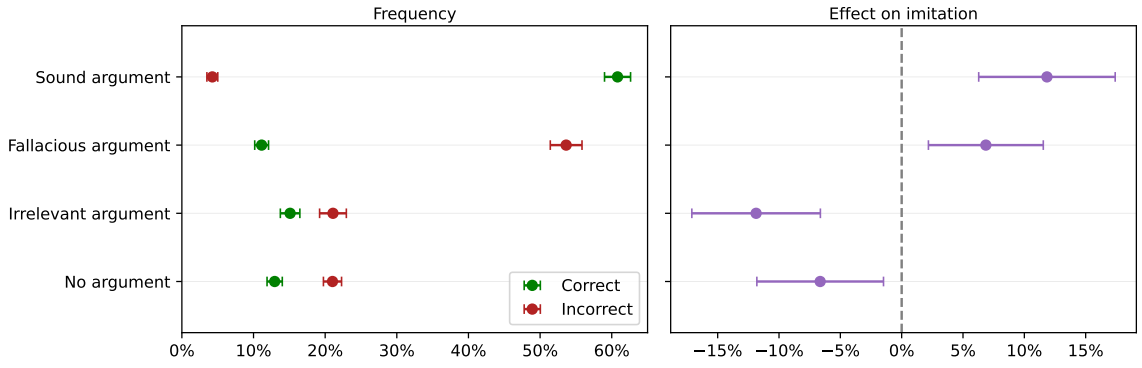


Figure 5: The argument quality gap. *Notes:* Left panel shows frequency of argument qualities in correct and incorrect explanations. Sample is the Orator experiment. Right panel shows effect on imitation, estimated via (2), jointly across argument qualities. *Explanation* sample is the main Receiver experiment, *Choice Only* is pooled from all collections. Whiskers show 95% CIs with clustering at the orator and receiver levels.

Argument quality and imitation. We next examine whether these different argument classes are associated with different imitation rates. The right panel of Figure 5 displays the treatment effects associated with each argument category. This effect is estimated across learning and unlearning situations, as in (2), jointly across argument qualities.

Sound arguments—which appear in more than half of correct explanations but are very rare

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

in incorrect ones—increase imitation by 11.8 p.p. ($p < 0.01$). They therefore very substantially contribute to aggregate learning. On the other hand, fallacious arguments—which analogously appear in more than half of incorrect explanations but are rare in correct ones—also increase imitation by 6.9 p.p. ($p < 0.01$). Fallacious arguments thus play an offsetting role, reducing aggregate learning. Both sound and unsound arguments are therefore imitated more, a heterogeneity which is masked by muted aggregate effects.

In turn, irrelevant arguments decrease imitation by 11.9 p.p. ($p < 0.01$) while the absence of arguments drives it down by 6.6 p.p. ($p = 0.01$). Since the latter two are slightly more frequent in incorrect explanations, they moderately contribute to aggregate learning.

Finally, we combine these supply- and demand-side results to assess how much of the asymmetric effect is driven by argument composition. Column (5) of Table 1 adds a control for each of the argument categories identified above. Comparing it with column (1) shows that differences in argument categories account for up to 25% of the overall asymmetry.

Substance therefore clears both bars but cannot close the gap. Within every argument category, learning explanations are imitated more than unlearning ones. Holding the quality of what is argued fixed, explanations still persuade more when the answer is right. The remaining driver cannot be which arguments are made; it must lie in how the reasoning is expressed.

4.4 How it is Argued: the Richness of Explanations

Two explanations may rest on the same underlying argument yet differ in how that reasoning is expressed—in coherence, specificity, elaboration, qualification, and linguistic structure. Such differences can matter for social learning precisely because receivers do not observe argument validity directly. They must infer how much understanding, effort, and reliability lie behind an orator's answer, and a more developed explanation may be read as a stronger signal of competence or credibility even when the underlying argument is held fixed. The question is whether some property of expression, independent of which arguments are made, separates truths from falsehoods on the supply side and is rewarded by receivers on the demand side.

We first remain agnostic about which property matters, and instead measure a broad set of textual features. We annotate instances of 25 text markers, such as filled pauses or uncertainty markers, with GPT-4 and compute a further 25 text metrics, like word count or readability scores. Appendix Table B1 provides an overview of all features, while Appendix Figure D7 compares their prevalence across correct and incorrect explanations. A striking pattern emerges: 36 of the 50

Table 1: Contribution of explanation features to learning effect

	Dependent variable: Imitation in Learning & Unlearning						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Explanation	-0.002	-0.068	0.012	0.011	0.042**	0.051***	0.074
	(0.013)	(0.042)	(0.014)	(0.014)	(0.018)	(0.015)	(0.047)
Learning	0.198***	0.196***	0.200***	0.202***	0.197***	0.206***	0.197***
	(0.017)	(0.017)	(0.017)	(0.017)	(0.021)	(0.018)	(0.022)
Explanation × Learning	0.132***	0.109***	0.113***	0.117***	0.099***	0.053**	0.049*
	(0.021)	(0.022)	(0.021)	(0.021)	(0.026)	(0.022)	(0.026)
Orator confidence			-0.005				-0.004
			(0.008)				(0.009)
Explanation × Orator confidence			0.055***				0.034***
			(0.011)				(0.012)
Argument count				-0.015*			-0.005
				(0.008)			(0.019)
Explanation × Argument count				0.060***			-0.006
				(0.010)			(0.023)
Richness						-0.011	-0.003
						(0.009)	(0.011)
Explanation × Richness						0.106***	0.079***
						(0.012)	(0.015)
Orator controls		✓					✓
Argument controls					✓		✓
Correct vs incorrect gap			0.42***	0.32***		0.77***	
Gap × Imitation			0.023	0.019		0.082	
Share explained	0%	17%	14%	11%	25%	60%	63%
Observations	8854	8854	8854	8854	8854	8854	8854
R^2	0.092	0.100	0.098	0.097	0.098	0.111	0.117

Notes: Sample is the main Receiver experiment for *Explanation* and all collections for *Choice Only*, both restricted to learning and unlearning situations. All features are standardized so that imitation coefficients β_f are directly comparable. *Correct vs incorrect gap* is the difference in features between learning and unlearning situations Δ_f, expressed in SD. *Gap × Imitation* is $\Delta_f \cdot \beta_f$. *Share explained* is $1 - \Delta_{I,-f}/\Delta_I$, cf. Section 4. *Orator controls* denotes dummies for *Republican*, *Higher education*, *Black*, *Working*, *Age above 35*, *Male* and *Optimality on all other tasks*. *Argument controls* denotes dummies for *No argument*, *Irrelevant argument*, *Fallacious argument* and *Sound argument*. Controls include the variables and interactions with *Explanation*. Standard errors are clustered at the orator and receiver levels.

features occur more often in correct explanations. Correct explanations are not set apart by one or two diagnostic cues but are more developed across the board. This broad pattern points to a single underlying dimension—how detailed, articulated, and internally developed the reasoning is—which we call the *richness* of an explanation. Because any individual feature is only a noisy proxy for this dimension, we summarize it in a single measure of richness and ask whether richer explanations are both more common among correct orators and more frequently imitated.

4.4.1 The Richness Gap

We first describe how we measure richness,[25] then establish the two facts the framework requires: a supply-side gap in richness between correct and incorrect explanations, and a demand-side response of receivers to richness.

Quantifying richness. We annotate richness using GPT-4, instructing it to apply the following pre-registered definition: "A rich explanation is detailed, comprehensive, logically structured, nuanced, and tailors the argument to fit the context. A sparse explanation is basic, narrow, unclear or disorganized, presents only surface-level understanding, lacks depth or specific details and fails to clearly relate to the context." Richness is then scored on a Likert scale ranging from 0 to 10 (both inclusive). Average richness of explanations is 3.2, with a standard deviation of 1.7. Appendix Figure D4 shows the full distribution.

We document a substantial richness gap: correct explanations have 0.77 SD higher richness than incorrect ones ($p < 0.01$). Importantly, this is not only a reflection of the length of transcripts, as the gap is still 0.66 SD after controlling for transcript length ($p < 0.01$). It is also not only a reflection of arguments used, as Appendix Figure D4 shows this pattern holds within each argument category: the gap is 0.30 SD ($p < 0.01$) for none, 0.48 SD ($p < 0.01$) for irrelevant, 0.67 SD ($p < 0.01$) for fallacious, and 0.16 SD ($p = 0.01$) for sound arguments.

What precisely does the richness score capture? In Appendix Figure D8, we investigate which of our 50 textual features are most predictive of an explanation's richness score. All features put together explain richness with a high R^2 of 0.57. Lexical richness—the complexity, variety, and uniqueness of vocabulary, as measured by the entropy of the distribution of words—appears as the most powerful predictor, as on its own it achieves an R^2 of 0.36. Conversely, all features only add little explanatory power once all other features are controlled for. This confirms that richness is a broad synthesis of many text features.

What does variation in richness capture? It is split roughly equally between orators (41.6%) and within orators across tasks (58.4%). A decomposition of the richness gap shows that going from an orator who is never right to one who is always right raises richness by 2.25 SD ($p < 0.01$), while within a given orator a correct answer raises it by 0.57 SD ($p < 0.01$). The latter is only

[25]The term richness is used in different disciplines and economics fields, perhaps most commonly to characterize the space of numerical messages in models of communication. While various definitions of richness are used in the literature, they often relate to the cardinality and/or granularity of the message space as well as the mapping between messages and states. Here, we attempt to characterize the richness of a given message (rather than a hypothetical message space) in natural language.

slightly smaller than the overall gap: richness tracks correctness *within* a person, not merely across abler and less able orators. The gap is therefore not an artifact of who tends to be right.

Richness and imitation. The richness score of an explanation is by far the most potent predictor of imitation. Column (6) of Table 1 shows that, pooling across learning and unlearning situations, a 1 SD increase in richness is associated with a 10.6 p.p. increase in imitation in the *Explanation* treatment ($p < 0.01$). After controlling for orator characteristics, orator confidence, the number of arguments used and the type of arguments used, it is still significant at 7.9 p.p. ($p < 0.01$), as can be seen in column (7). Moreover, it is significant and similar in both learning and unlearning situations at 11.3 p.p. ($p < 0.01$) and 9.4 p.p. ($p < 0.01$). Finally, a decomposition shows a 1 SD increase in orator-level richness predicts 11.3 p.p. ($p < 0.01$) higher imitation, while a 1 SD within-orator increase predicts 10.1 p.p. ($p < 0.01$). Thus, receivers respond almost identically to generally richer orators and to richer-than-usual explanations by the same orator.

Does the richness gap explain the effect of explanations? Learning explanations are richer, and richness strongly predicts imitation; together these account for much of the asymmetric effect. Comparing columns (1) and (6) of Table 1 shows that 60% of the asymmetric effect is explained by the richness gap. Richness remains a powerful determinant of the asymmetric effect once we control for argument categories. Column (7) shows that, after accounting for richness, the explained share of the asymmetric effect only increases to 63% when additionally controlling for argument content, number of arguments, orator confidence and orator characteristics.[26]

Cross-task variation in the richness channel. The framework makes a prediction we can test across tasks. The task-level contribution of richness is the product of three quantities—the frequency of learning and unlearning $p(1-p)$, the richness gap between correct and incorrect explanations Δ_f, and the imitation of richness β_f—so the explanation effect should be largest precisely where these are jointly large. If richness is the operative channel, tasks should line up with this prediction; if not, they should not.

First, Figure 6a plots the frequency of learning and unlearning situations against the optimality rate of each task, which is well fit by the $p(1-p)$ frequency implied by random matching.

Second, Figure 6b reveals a strong positive relation between the richness gap and optimality. Although this paper takes the richness gap as given and hence does not attempt to explain it, this

[26]Online Appendix H.5 explores whether the imitation of richness depends on orator and receiver characteristics, but this does not appear to be the case when using naturally occurring richness variation.

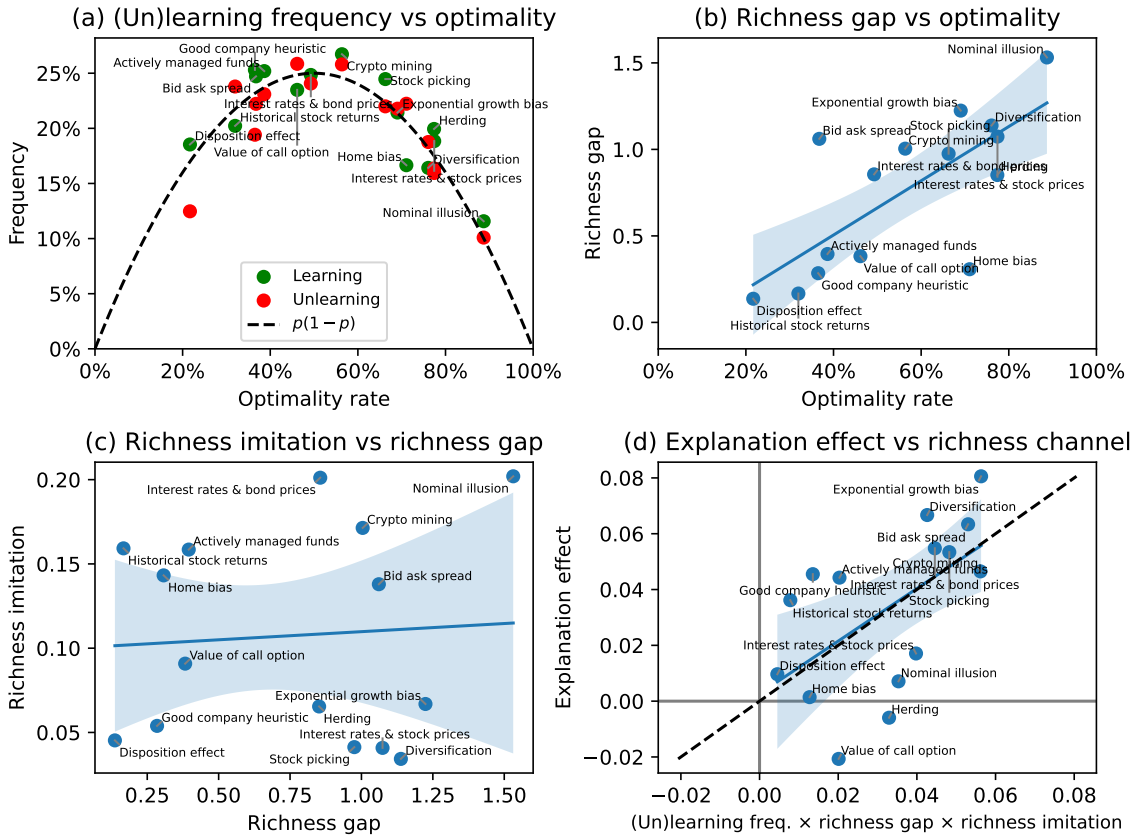


Figure 6: Cross-task variation in the richness channel. *Notes:* Panel (a) plots frequency of learning and unlearning situations against optimality. Panel (b) plots richness gap between correct and incorrect explanations against optimality. Panel (c) plots richness imitation, estimated for each task as in column (6) of Table 1, against the richness gap. Panel (d) plots increase in optimality from prior to posterior in *Explanation* relative to *Choice Only* against the task-level richness channel, defined as the product of the frequency of learning and unlearning, the richness gap, and pooled richness imitation (from column (6) of Table 1). Shaded areas show 95% CIs.

suggests that in easier tasks, more respondents are able to articulate meaningful explanations.

Third, we find that the imitation of richness is robust and relatively homogeneous, as shown in Figure 6c. Furthermore, no association with the richness gap is discernible. To avoid the circularity implied by regressing task-level optimality increases on task-level imitation, we use the pooled effect of richness imitation of 10.6 p.p. from column (6) of Table 1.

Finally, Figure 6d multiplies each of these three components and plots this predicted effect against the observed optimality effect of explanations. This shows a strong positive association, with a slope of 0.94, whose confidence interval excludes zero ($p < 0.01$) but contains 1. Given the level of noise in these cross-task regressions, the R^2 is relatively low at 0.32.

More (un)learning opportunities and a larger richness gap are therefore associated with a stronger effect of explanations, in line with our central mechanism. However, this correlational evidence is only suggestive, and we hence complement it with exogenous variation in richness.

4.4.2 Exogenous Variation in Richness

The descriptive results above leave open the question whether exogenous variation in the richness of an explanation also shapes imitation. To isolate this channel, we conduct an additional pre-registered experiment that varies linguistic richness while holding the argumentative content of explanations constant. This delivers the decisive demand-side test. Because induced richness is orthogonal to accuracy by construction, any effect on imitation shows that receivers reward richness *as such*—even when it carries no information about whether the answer is correct.

In the language of our framework, richness satisfies the supply condition (it is diagnostic of correctness in natural supply) yet receivers condition on it even when we strip it of all diagnostic value. The demand response does not require the cue to be informative; this experiment exactly isolates the wedge between what is diagnostic and what receivers treat as diagnostic.

Design. The experiment shares the Receiver design with the *Transcript* treatment of Section 4.2 and uses the same orator-generated explanations. The main difference is that we randomly assign respondents at the task level to either a *Rich version* or a *Sparse version* of each transcript. Both versions are generated using GPT-4.1, instructed to preserve all arguments, factual statements, and source references while varying linguistic style. The sparse version mimics conversational uncertainty through mid-sentence self-corrections, informal vocabulary, and fragmented structure; the rich version uses precise terminology, explicit logical connectors, illustrative examples, and contextual framing. Full prompts and details are in Online Appendix I.

Results. Figure 7a reports the headline result: the *Rich version* substantially raises imitation in both learning and unlearning situations. Imitation increases from 39.7% to 58.4% in learning ($p < 0.01$) and from 15.0% to 26.3% in unlearning ($p < 0.01$). The effect also appears in confirming cases: it rises from 78.8% to 81.8% when orator and receiver are both incorrect ($p = 0.02$), and from 98.2% to 98.9% when both are correct ($p = 0.05$). That richness raises imitation in unlearning situations is the key point: receivers reward it even when the orator is wrong and richness is therefore actively misleading.

Effect size. To interpret these magnitudes, we rate the richness of each version using an LLM as in Section 4.4. On a 0–10 scale, original explanations averaged 3.2, sparse versions 2.9, and

rich versions 5.9, implying that rich versions are 1.71 SD (of original richness) richer than sparse versions ($p < 0.01$). For comparison, the unconditional richness gap between correct and incorrect original explanations is 0.77 SD. Across learning and unlearning, the rich-version condition raises imitation by 15.4 p.p. ($p < 0.01$). Hence, an experimentally induced 1 SD increase in linguistic richness increases imitation by 9.0 p.p. ($p < 0.01$), about 85% of the 10.6 p.p. increase associated with naturally occurring richness variation in the observational data.

Online Appendix I.8 examines how receiver characteristics correlate with effects of the richness manipulation experiment, documenting that less confident and (to a weaker extent) less financially able receivers are more sensitive to the rich version.

4.4.3 Human-annotated Argument Equivalence

One objection remains: the manipulation was instructed to preserve arguments, but if the rich versions nonetheless introduced stronger argumentative substance, their effect on imitation could reflect better arguments rather than richness as such. We investigate this possibility with an additional pre-registered experiment that measures the arguments humans actually *perceive* in each version.

Design. Participants were shown transcripts from the richness manipulation experiment and asked to identify which arguments were present, using the task-specific lists of arguments from Table F1. Instructions emphasized the substance of arguments rather than wording. Each annotator saw either the rich or the sparse version of an explanation, never both. Participants annotated 10 explanations from a single task; in line with the pre-registration, we discard the first two annotations so annotators could familiarize themselves with the set of arguments. Participants were eligible for a "beauty contest" style bonus based on agreement with other annotators. The final data contain one annotation each for both the rich and sparse variants of 6,883 of the 6,910 explanations, corresponding to 99.6% coverage. Further details are in Online Appendix J.

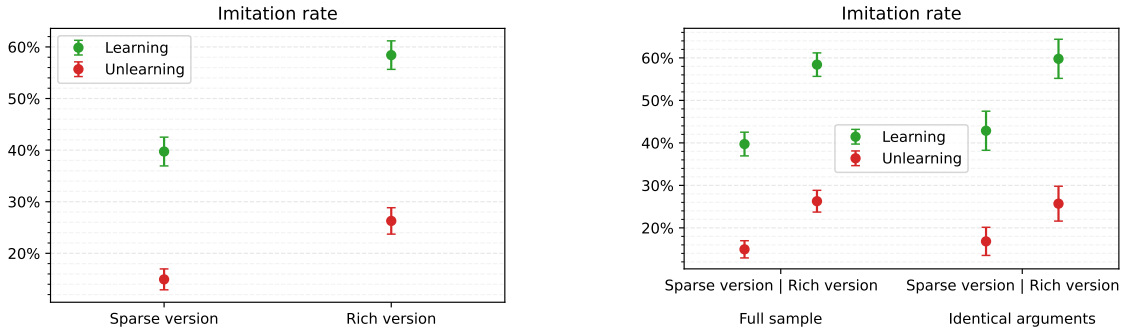
Imitation of richness under argument equivalence. Across these 6,883 explanation pairs, human annotators identify exactly the same set of arguments in the rich and sparse versions for 2,069 pairs, or 30.1% (95% CI: 28.9%–31.3%).

Although this share may seem low, it should be noted that it reflects the fact that human annotation is inherently noisy. Importantly, it is significantly higher than the 7.6% to be expected if arguments were identified at random ($p < 0.01$). Moreover, when comparing two human annotations on the exact same explanation-version, the argument set is judged identical in only

34.7% of cases. If we interpret this same-text agreement rate as the maximum attainable with noisy human coders, this implies that the two versions preserve the same underlying argument set for as much as 86.4% of explanations. In Online Appendix I.5, we perform the same annotation with GPT-4, which appears to have less noise, and therefore identifies exactly the same arguments in 62.6% of version pairs.

We use this exact-agreement subset as our most conservative test of whether richness affects imitation even when perceived argument content is held fixed. Figure 7b compares the richness manipulation in the full sample and this subset.

The effect on imitation remains large and significant in this subset. In learning situations, the rich version raises imitation by 16.9 p.p. ($p < 0.01$). In unlearning situations, it raises imitation by 8.9 p.p. ($p < 0.01$). We find no evidence that the effect differs significantly between the full sample and the identical-argument subset ($p = 0.23$). The strong effect of the richness treatment on imitation is therefore robust to keeping the argument content of explanations constant.[27]



(a) Imitation in richness manipulation

(b) Imitation with identical arguments

Figure 7: Effect of richness manipulation on imitation. *Notes:* Left panel shows imitation rates in learning, unlearning, and confirmatory situations. Right panel restricts the sample to the 30.1% of explanations for which human annotators identified exactly the same set of arguments in the rich and sparse versions. Sample is the Richness receiver survey (972 receivers, 14,580 obs.), with annotations from the Argument annotation survey (1,967 annotators, 13,766 obs.). Whiskers show 95% CIs with clustering at the orator and receiver levels.

Channels of richness imitation. Richness could increase imitation through two channels. First, it could increase the perceived quantity or quality of arguments in an explanation. Second, it could lead receivers to imitate more even holding informational content fixed, perhaps because

[27]An LLM-based check yields the same conclusion: arguments are largely preserved across versions, and restricting the imitation analysis to transcripts with exact LLM-annotated argument overlap leaves the effect unchanged. See Online Appendix I.5.

linguistic richness acts as a signal of the speaker's competence, confidence, or clarity.[28]

We verify that the first mechanism is present by comparing the human annotation with one performed by an LLM. Online Appendix Table J5 shows that, while both identify around one argument on average in the sparse version, humans identify 0.087 more arguments in the rich than in the sparse version relative to GPT-4 ($p < 0.01$). Although this effect is modest, it confirms human perceptions of argument content are themselves sensitive to linguistic richness.

The argument annotation experiment allows us to cautiously shed some light on the relative importance of each channel. Indeed, we can investigate how the effect of the linguistic richness treatment evolves as we add increasingly fine controls for explanations' perceived argument content. Online Appendix Table J6 regresses imitation on the richness treatment across learning and unlearning situations. Adding controls for an identical argument set, for the number of human-identified arguments, or fixed effects for each argument, only modestly reduces the magnitude of the main effect. In this decomposition, variation in linguistic richness after controlling for argument content makes up 87%–92% of the total effect.

This shows that both perceived argument content and linguistic richness conditional on content drive imitation, and suggests the latter accounts for a larger share of the total effect. This decomposition is necessarily imperfect. It is impossible to measure the informational content of an explanation as perceived by each receiver, and human annotations are only a noisy proxy. Subject to these caveats, the annotation experiment corroborates our central conclusion: the form of an explanation, not only its argument content, shapes social learning from it.

Richness is the property whose supply gap and demand response are both large. Correct explanations are reliably richer: within each argument category, across tasks, across orators, and even within the same orator answering different questions. And receivers reward richness for its own sake. The experiment shows they imitate richer explanations more even when richness is made orthogonal to correctness, when perceived arguments are identical, and net of perceived argument content. The asymmetry is the product of these two facts. Truths come with richer explanations, receivers reward richness whether or not the answer is right, and so explanations spread correct answers without equally spreading incorrect ones.

Why do receivers reward richness? Our design identifies that receivers reward richness, but not *why*, which is something we cannot fully settle. At least two non-exclusive accounts are consistent with the evidence. The first is that richness is a shortcut to evaluate accuracy used

[28]We thank an anonymous referee for prompting us to think about these two mechanisms.

most by those least able to evaluate content, given that it is cognitively costly to evaluate the substance of explanations; consistent with this, Online Appendix I.8 shows that less confident, and to a weaker extent less financially able, receivers anchor on richness more strongly when arguments are held fixed. The second account is processing fluency: more articulate, structured, and precise explanations are easier to process, and fluency is known to be misattributed as a signal of truth and competence (e.g., Reber et al., 1998; Alter and Oppenheimer, 2009). We cannot separate these accounts, and read them as complementary descriptions of a reliance on form that substitutes, at least in part, for evaluating content.

Result 4. *Explanations behind correct answers are linguistically richer, accounting for roughly 60% of the asymmetric treatment effect. Experimentally varying richness while holding arguments fixed raises imitation in both learning and unlearning situations.*

5 When Do Explanations Work? An Abstract Application

Do the asymmetric effect and the central role of richness extend beyond financial reasoning, and where does the mechanism break down? We take both questions to a canonical balls-and-urns task (Edwards, 1968): a tractable binary version and a harder continuous version.

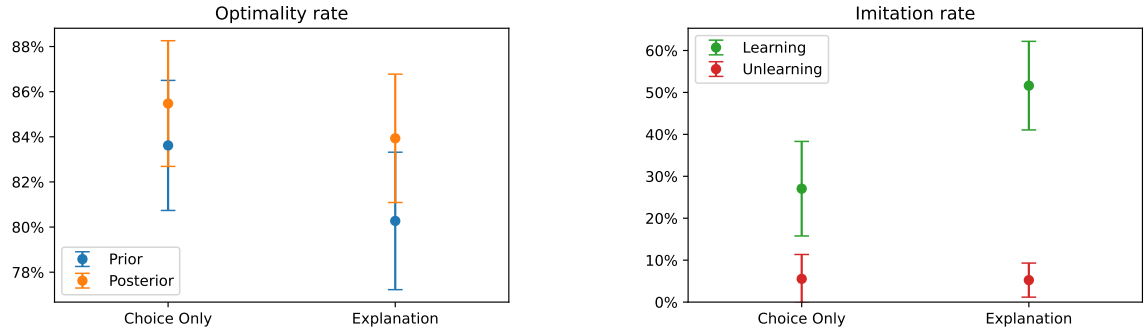
Design and logistics. We use a standard balls-and-urns design. Two bags contain colored balls: one holds D% blue and 100-D% red, the other 100-D% blue and D% red. One bag is randomly selected, with the prior probability of choosing the blue bag set to either 40% or 60%. A single ball is then drawn from the bag and its color is revealed. The participant's task is to state which bag was more likely to be selected. Diagnosticity D is either 55% or 65%, so that the correct answer varies across configurations and ball draws.

The binary elicitation mirrors our main experiment, allowing us to define optimality and imitation in the same way.[29] The remaining structure parallels our main design: the Orator and Receiver experiments, aligned incentives, and the contrast between *Choice Only* and *Explanation*. We collect data from Prolific, and our final sample consists of 483 orators and 1,914 receivers. Appendix K.1 provides additional details.

Results. Four findings emerge. First, prior optimality lies between 80% and 84% across treatments, well above the 50% random-choice benchmark. Second, explanations modestly improve posterior optimality: *Explanation* increases posterior optimality by 3.7 p.p. ($p < 0.01$) relative

[29]We thank an anonymous referee for suggesting this approach.

to the prior, 1.8 p.p. more than *Choice Only* ($p = 0.09$).[30] Third, the average effect is driven by the same asymmetric pattern: imitation in learning opportunities rises from 27.0% to 51.6% ($p < 0.01$), while imitation in unlearning is flat around 5%. Receivers therefore become more responsive to correct orators when they hear an explanation, but no better at identifying incorrect ones. Fourth, the richness mechanism also replicates. Explanations behind correct choices are 0.54 SD richer than those behind incorrect ones ($p < 0.01$), comparable to the 0.77 SD gap we document in the financial tasks. Increasing richness by 1 SD increases imitation by 7.4 p.p. ($p < 0.01$)—slightly less than in the financial tasks—accounting for 22.0% of the learning effect; the smaller share hence reflects a smaller richness gap and a slightly weaker response to richness. Additional details are relegated to Appendix K.



(a) Prior and posterior optimality by treatment

(b) Imitation in learning and unlearning

Figure 8: Effect of *Explanation* on optimality and imitation in binary Balls-and-urns experiment. *Notes:* See notes for Figure 1. Sample is the Balls-and-urns receiver experiment (1,914 receivers and obs.). Whiskers show 95% CIs with clustering at the orator level.

The role of task difficulty. A continuous variant of the task, with posteriors elicited on a 0–100 scale, is far harder: prior optimality is only 16.7%, and treatment effects on imitation and optimality are muted. This boundary is itself explained by the two-sided logic. The supply side does not break down: correct explanations remain richer than incorrect ones, with a 0.75 SD gap ($p < 0.01$) comparable to the other tasks. The demand side, by contrast, does. Receivers respond much less to richness, with a 1 SD increase raising full imitation by only 3.6 p.p. ($p = 0.05$), much lower than in the binary balls-and-urns and financial tasks. Where the task is too hard for receivers to engage with the reasoning they hear, the richness gap on the supply side persists

[30] In both balls-and-urns experiments, respondents only answer one task. Hence orator-level analyses use standard errors that are not clustered, while receiver-level analyses always cluster at the orator level.

but receivers no longer reward it, and the asymmetric benefit disappears. Online Appendix L provides a detailed analysis of this additional experiment.

Two caveats about the interpretation of the differences across the two variants are worth noting. First, across our two variants difficulty co-varies with the dimensionality of the action space, since the tractable task is also the binary one; we therefore cannot fully separate a pure difficulty account from one in which what matters is the match between the dimensionality of messages and of the environment. Second, changing the elicitation also shifts which numerical modes are salient (Bordalo et al., 2025), so difficulty and framing are not fully separable.

6 Discussion and Conclusion

We examine how explanations influence the propagation of truths and falsehoods across 15 financial decision-making problems. In a first experiment, participants record an explanation for each of their answers, incentivized by the accuracy of their listeners' responses. In a second experiment, separate respondents either observe an orator's choice alone or also hear one of more than 6,900 verbal explanations before revising their own. Explanations raise average optimality, but the improvement is driven by the greater spread of truths: falsehoods do not become less contagious. The asymmetry reflects a property of explanations whose supply gap and demand response are both large. Explanations for correct answers are richer—more detailed, structured, and articulated—than for incorrect ones, and receivers imitate richer explanations more regardless of whether the underlying answer is right. Richness differences account for roughly 60% of the asymmetric effect. An experiment that varies linguistic richness exogenously, holding the underlying arguments fixed and verifying argument-equivalence through independent human annotation, shows this response to richness is partly causal. We examine different mechanisms that could lead receivers to reward the richness. Explanations help groups converge on truths largely because, in our setting, that richness happens to accompany correct reasoning. We finally shed light on the boundary conditions of the effectiveness of explanations. When a task is too hard for receivers to engage with the reasoning they hear, they respond to the richness in a more muted way, and the benefit of explanations disappears. Explanations propagate truths only where listeners can act on the reasoning they contain.

Future directions. The evidence in this paper may be extended in various directions. We find that explanation richness is correlated with accuracy, which is a central relationship underlying the overall beneficial effect of explanations. This relationship may be specific to explanations in

settings with aligned incentives. Our setup might serve as a blueprint for studying analogous patterns in the case of *persuasive messages*, where the orator wants the receiver to take a specific action. In the case of persuasion, it is conceivable that the richness-truth association in the supply of arguments weakens, or, in some situations, even reverses. The direction is theoretically ambiguous: a sender who understands that richness raises imitation has an incentive to make misleading explanations richer, while more elaborate explanations may also invite closer scrutiny of the underlying arguments and push senders toward shorter, more concise messages. Which force dominates is an empirical question that our design offers a natural way to study.

Our design could also be extended to separate how much of an explanation's information about correctness is carried by the choice it justifies versus by the justification itself: a treatment presenting the explanation without the accompanying choice, and asking receivers to infer the orator's answer, would reveal which features are informative about that choice, sharpening the distinction between what is diagnostic of correctness and what receivers respond to regardless. Relatedly, future designs could elicit receivers' beliefs about the probability that the orator's answer is correct directly, rather than inferring them from imitation and confidence movements; this would measure perceived diagnosticity feature by feature and sharpen the mapping from explanation characteristics to inferred correctness. Finally, the LLM-based annotation we develop could be applied to high-dimensional language data in other settings to study mechanisms underlying economic decisions. The two-step identification of arguments, the extraction of features and the measurement of richness could all be applied to field settings, such as social media posts explaining financial decisions. This opens up valuable opportunities to validate experimental findings in more naturalistic contexts.

References

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

Alter, Adam L and Daniel M Oppenheimer, "Uniting the tribes of fluency to form a metacognitive nation," *Personality and social psychology review*, 2009, *13* (3), 219–235.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Kendall, Chad W and Constantin Charles, "Causal narratives," *NBER Working Paper No. 30346*, 2026.

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

Lerner, Jennifer S and Philip E Tetlock, "Accounting for the effects of accountability.," *Psychological bulletin*, 1999, *125* (2), 255.

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

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

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

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

Reber, Rolf, Piotr Winkielman, and Norbert Schwarz, "Effects of perceptual fluency on affective judgments," *Psychological science*, 1998, *9* (1), 45–48.

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

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

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

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

Serra-Garcia, Marta and Uri Gneezy, "Mistakes, overconfidence, and the effect of sharing on detecting lies," *American Economic Review*, 2021, *111* (10), 3160–3183.

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

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

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

_ **, Mattie Toma, and Victor Yaneng Wang**, "Numbers Tell, Words Sell," *Working Paper*, 2026.

Appendix for "Explanations"

Thomas Graeber Christopher Roth Constantin Schesch

A Explanations in a belief updating framework

We briefly cast our experimental setup in a standard belief formation model that speaks to the existing economics literature.

Beliefs and imitation in *Choice Only*. We represent the financial reasoning questions as binary choices, and answers as belief probabilities in the correct option $\pi \in [0,1]$. Let $\ell = \text{logit}(\pi) \in \mathbb{R}$ denote an agent's log-odds belief, so that $\ell > 0$ means the agent's current answer is correct and $\ell < 0$ means it is incorrect. Let G denote the population distribution of ℓ, and define

$$G_1(\ell) = G(\ell \mid \ell > 0), \qquad G_0(\ell) = G(\ell \mid \ell < 0),$$

with densities g_1 and g_0. Receivers and orators are drawn from the same population distribution.

In *Choice Only*, the receiver observes only the sign of the orator's belief, $\text{sign}(\ell_j)$. If the optimality rate is $p = \mathbb{P}(\ell_j > 0)$, the diagnosticity of that choice is

$$\delta_c := \log \frac{p}{1-p}$$

The receiver forms the posterior $\ell_i' = \ell_i + \delta_c \cdot \text{sign}(\ell_j)$. In a learning situation, the receiver is initially incorrect ($\ell_i < 0$) and the orator is correct ($\ell_j > 0$), so that the receiver imitates if $\ell_i + \delta_c > 0$. In unlearning situations, the receiver imitates if $\ell_i - \delta_c < 0$.

Imitation in *Explanation*. The *Explanation* treatment additionally reveals a feature f of the orator's reasoning. Let H_1 and H_0 denote its distribution among correct and incorrect explanations. The true diagnosticity of feature f is then

$$\delta(f) = \log \frac{dH_1(f)}{dH_0(f)}.$$

In a learning situation, the receiver then imitates if $\ell_i + \delta_c + f \cdot \delta(f) > 0$, which happens with probability $1 - G_0(-\delta_c - \delta(f))$. Symmetrically, the unlearning probability is $G_1(\delta_c + \delta(f))$.

Define the rational slope β_f as the projection of the Bayesian response on the feature f:

$$\beta_f = \frac{\mathrm{Cov}_{H_1}(f, 1 - G_0(-\delta_c - \delta(f))) + \mathrm{Cov}_{H_0}(f, G_1(\delta_c + \delta(f)))}{\mathrm{Var}_{H_1}(f) + \mathrm{Var}_{H_0}(f)}.$$

This is the model-implied analogue of the reduced-form coefficient β_f when receivers correctly perceive the incremental diagnosticity of feature f.

Feature gap. The feature supply gap Δ_f is connected to the slope β_f. Writing out its definition,

$$\Delta_f = \mathbb{E}_{H_1}[f] - \mathbb{E}_{H_0}[f] = \int f\left(e^{\delta(f)} - 1\right) dH_0(f) = \mathrm{Cov}_{H_0}\left(f, e^{\delta(f)}\right) \approx \mathrm{Cov}_{H_0}(f, \delta(f)),$$

where we used $dH_1(f) = e^{\delta(f)} dH_0(f)$, $\int e^{\delta(f)} dH_0(f) = 1$ and $e^{\delta(f)} \approx 1 + \delta(f)$ for small signal diagnosticities $\delta(f)$. Under the same small-signal approximation,

$$\beta_f \approx \frac{g_0(-\delta_c)\,\mathrm{Cov}_{H_1}(f, \delta(f)) + g_1(\delta_c)\,\mathrm{Cov}_{H_0}(f, \delta(f))}{\mathrm{Var}_{H_1}(f) + \mathrm{Var}_{H_0}(f)} \approx \frac{g_0(-\delta_c) + g_1(\delta_c)}{\mathrm{Var}_{H_1}(f) + \mathrm{Var}_{H_0}(f)} \cdot \Delta_f$$

whenever H_1 and H_0 are not too far apart so that $\mathrm{Cov}_{H_1}(f, \delta(f)) \approx \mathrm{Cov}_{H_0}(f, \delta(f)) \approx \Delta_f$. Thus, both Δ_f and β_f are driven by the covariance between the feature and its true diagnosticity. If a feature f is more diagnostic, the gap between correct and incorrect explanations is larger (higher Δ_f) and a rational receiver imitates it more strongly (higher β_f).

Misperceptions of the signal structure. However, the two might diverge if the diagnosticity is misperceived by the receiver as $\widehat{\delta}(f)$. A simple form of misperception is proportional under- or over-reaction via

$$\widehat{\delta}(f) = \kappa\delta(f),$$

with $\kappa > 1$ for over- and $\kappa < 1$ for under-reaction. Linearizing around the *Choice Only* signal gives $1 - G_0(-\delta_c - x) \approx 1 - G_0(-\delta_c) + g_0(-\delta_c)x$, $G_1(\delta_c + x) \approx G_1(\delta_c) + g_1(\delta_c)x$. Therefore,

$$\begin{aligned}\widehat{\beta}_f &= \frac{\mathrm{Cov}_{H_1}\left(f, 1 - G_0(-\delta_c - \widehat{\delta}(f))\right) + \mathrm{Cov}_{H_0}\left(f, G_1(\delta_c + \widehat{\delta}(f))\right)}{\mathrm{Var}_{H_1}(f) + \mathrm{Var}_{H_0}(f)} \\ &\approx \frac{g_0(-\delta_c)\,\mathrm{Cov}_{H_1}(f, \kappa\delta(f)) + g_1(\delta_c)\,\mathrm{Cov}_{H_0}(f, \kappa\delta(f))}{\mathrm{Var}_{H_1}(f) + \mathrm{Var}_{H_0}(f)} \approx \kappa\beta_f.\end{aligned}$$

Thus, under small incremental signals, the revealed response to a feature is approximately proportional to the fraction of its true diagnosticity perceived by receivers. Under-reaction to

diagnosticity implies a low $\widehat{\beta}_f$ relative to β_f, while over-reaction implies a too high $\widehat{\beta}_f$.

Relation to reduced-form results. This simple model therefore connects the reduced-form coefficients Δ_f and β_f to a standard belief updating framework. It shows that, if agents correctly perceive the signal structure, both should co-move and hence explain the effect of explanations. Conversely, if agents may misperceive the signal structure, the effect can only be explained by a feature that has both (i) a large correctness gap Δ_f and (ii) strong imitation β_f. This motivates our search for a corresponding feature, which concludes that richness is the strongest candidate.

At the same time, the model makes clear that, in general, Δ_f and β_f depend on the full distribution of beliefs, of features and of diagnosticities, so that their exact relationship cannot be pinned down in our naturalistic setting. This forbids strong conclusions about over- vs. under-inference, i.e., a normative comparison to the Bayesian benchmark. Instead, we focus on each feature's ability to explain the observed treatment effect of explanations.

B Annotation of explanations

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

We annotate four different dimensions of explanations. First, we categorize explanations into broad categories, e.g., to distinguish restatements of the answer from non-substantive or substantive argumentation. Second, we identify a large set of 50 features in the explanations, e.g., the word count, the number of uncertainty markers or of analogical arguments. Third, we rate the general richness of explanations using a pre-registered definition. Fourth, we identify the different arguments appearing in each task and tag their presence in each explanation.

B.1 Explanation categorization

We first categorize speeches into general categories to acquire a broad overview of the different types of explanations. For that, we asked a team of RAs to identify whether an explanation fell into one of the following categories: *Only Restatement*, *Any Uncertainty*, *Non-Substantive Explanation*, *Substantive Explanation*, *Correct Explanation*, *Incorrect Explanation*, *Unclear Explanation*, *Invalid Explanation* (see Table B2 for a detailed overview). They are not necessarily mutually

exclusive.

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

B.2 Text features

We identify 50 domain-general text and speech features of explanations, which apply across all explanations independently of the specific question or answer and were largely taken from the vast existing research on text analysis and natural language data.

We first extract 25 markers in five categories: language markers, disfluencies, certainty markers, reasoning content, and addresses to the Receiver. These include explicit uncertainty markers (modal verbs such as could, might; epistemic stance markers such as I believe; hedges such as probably, perhaps), relative and absolute language (almost, always), and certainty references (definitely, certainly). We also capture implicit markers of uncertainty (hesitations, filled pauses, repetitions, self-corrections), as well as mentions of sources, personal experience, authority, direct addresses, and apologetic phrases. Some features potentially overlap: for example, we simultaneously extract high-, low-, and any confidence markers.

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

We generate 25 textual and speech characteristics via direct computation, including word count, delivery speed, and the Flesch–Kincaid readability score. We also include more advanced features, notably based on the distribution of words, part-of-speech tagging, named entity recognition, and syntactic structure identification. Appendix Table B1 provides an overview of all features.

B.3 Richness rating

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

B.4 Argument identification

Section 4.3 describes the argument identification and annotation scheme. It also provides statistics on inter-rater reliability, from a second blind human annotation and from an annotation via GPT-4, all showing substantial agreement. Online Appendix Table F1 shows all arguments appearing in the final scheme. Each has a title used to denote it in Figures D6 and a detailed description used in the annotation. Table B3 further shows the four types of argument we have identified. Each speech is then associated with a specific argument category based on the strongest type of argument it contains.

Appendix Table B1: Text features of explanations

Feature	Description
	Features annotated with GPT-4
Language Markers	
Modal verbs	Verbs indicating possibility, probability, or necessity. Example: “might”, “could”, “would”.
Certainty adverbs	Adverbs indicating certainty or doubt. Example: “possibly”, “probably”, “likely”.
Hedging language	Phrases indicating hedged claims. Example: “it seems”, “appears to be”, “to the best of our knowledge”.
Relative language	Words indicating qualifiers or comparisons. Example: “almost”, “nearly”, “more or less”.
Absolute language	Words indicating absolutes or superlatives. Example: “Always”, “Best”.
Epistemic stance markers	Phrases indicating subjective judgment. Example: “I believe”, “we assume”, “in my opinion”.
Conditional statements	Sentences indicating “If-Then” constructs. Example: “If we don’t act now, then”, “Assuming X, then Y”.
Interrogation markers	Words indicating questions or uncertainty. Example: “who”, “what”, “where”, “when”.
Numerical expressions	Phrases indicating quantitative or probabilistic information. Example: “more than 100 banks”, “95% chance that”.
Disfluencies	
Filled pauses	Instances of filled pauses. Example: “um”, “ah”, “er”.
False starts	Sentences starting but not completed. Example: “If you look at - I believe that”.
Repetitions	Instances of word or phrase repetition. Example: “I I mean”, “this is, this is wrong”.
Repairs	Instances where the speaker corrects themselves. Example: “I have two- three dogs”.
Certainty Markers	
Certainty markers	Statements indicating overall confidence. Example: “Without a doubt”, “I am certain that”.
High certainty markers	Statements indicating high confidence. Example: “I am certain that”, “I am sure that”.
Low certainty markers	Statements indicating low confidence. Example: “It might”, “I’m not sure but”.
Reasoning Content	
Indications of origin	Statements indicating information origin. Example: “According to”, “My grandmother has always said that”.
Personal experience args.	Arguments based on personal experience. Example: “I have often found that”.
External authority args.	Arguments based on external authority. Example: “My girlfriend works at a bank and said”.
Empirical args.	Arguments based on empirical facts. Example: “I remember reading a newspaper article saying”.
Analogical args.	Arguments based on analogies. Example: “Investment funds are like babies”.
Logical reasoning args.	Arguments based on logical reasoning. Example: “Since active managers put in more research”.
Normative args.	Arguments based on ethical considerations. Example: “It would not be fair if”.
Addresses to Receiver	
Directive addresses	Directives to the listener. Example: “You should definitely say that”.
Apologetic or humble addresses	Apologetic or humble addresses. Example: “I apologize for not knowing more”.
	Computed features
Simple Features	
Word count	Total number of words.
Word length	Average length of words.
Words per minute	Average number of words per minute.
Sentence count	Total number of sentences.
Sentence length	Average length of sentences.
Lexical Metrics	
Word Diversity	Ratio of unique words to total number of words.
Word entropy	Entropy of distribution of words.
Hapax Legomena Ratio	Percentage of words that appear only once.
Share of long words	Percentage of words that have more than 10 letters.
Readability	
Flesch-Kincaid Index	Grade required to understand the text, based on sentence length and word length.
Gunning Fog Index	Grade required to understand the text, based on sentence length and percentage of complex words.
SMOG Index	Grade required to understand the text, based on polysyllabic word counts.
Automated Readability Index	Grade required to understand the text, based on characters per word and words per sentence.
Cohesion Metrics	
Referential Cohesion	Mean Jaccard word overlap between adjacent sentences.
Syntactic Complexity Metrics	
Part-of-Speech Entropy	Entropy of parts of speech (e.g., nouns, verbs, adjectives) in the text.
Mean Clause Length	Average number of words per clause, including both standalone and dependent clauses.
Subordination Index	Ratio of subordinate clauses to main clauses.
Clause Density	Average number of clauses per sentence.
Named Entity Recognition	
Entity Density	Ratio of number of named entities (e.g., people, organizations, locations) to total words.
Entity Type Density	Ratio of number of different types of named entities (e.g., person, organization, location) to total words.
Sentence Structure and Syntax Metrics	
Number of Clauses	Total number of clauses.
Syntactic Tree Depth	Maximum depth of the syntactic dependency tree.
Syntactic Tree Mean Branching	Average number of branches per node in the syntactic tree.
Noun Density	Ratio of number of nouns to total number of words.
Verb Density	Ratio of number of verbs to total number of words.

Appendix Table B2: Overview of explanation categories

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

Appendix Table B3: Overview of argument types

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

C Robustness checks

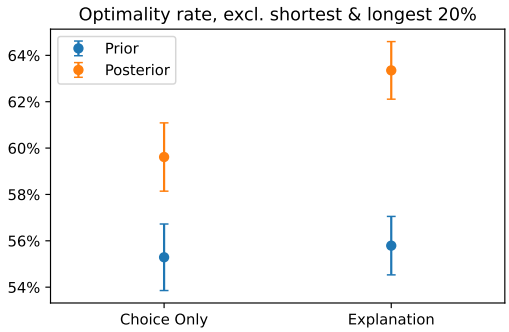
C.1 Excluding the shortest and longest recordings

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

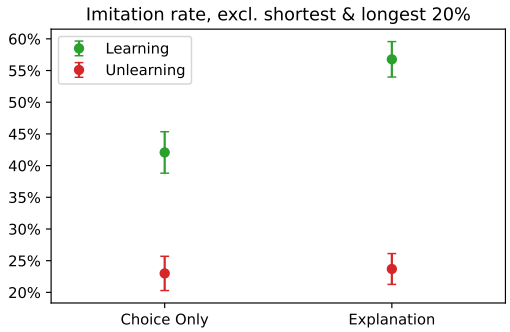
Figure C1b shows the resulting optimality rates, mirroring Figure 1b. The share of receivers giving the correct answer before having been exposed to the orator is 55.3% in *Choice Only* and 55.8% in *Explanation* ($p = 0.59$). Afterwards, the share of receivers giving the correct answer is 59.6% in *Choice Only* and 63.4% in *Explanation* ($p < 0.01$), corresponding to a 3.7 p.p. posterior gap. Netting out the small prior difference, being exposed to an orator's explanation instead of only their choice increases the improvement in optimality by 3.2 p.p. ($p < 0.01$).

Figure C1b repeats the analysis by learning and unlearning situations as in Figure 1b. In unlearning situations, being exposed to an orator's explanation increases the likelihood of imitating their answer by an insignificant 0.7 p.p. ($p = 0.70$) relative to only seeing their choice. On the other hand, in learning situations, being exposed to an orator's explanation in addition to their choice increases the likelihood of taking over their answer by 14.7 p.p. ($p < 0.01$).

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



(a) Prior and posterior optimality



(b) Imitation by situation

Appendix Figure C1: Robustness to dropping shortest and longest recordings. *Notes:* Starting samples are as in Figures 1a and 1b. We additionally drop the 20% shortest and 20% longest explanations. By construction, receiver prior optimality is 0% in learning situations and 100% in unlearning situations. Whiskers show 95% CIs with clustering at the orator and receiver levels.

C.2 Distinguishing by number of options

While twelve of the tasks gathered from the literature have three answer options, three of them are effectively true/false questions and therefore only have two options: "Actively managed funds", "Crypto mining" and "Stock picking" (cf. Table E4).

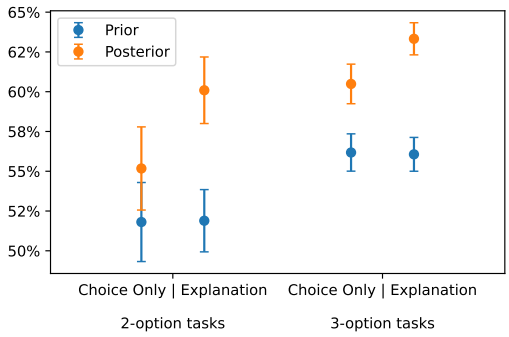
It is natural to expect different optimality rates if some respondents randomize between answers. However, the prior optimality rate is 51.8% for the two-option tasks but 56.2% for the three-option tasks ($p < 0.01$), showing that cross-task heterogeneity dominates. Prior confidence is 63.6% on average for two-option tasks, against 68.8% for three-option tasks ($p < 0.01$). Posterior confidence exhibits a similar gap at 68.9% and 74.0%, respectively ($p < 0.01$).

Although the number of options is constant across conditions and should therefore not influence treatment effects, Figure C2 confirms that our main effects hold for both types of tasks. The left panel shows that *Explanation* increases posterior optimality by 4.8 p.p. relative to *Choice Only* among two-option tasks ($p < 0.01$), and by 3.0 p.p. among three-option tasks ($p < 0.01$). This difference in treatment effects between the two types of tasks is not statistically significant ($p = 0.18$). In the right panel, the effect of *Explanation* over *Choice Only* is not significant in unlearning situations for two- or three-option tasks ($p = 0.24$ and $p = 0.66$). On the other hand, in learning situations it is 15.6 p.p. for two-option tasks ($p < 0.01$) and 12.2 p.p. for three-option tasks ($p < 0.01$). As before, this difference is not statistically significant ($p = 0.33$).

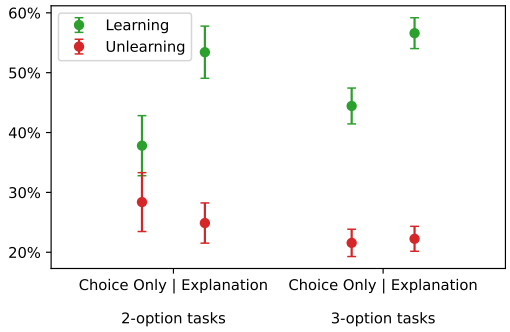
Finally, we verify that richness drives imitation robustly across tasks by repeating the specification of column (6) of Table 1 separately. In two-option tasks a 1 SD increase in richness raises imitation by 15.2 p.p. ($p < 0.01$), while it causes a 10.6 p.p. increase in three-option tasks ($p < 0.01$). This difference is not statistically significant ($p = 0.07$).

The fact that we use two- and three-option tasks introduces an interesting source of variation when analyzing receivers' reaction to confirmatory signals, as discussed in Section 3.3. Importantly, we define imitation as the receiver picking the exact same option as the orator. When the receiver and the orator are both right, they have necessarily chosen the same answer. However, when both are wrong, they will sometimes have chosen two different wrong answers.

When the receiver and orator are both wrong, imitation rates are 82.6% in *Choice Only* and 82.2% in *Explanation* ($p = 0.75$), substantially lower than when both are right. However, in tasks with only two options, imitation rates are 97.5% and 96.7%, respectively ($p = 0.44$). They are similarly high in tasks with three options when the orator chose the same wrong option as the receiver, at 97.7% and 96.3% ($p = 0.02$). On the other hand, imitation is much lower at



(a) Prior and posterior optimality



(b) Imitation rate

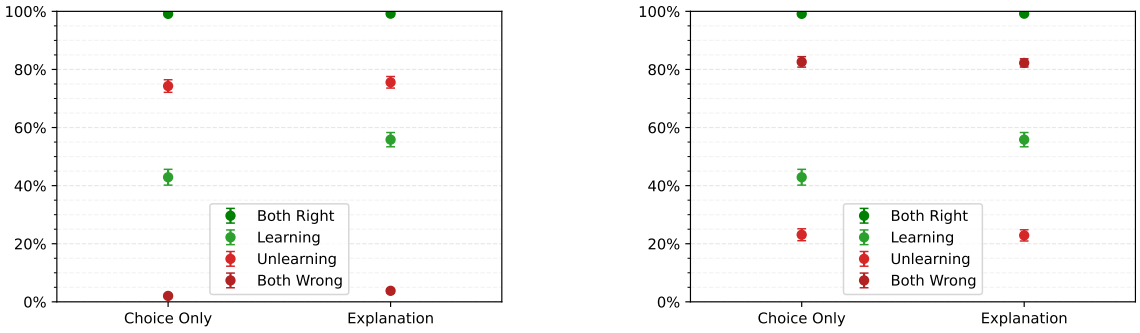
Appendix Figure C2: Effect of *Explanation* by number of options. *Notes:* See notes for Figure 1. Two-option tasks are "Actively managed funds", "Crypto mining" and "Stock picking", see Table E4. Whiskers show 95% CIs with clustering at the orator and receiver levels.

36.0% and 37.8% ($p = 0.54$) in three-option tasks when the orator chose a different wrong option. Interestingly, the imitation rate for an alternative wrong option is between the learning and unlearning rates. Explanations therefore slightly reduce imitation in three-option tasks when receiver and orator initially picked the same wrong answer.

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

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

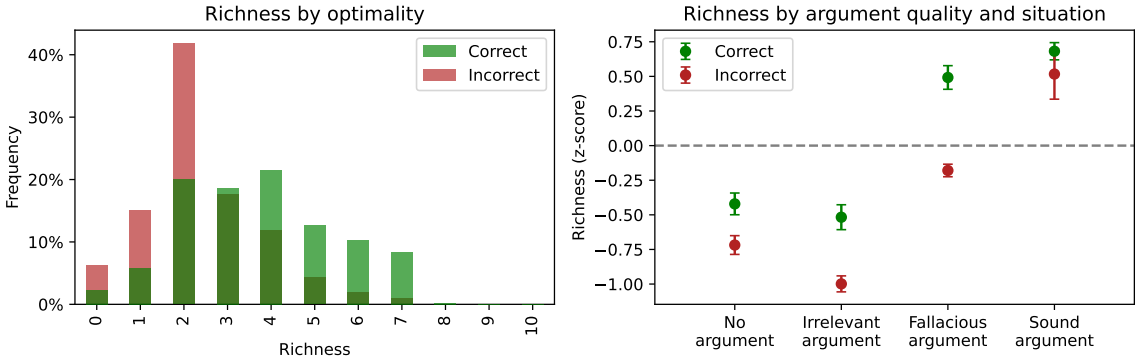
D Additional figures



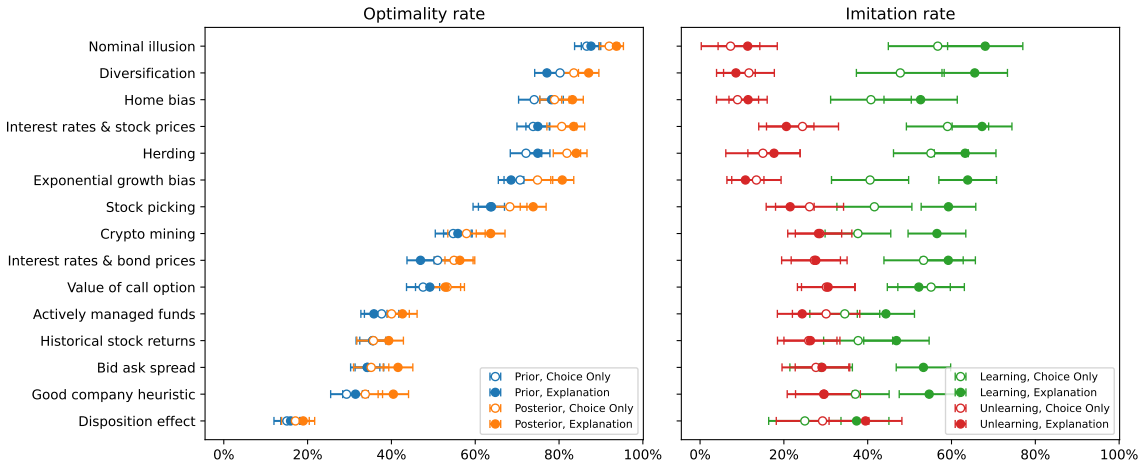
(a) Posterior optimality

(b) Imitation rate

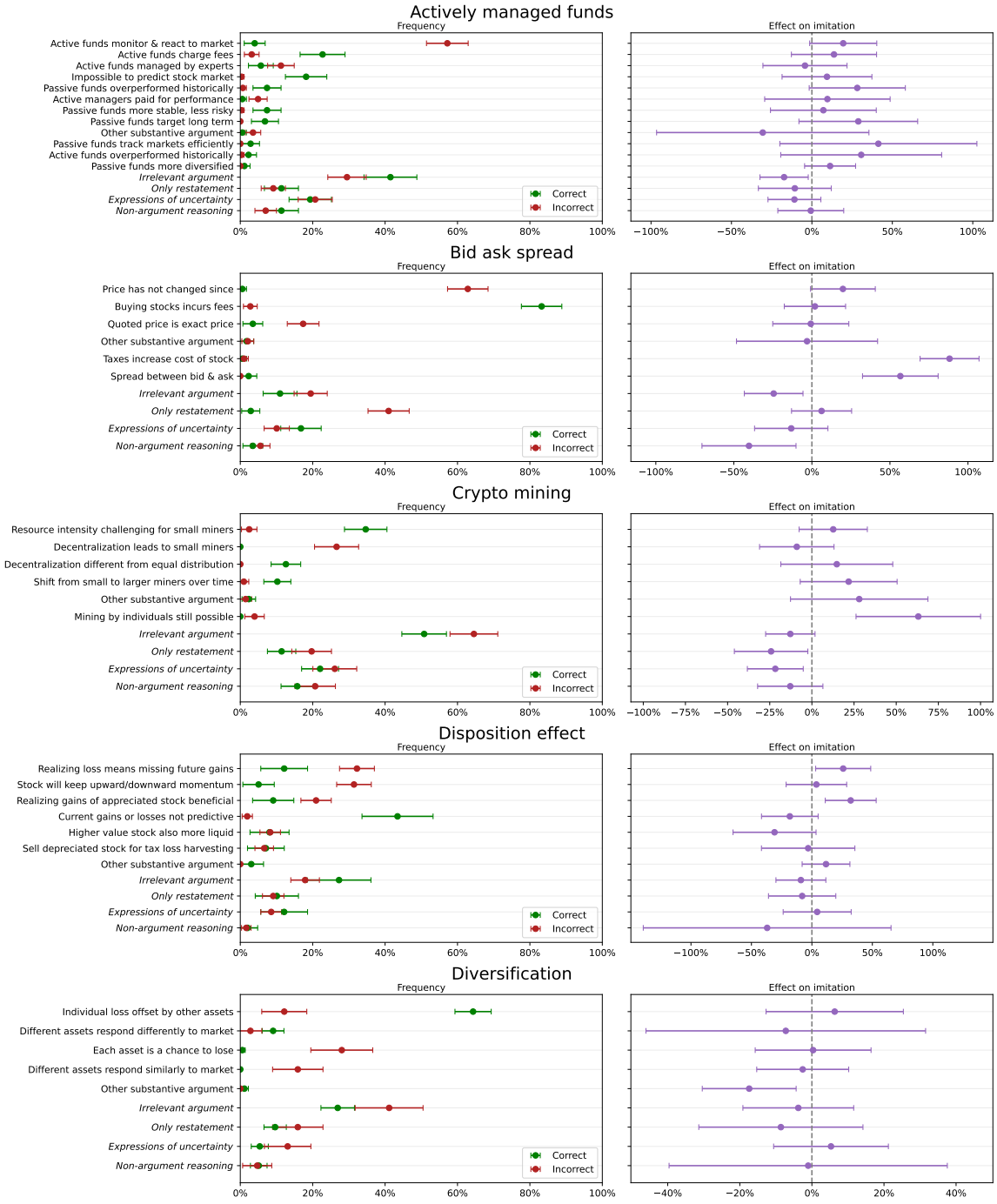
Appendix Figure D3: Effect of *Explanation* in conflicting and confirming situations. *Notes:* By construction, receiver prior optimality is 0% in *Learning* and *Both Wrong*, and 100% in *Unlearning* and *Both Right*. Whiskers show 95% CIs with clustering at the orator and receiver levels.



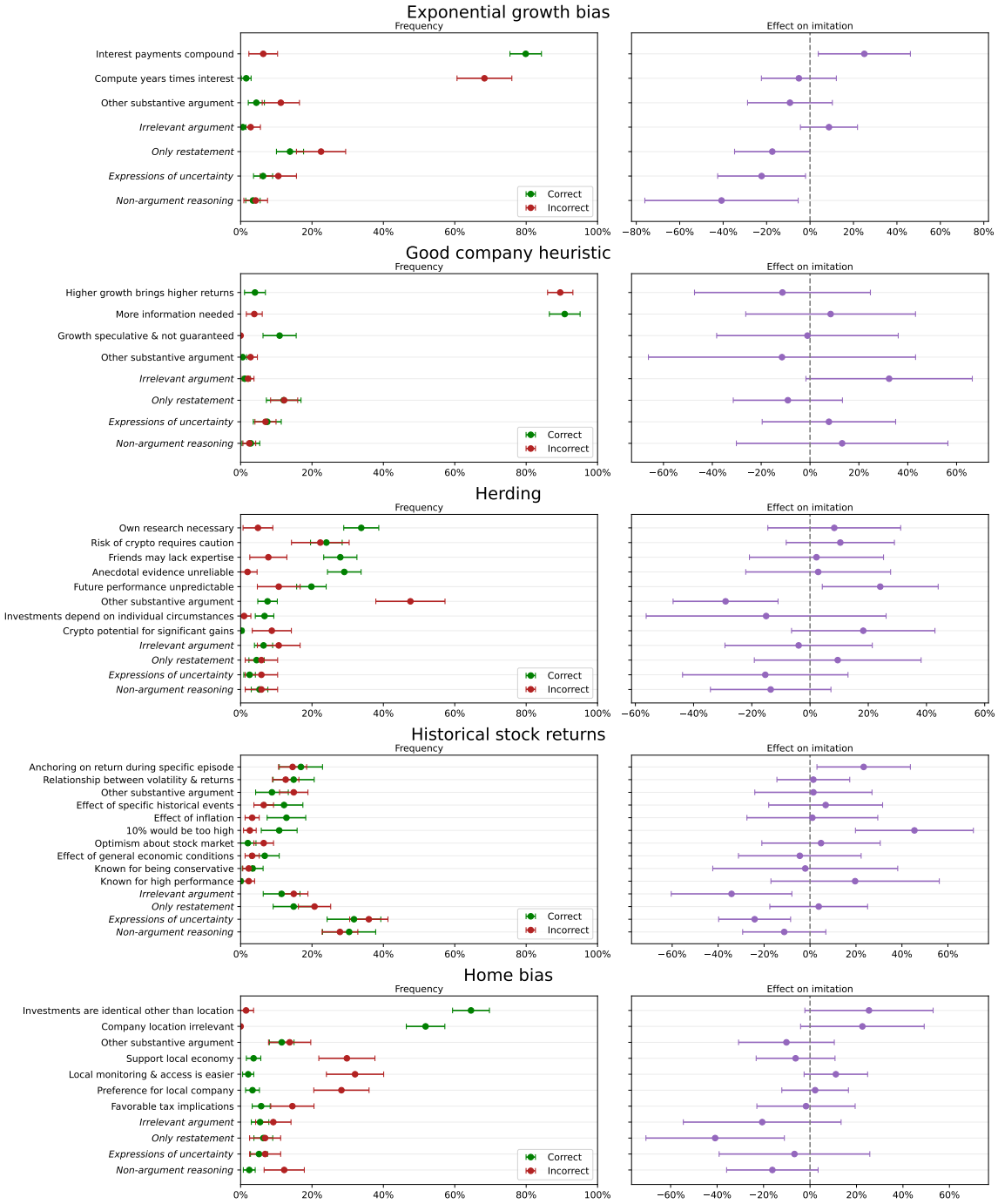
Appendix Figure D4: Richness gap by orator optimality and argument quality. *Notes: Explanation* sample is the main Receiver experiment, *Choice Only* is pooled from all collections. Whiskers show 95% CIs with clustering at the orator level.



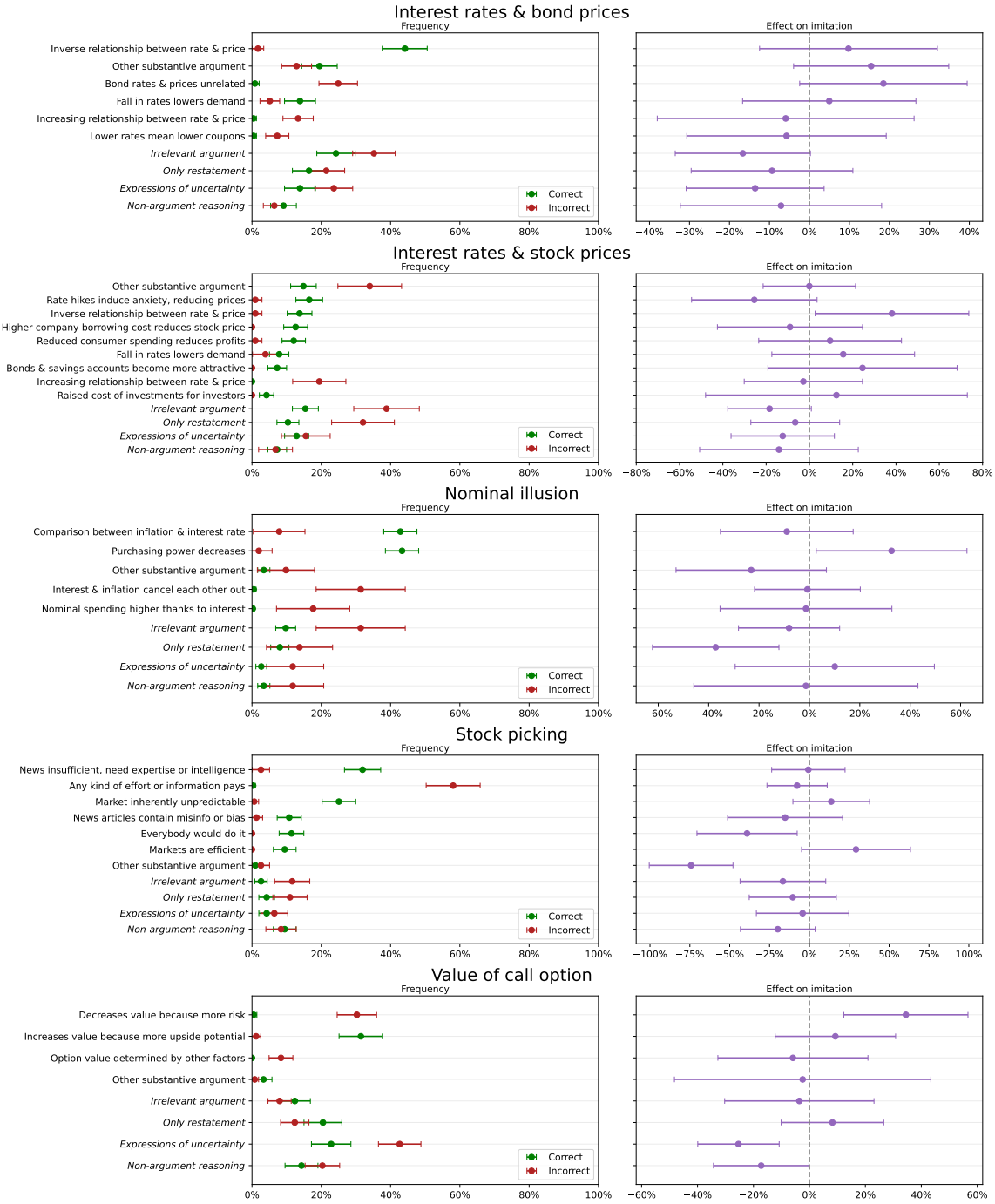
Appendix Figure D5: Effect of *Explanation* on optimality and imitation by task. *Notes:* See notes for Figure 1. Whiskers show 95% CIs with clustering at the orator and receiver levels.



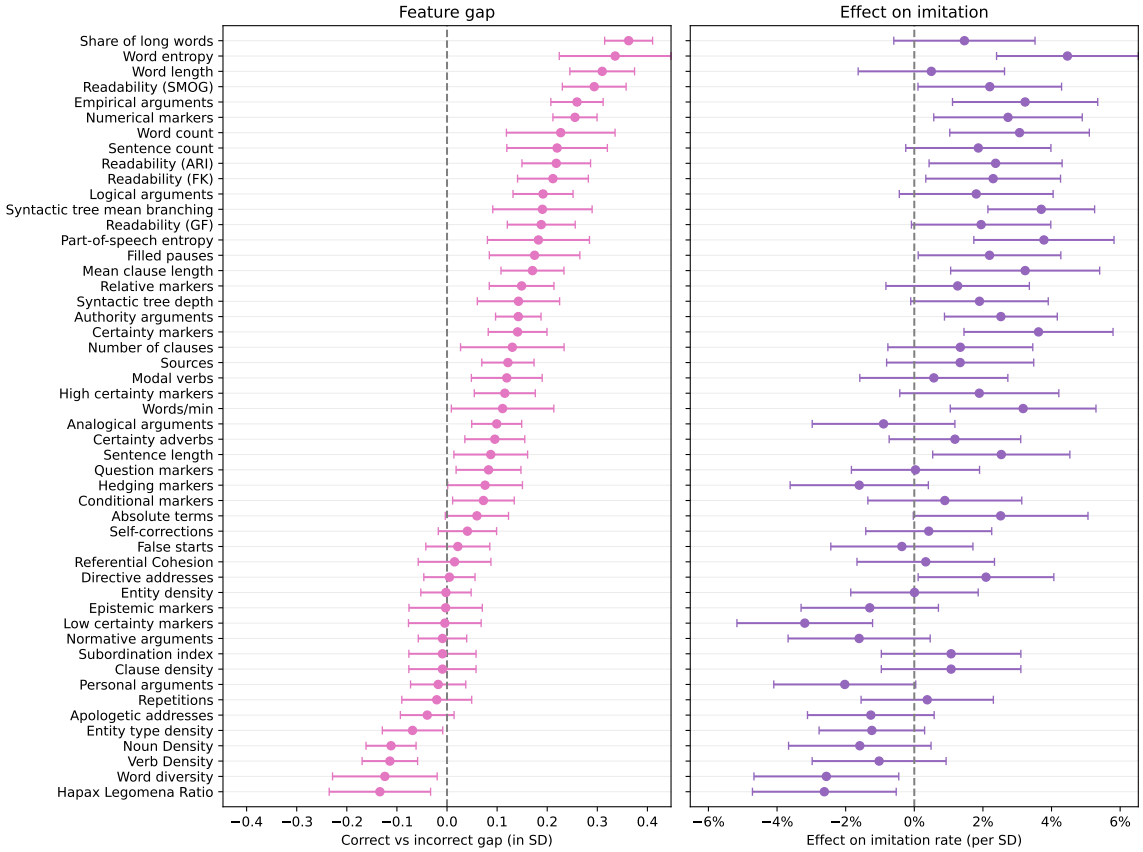
Appendix Figure D6: Arguments by task I/III. *Notes:* See notes for Figure 4.



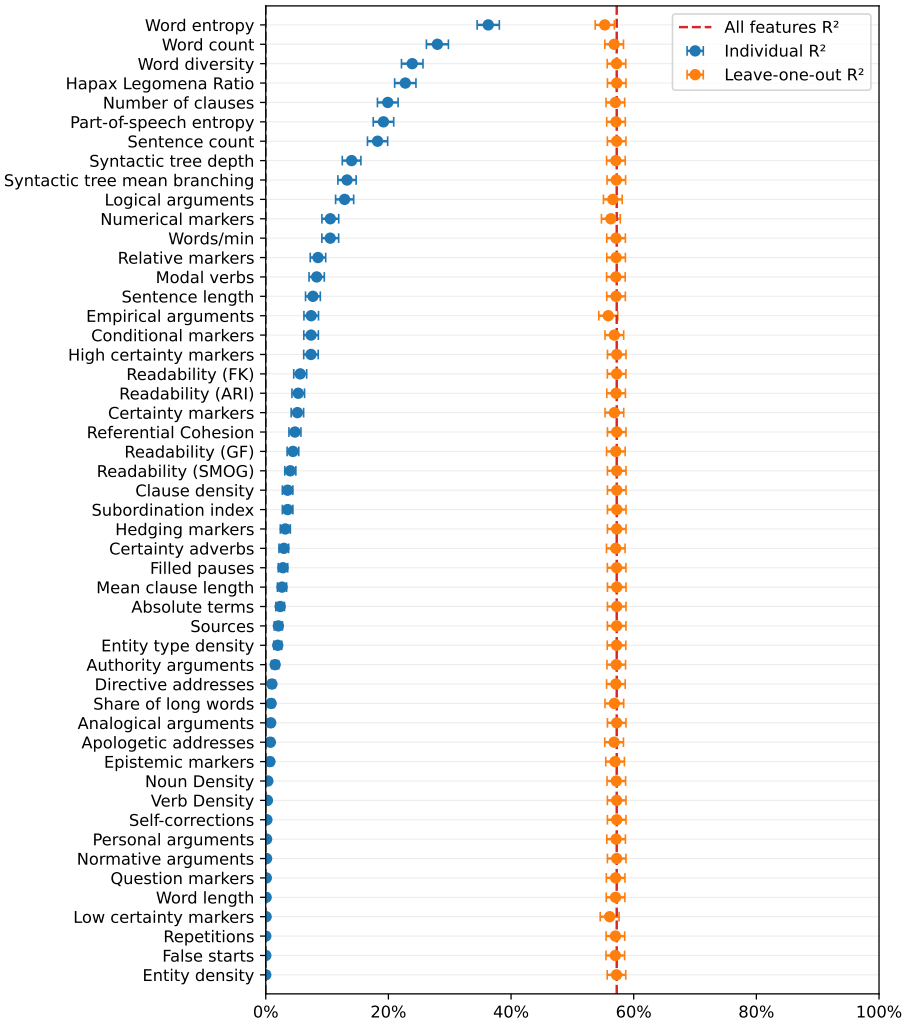
Appendix Figure D6: Arguments by task II/III. *Notes:* See notes for Figure 4.



Appendix Figure D6: Arguments by task III/III. *Notes:* See notes for Figure 4.



Appendix Figure D7: Text feature gaps and effects on imitation. *Notes:* Left panel shows difference in text features between correct and incorrect explanations, expressed in SD. Right panel shows effect of 1 SD on imitation, estimated via (2), separately for each feature. See Table B1 for list of features. Whiskers show 95% CIs with clustering at the orator and receiver levels.



Appendix Figure D8: Explanatory power of text features for richness. *Notes: All features* R^2 is the R^2 in a regression of richness on all features and a constant. *Individual* R^2 is the R^2 in a regression on a given feature and a constant. *Leave-one-out* R^2 is the R^2 in a regression on all features except the given feature. See Table B1 for list of features. Whiskers show 95% CIs with clustering at the orator levels.

E Additional tables

Appendix Table E4: Financial Decision Questions

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

Notes: Correct answers are underlined.

Appendix Table E5: Overview of data collections

Collection	Sample	Respondents	Treatments	Main outcomes	Pre-analysis plan
Baseline experiments					
Orator Experiment	Prolific	466	None	Choices in 15 financial decision tasks and voice recordings of explanations for choices	https://aspredicted.org/4t5b-42ws.pdf
Explanation Receiver Experiment	Prolific	1,103	*Choice Only* *Explanation*	Choices in 15 financial decision tasks	https://aspredicted.org/4t5b-42ws.pdf
Mechanism experiments					
Confidence Receiver Experiment	Prolific	714	*Choice Only* *Choice & Confidence*	Choices in 15 financial decision tasks	https://aspredicted.org/fwv8-3gzs.pdf
Transcript Receiver Experiment	Prolific	918	*Choice Only* *Transcript*	Choices in 15 financial decision tasks	https://aspredicted.org/y2f5-8zdp.pdf
Richness Receiver Experiment	Prolific	972	*Sparse version* *Rich version*	Choices in 15 financial decision tasks	https://aspredicted.org/9t49-kxbc.pdf
Argument Annotation Experiment	Prolific	1,967	None	Argument annotations of rich and sparse explanation transcripts across 15 financial decision tasks	https://aspredicted.org/46gr4k.pdf
Balls-and-urns experiments					
Balls-and-Urns Orator Experiment	Prolific	474	None	Estimate in standard balls-and-urns task	https://aspredicted.org/p6sd-p88k.pdf
Balls-and-Urns Receiver Experiment	Prolific	1,826	*Choice Only* *Explanation*	Estimate in standard balls-and-urns task	https://aspredicted.org/p6sd-p88k.pdf
Binary Balls-and-Urns Orator Experiment	Prolific	483	None	Binary choice in standard balls-and-urns task and voice recordings of explanations	https://aspredicted.org/5na3fq.pdf
Binary Balls-and-Urns Receiver Experiment	Prolific	1,914	*Choice Only* *Explanation*	Binary choice in standard balls-and-urns task	https://aspredicted.org/5na3fq.pdf

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

Appendix Table E6: Treatment effects on optimality

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

Notes: Choice Only is the omitted category. The p-value in a t-test for equality between *Transcript* and *Explanation* is $p < 0.01$. Samples are the corresponding Receiver experiments for *Explanation* (1,103 receivers, 13,111 obs.), *Choice & Confidence* (714 receivers, 8,531 obs.) and *Transcript* (918 receivers, 10,977 obs.), while *Choice Only* is pooled from all collections (2,642 receivers, 8,240 obs.). Standard errors are clustered at the orator and receiver levels.

Appendix Table E7: Treatment effects on imitation in learning & unlearning situations

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

Notes: Choice Only is the omitted category. The p-value in a t-test for equality between *Learning* × *Transcript* and *Learning* × *Explanation* is $p = 0.02$. Samples are as in Table E6, restricted to learning and unlearning situations. Standard errors are clustered at the orator and receiver levels.

Appendix Table E8: Decomposition of learning effect in *Transcript* treatment

	Dependent variable: Imitation in Learning & Unlearning						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Transcript	-0.010	-0.020	-0.002	0.002	0.021	0.038**	0.100**
	(0.014)	(0.042)	(0.014)	(0.014)	(0.018)	(0.015)	(0.048)
Learning	0.198***	0.196***	0.200***	0.202***	0.197***	0.206***	0.197***
	(0.017)	(0.017)	(0.017)	(0.017)	(0.021)	(0.018)	(0.022)
Transcript × Learning	0.090***	0.079***	0.080***	0.077***	0.045	0.023	0.009
	(0.022)	(0.023)	(0.022)	(0.022)	(0.027)	(0.023)	(0.028)
Orator confidence			-0.005				-0.004
			(0.008)				(0.009)
Transcript × Orator confidence			0.038***				0.019
			(0.011)				(0.011)
Argument count				-0.015*			-0.005
				(0.008)			(0.019)
Transcript × Argument count				0.051***			-0.013
				(0.011)			(0.024)
Richness						-0.011	-0.003
						(0.009)	(0.011)
Transcript × Richness						0.095***	0.078***
						(0.012)	(0.015)
Orator controls		✓					✓
Argument controls					✓		✓
Correct vs incorrect gap			0.42***	0.32***		0.77***	
Gap × Imitation			0.016	0.017		0.073	
Share explained	0%	12%	11%	15%	51%	74%	90%
Observations	7900	7900	7900	7900	7900	7900	7900
R^2	0.071	0.075	0.074	0.075	0.078	0.085	0.089

Notes: See notes for Table 1. *Transcript* sample is the *Transcript* survey, *Choice Only* sample is pooled from all collections. Standard errors are clustered at the orator and receiver levels.

Online Appendix for "Explanations"

Thomas Graeber Christopher Roth Constantin Schesch

June 29, 2026

This Online Appendix contains a full list of identified arguments (Section F) and additional results on confidence (Section G) and respondent characteristics (Section H) in our main experiment. Furthermore, it provides additional results on our richness manipulation experiment (Section I), the argument annotation experiment (Section J), binary Balls-and-urns (Section K) and continuous Balls-and-urns (Section L) experiments. Finally, it contains survey screens (Section M). The full text of all surveys is reproduced in an Online Supplement.

F List of identified arguments

Table F1 shows all naturally-occurring arguments in the 15 financial reasoning tasks we study in our main experiment, as identified by our annotation scheme detailed in Appendix B.

Table F1: Arguments Table

Argument	Description	Category
Task: Actively managed funds		
Active funds monitor & react to market	Actively managed funds can monitor and quickly adapt to market changes.	Fallacious
Impossible to predict stock market	Human inability to predict market movements, performance pressure, errors or overconfidence limit the effectiveness of active management.	Sound
Active funds managed by experts	Expertise in active management can lead to better investment decisions.	Fallacious
Active managers paid for performance	Active managers get paid because clients expect them to bring higher results than passive funds.	Fallacious
Active funds overperformed historically	References to historical data showing active management's performance.	Fallacious
Passive funds overperformed historically	References to historical data showing passive management's performance.	Fallacious
Passive funds more stable, less risky	Passively managed funds maintain stability by not frequently changing investments, while actively managed funds are risky investments.	Sound
Passive funds more diversified	Passive management benefits from diversification across a broad market index.	Sound

Continued on next page

Argument	Description	Category
Active funds charge fees	Investment fees of actively managed funds are higher than for passive management. They reduce net returns and negate potential gains.	Sound
Passive funds target long term	Passively managed funds tack market trends over the longer term, so that they are better at delivering long-term growth.	Fallacious
Passive funds track markets efficiently	Passively managed funds can achieve long-term growth by following market trends. Passive management is efficient in tracking market performance with minimal intervention.	Sound
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant
Task: Bid ask spread		
Spread between bid & ask	Stocks have a bid and an ask price, and one can only buy the stock at the ask price which is always higher than the midpoint.	Sound
Buying stocks incurs fees	Buying stock through an online broker incurs additional fees, leading to a cost higher than the stock's listed price.	Sound
Quoted price is exact price	The cost of purchasing a stock is exactly the listed trading price if no fees are applied.	Fallacious
Taxes increase cost of stock	The cost of purchasing the stock is higher because of taxes.	Fallacious
Price has not changed since	Since the price hasn't changed since it was quoted, the stock can be bought at this exact price.	Fallacious
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant
Task: Crypto mining		
Resource intensity challenging for small miners	Bitcoin mining requires significant energy and resources, making it difficult for small miners. Large miners have an economic advantage in Bitcoin mining due to their scale and resources. Mining may not be profitable for small miners.	Sound
Mining by individuals still possible	Despite challenges, mining Bitcoin by individuals on a small scale is still possible, so that small miners dominate.	Fallacious
Decentralization different from equal distribution	Decentralization means that everyone can mine, but not that everyone mines equally, so that in practice larrge miners dominate.	Sound
Decentralization leads to small miners	Decentralization means there is no central planner, so that it leads to a diversity of miners, in which small miners dominate.	Fallacious
Shift from small to larger miners over time	There has been a historical shift from small miners to large mining operations over time.	Sound
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant
Task: Disposition effect		
Sell depreciated stock for tax loss harvesting	Selling a stock that has lost value can be beneficial for tax purposes, allowing for tax loss harvesting.	Sound
Realizing loss means missing future gains	A stock that has lost value may have the potential to increase in value in the future, making it unwise to sell. Avoid selling stocks at a loss to prevent realizing the loss and potentially missing out on future gains.	Fallacious
Realizing gains of appreciated stock beneficial	Selling a stock that has gained value realizes the profit, ensuring a positive return on investment.	Fallacious
Stock will keep upward/downward momentum	One should keep the stock that has gone up and sell the stock that has gone down, because these trends can be expected to continue in the future.	Fallacious
Current gains or losses not predictive	Stock values fluctuate, so current losses or gains do not reflect future performance.	Sound
Higher value stock also more liquid	The stock with the highest value will also be more liquid, one should therefore sell that one.	Fallacious
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant

Continued on next page

Argument	Description	Category
Task: Diversification		
Individual loss offset by other assets	Investing in multiple assets prevents total loss if one specific investment fails, akin to not putting all eggs in one basket.	Sound
Different assets respond differently to market	Different assets respond differently to market changes, so spreading investments can mitigate losses due to geopolitical or macroeconomic events.	Fallacious
Different assets respond similarly to market	Different assets usually respond similarly to market changes, so that it does not change much to invest in multiple assets instead of a single one.	Fallacious
Each asset is a chance to lose	Each asset is a chance to lose, so investing in multiple assets increases the chances of losing money.	Fallacious
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant
Task: Exponential growth bias		
Interest payments compound	The total amount in the savings account increases due to compound interest, where interest is earned on both the initial principal and the accumulated interest from previous periods.	Sound
Compute years times interest	A simple calculation of 2% interest per year on the initial $100, leading to a total of $110 after five years without considering compound interest.	Fallacious
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant
Task: Good company heuristic		
Higher growth brings higher returns	Investing in the firm with higher growth prospects will yield higher returns due to its potential for growth.	Fallacious
Growth speculative & not guaranteed	Growth prospects are speculative and not a guaranteed indicator of future success, thus more information is needed.	Fallacious
More information needed	More information is needed to make a decision because the provided details are insufficient.	Sound
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant
Task: Herding		
Future performance unpredictable	Past performance of cryptocurrencies does not guarantee future results. Timing the market correctly when investing in cryptocurrencies is not possible.	Sound
Own research necessary	It is important to conduct one's own research before investing in cryptocurrencies.	Fallacious
Risk of crypto requires caution	High volatility, risk of scams, lack of backing and other risks associated with cryptocurrencies are a reason for caution.	Fallacious
Friends may lack expertise	Friends providing advice may lack expertise in financial markets or cryptocurrencies.	Sound
Anecdotal evidence unreliable	Anecdotal evidence from friends is not a reliable basis for investment decisions, can be coincidence, luck etc.	Sound
Investments depend on individual circumstances	Investment decisions should be based on individual circumstances and not influenced by others. Cryptocurrencies may not be suitable for all investors.	Sound
Crypto potential for significant gains	Cryptocurrencies have the potential for significant gains from investing.	Fallacious
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant
Task: Historical stock returns		
Effect of inflation	Arguments that consider the impact of inflation on the average annual return.	Other
Relationship between volatility & returns	Arguments about how economic volatility affects the stock market's performance.	Other
Optimism about stock market	Arguments expressing a general optimism about the stock market's performance and long-term growth.	Other

Continued on next page

Argument	Description	Category
Effect of general economic conditions	Arguments considering the general economic conditions and their impact on the stock market.	Other
Effect of specific historical events	Arguments considering the impact of specific historical economic events on the stock market, such as COVID-19 pandemic, recessions and subsequent recoveries etc.	Other
Anchoring on return during specific episode	Arguments where some remembrance of a specific or general stock returns is used as an anchor for the average return of the S&P 500.	Other
Known for high performance	The S&P500 is known for its high performance, which is why it has a historical average return above 10%.	Other
Known for being conservative	The S&P500 is known for being a popular, steady and conservative investment, which is why it has a historical return below 10%.	Other
10% would be too high	Arguments based on the idea that a historical return above 10% seems too high. This can also involve the idea that, if that were true, everybody would be investing in the S&P500, which is not true and/or would reduce the return.	Other
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant
Task: Home bias		
Company location irrelevant	The location of a company's headquarters does not impact its investment value.	Sound
Support local economy	Investing in a company headquartered in one's home state supports the local economy and community. This can also happen via taxes being paid in one's home state.	Fallacious
Local monitoring & access is easier	Investing in a company in one's home state allows for easier monitoring and access to the company.	Fallacious
Favorable tax implications	The choice between investing in a home state or out-of-state company may be influenced by different tax implications.	Fallacious
Preference for local company	A preference or bias towards investing in companies headquartered in one's home state.	Fallacious
Investments are identical other than location	Both investment options are considered equally good due to the companies being identical except for location.	Sound
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant
Task: Interest rates and bond prices		
Inverse relationship between rate & price	Since there is an inverse relationship between interest rates and bond prices, bond prices will increase when the interest rate falls.	Sound
Increasing relationship between rate & price	Since there is a relationship in the same direction between interest rates and bond prices, bond prices will fall when the interest rate falls.	Fallacious
Fall in rates lowers demand	A fall in the interest rate leads to less demand and therefore a higher price of bonds.	Fallacious
Bond rates & prices unrelated	Bond prices remain stable and are not influenced by fluctuations in interest rates.	Fallacious
Lower rates mean lower coupons	Since the interest rate determines the interest payment that bondholders get from holding the bond, the bond's value will go down if the interest rate goes down.	Fallacious
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant
Task: Interest rates and stock prices		
Inverse relationship between rate & price	Since there is an inverse relationship between interest rates and stock prices, stock prices will increase when the interest rate falls.	Sound
Increasing relationship between rate & price	Since there is a relationship in the same direction between interest rates and stock prices, stock prices will fall when the interest rate falls.	Fallacious
Fall in rates lowers demand	A fall in the interest rate leads to less demand and therefore a lower price of stocks.	Fallacious
Higher company borrowing cost reduces stock price	Higher interest rates increase borrowing costs for companies, reducing their profitability and negatively affecting stock prices.	Sound
Bonds & savings accounts become more attractive	Higher interest rates make bonds and savings accounts more attractive compared to stocks, leading investors to shift their investments.	Fallacious
Reduced consumer spending reduces profits	Higher interest rates reduce consumer spending (e.g. due to borrowing constraints), negatively affecting company profits and stock prices.	Sound

Continued on next page

Argument	Description	Category
Raised cost of investments for investors	Interest rate increases raise the cost of investments, making it more expensive for investors and negatively affecting stock prices.	Fallacious
Rate hikes induce anxiety, reducing prices	Interest rate hikes make market participants uncertain and anxious, which reduces stock prices.	Fallacious
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant
Task: Nominal illusion		
Comparison between inflation & interest rate	Since the inflation rate is higher than the interest rate, one would be able to buy less tomorrow than today. This argument is distinct from PurchasingPowerDecrease because it displays no understanding of the mechanisms behind inflation and interest, and is based solely on a comparison of numbers.	Sound
Purchasing power decreases	Even though the amount of money in a savings account has increased thanks to the interest rate, the price at which one needs to buy goods and services will have increased more because of the comparatively higher inflation, so that the net effect on real purchasing power is higher. This argument is distinct from NumericalComparison because it displays an understanding of the mechanisms behind inflation and interest, not just a comparison of numbers.	Sound
Nominal spending higher thanks to interest	Because the amount of money in the savings account has increased thanks to the interest rate, one would be able to spend more than today.	Fallacious
Interest & inflation cancel each other out	Because the interest rate and inflation rate both cancel out, one would be able to buy exactly as much tomorrow as today.	Fallacious
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant
Task: Stock picking		
Everybody would do it	If it was possible to outperform the stock market by reading free online news, everybody would be doing it.	Sound
Markets are efficient	All publicly available information is already factored into stock prices, so that markets will already have adjusted to stale news.	Sound
News articles contain misinfo or bias	News articles can contain misinformation or bias, leading to poor investment decisions.	Sound
Market inherently unpredictable	The stock market is inherently volatile and unpredictable, making systematic outperformance difficult.	Sound
News insufficient, need expertise or intelligence	News are not enough for everyday people to outperform the stock market, since, for example, they also need to be specially smart, to have financial expertise and/or to have access to other sources of information.	Fallacious
Any kind of effort or information pays	Any kind of effort, research or information will help to outperform the stock market.	Fallacious
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant
Task: Value of call option		
Increases value because more upside potential	Higher volatility in a stock increases the potential for larger price movements, which can be advantageous for call option holders seeking to profit from upward stock movements.	Sound
Decreases value because more risk	Higher volatility is seen as increasing risk, making the call option less attractive and decreasing its value due to the unpredictability of stock price movements.	Fallacious
Option value determined by other factors	The volatility of a stock has no direct effect on the value of a call option because the call option's value is determined by other factors, not just the stock's volatility.	Fallacious
Other substantive argument	Any other substantive argument not part of the other categories.	Other
Irrelevant argument	Argument unrelated to the question; or no answer is implied by the argument.	Irrelevant

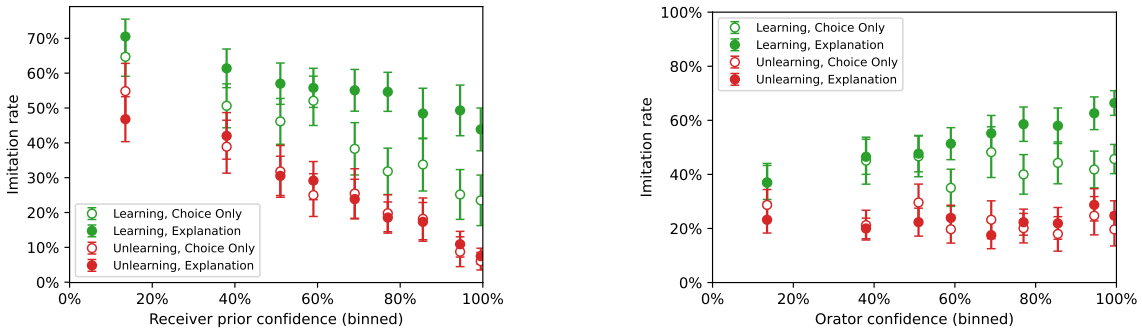
G Additional results on confidence

G.1 Heterogeneity by orator confidence

Section 4.1 analyzes the role of orator confidence in the treatment effect of explanations and asks whether they are equivalent to confidence statements. It documents that confidence plays a modest role in the overall effect. We nonetheless investigate more finely how orator confidence shapes social learning. Figure G1b shows that imitation generally slopes up in orator confidence, with an especially strong slope in learning situations. Among these, there is no visible explanation effect at any confidence level. Conversely, although learning effects appear weaker up to 60% confidence, they are strong and growing above this threshold. Across learning and unlearning situations, imitation is 11.6 p.p. higher for explanations where the orator has above-median confidence ($p < 0.01$). This gap is 6.5 p.p. ($p = 0.02$) in unlearning situations. For listeners exposed to orators with confidence below the median, the learning effect is 8.6 p.p. ($p < 0.01$), but twice as large at 15.6 p.p. ($p < 0.01$) above the median, also a 7.1 p.p. difference ($p = 0.09$). The learning benefit of explanations is thus amplified by confident and correct orators.

G.2 Heterogeneity by receiver confidence

We turn to heterogeneity in receiver confidence. Receiver prior confidence reflects both individual differences and task-level effects: a random effects decomposition attributes 43.2% of the variation in receiver prior confidence to persistent between-receiver differences, with the remainder reflecting within-receiver variation across tasks. One intuitive possibility is that explanations reduce the confidence threshold below which individuals are willing to imitate others. Figure G1a displays the imitation rate by bins of receiver prior confidence in *Choice Only* and *Explanation*, separately for learning and unlearning situations. As one would expect, imitation generally slopes down in receiver prior confidence. Just as there is no average effect of *Explanation* in unlearning situations, there is no statistically significant effect within any of the confidence bins. On the other hand, in learning situations, effects are relatively weak in the first bins up to 60% confidence, after which the treatment effect becomes much stronger. More precisely, splitting the sample by the overall median confidence of 73%, the net learning rate of explanations is 9.8 p.p. ($p < 0.01$) for unconfident receivers and twice as large at 19.6 p.p. ($p < 0.01$) for confident ones, a significant 9.8 p.p. difference ($p = 0.02$). The learning benefit of explanations is therefore even greater for confident but wrong receivers.



(a) Imitation by prior confidence

(b) Imitation by orator confidence

Figure G1: Effect of prior and orator confidence on imitation in learning and unlearning situations. *Notes:* Bin sizes range from 143 to 535 observations for receiver prior confidence and from 139 to 601 for orator confidence. *Explanation* sample is the main Receiver experiment (1,103 receivers, 13,111 obs.), *Choice Only* is pooled from all collections (2,642 receivers, 8,240 obs.). Whiskers show 95% CIs with clustering at the orator and receiver levels.

G.3 Interaction of orator and receiver confidence

Figure G2 shows how imitation by receiver and orator confidence combine by showing the difference in net learning rate between *Explanation* and *Choice Only* by orator and receiver prior confidence. The net learning rate is defined as the difference in imitation rates between learning and unlearning situations. We split both variables by the aggregate median confidence of 73%. We find that confidence effects largely compound: the explanation effect on net learning is not significant when both orator and receiver are below the median; it is significant and of similar magnitude when either one is above the median; and it is highest when both are confident.

G.4 Confidence and the content of explanations

We shed light on the mechanism behind these patterns by analyzing the interplay of confidence and the content of explanations. We first repeat our previous analysis while also including the *Choice & Confidence* treatment. In Figure G3a, the effects of *Choice & Confidence* and *Choice Only* by receiver prior appear identical. Confidence signals do not help confident but wrong receivers adjust their answer. Although Figure G3b may appear to show a slightly larger effect of *Choice & Confidence* relative to *Choice Only* for confident orators in learning situations, the difference is not significant above ($p = 0.37$) or below ($p = 0.08$) median confidence. Numerical certainty statements therefore do not explain the effect of confidence, highlighting the importance of

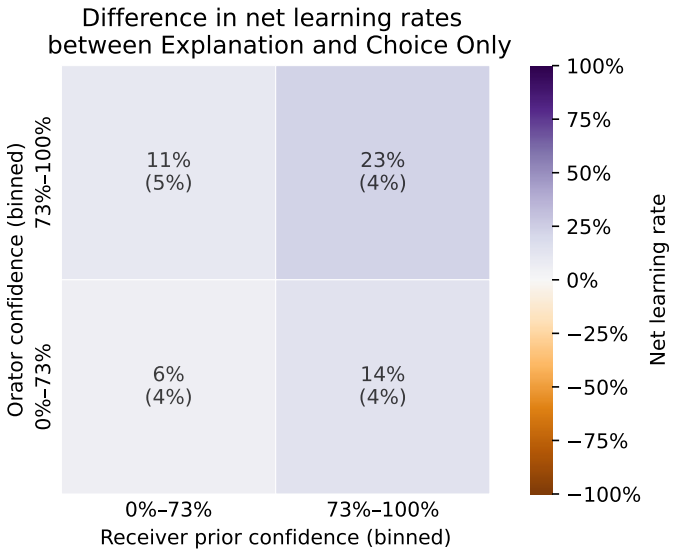
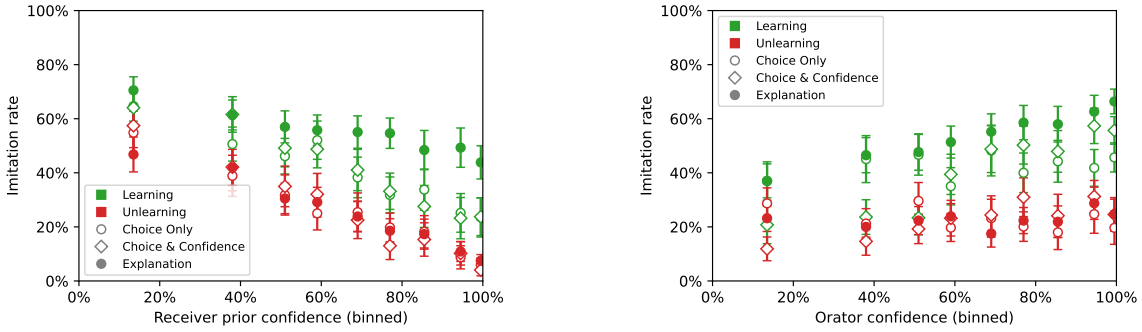


Figure G2: Effect of prior and orator confidence on imitation. *Notes:* Parentheses show SEs. *Explanation* sample is the main Receiver experiment (1,103 receivers, 13,111 obs.), *Choice Only* is pooled from all collections (2,642 receivers, 8,240 obs.).

explanation content.[31]



(a) Imitation by prior confidence

(b) Imitation by orator confidence

Figure G3: Effect of prior and orator confidence on imitation in *Choice & Confidence* treatment. *Notes:* See notes for Figure G2. Sample for *Choice & Confidence* is the corresponding receiver experiment (714 receivers, 8,531 obs.). Whiskers show 95% CIs with clustering at the orator and receiver levels.

We therefore turn to the effect of orator confidence on content differences in explanations. Figure G4a plots the average number of high certainty markers (e.g., "I am certain that", "I am

[31]Interestingly, even in the *Choice & Confidence* treatment which makes the latter very salient, there is no effect of orator confidence on imitation in unlearning situations.

sure that") by orator confidence. The number of markers increases with confidence, but there is no clear difference between correct and incorrect explanations. Low certainty markers (e.g., "It might", "I'm not sure but"), shown in Figure G4b, conversely decrease with confidence, though right and wrong explanations do not significantly differ over most of the confidence range.

We then examine the effect of confidence on intellectual content by computing the shares of argument qualities in Figure G4c. The presence of sound arguments for correct explanations and of fallacious arguments for incorrect explanations clearly trends up in orator confidence. Other categories, and in particular no argument, in turn trend down, though there is no significant differentiating pattern between situations or confidence.

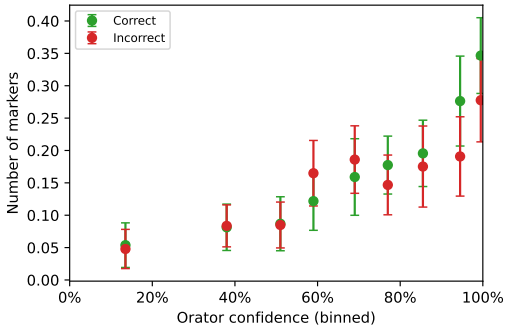
Finally, we focus on our main predictive variable by displaying richness by orator confidence in Figure G4d. We document three patterns. First, richness slopes up in orator confidence. Second, there is a richness gap between correct and incorrect explanations at all levels of confidence. Third, this gap becomes larger when orators become more confident. More precisely, the richness gap between correct and incorrect explanations is 0.48 SD ($p < 0.01$) for orators with confidence below the median against 0.91 SD ($p < 0.01$) with confidence above it, a 0.42 SD difference ($p < 0.01$) which is slightly more than half the size of the unconditional richness gap.

In conclusion, even within learning and unlearning situations, we uncover important differences in the content of explanations given by orators with differing certainties.

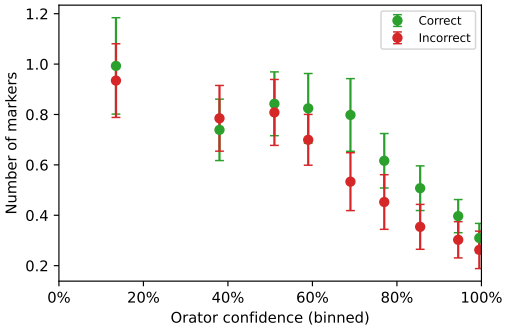
G.5 Richness imitation and confidence

Just as there are substantial content differences driven by the orator side, receivers could differentially imitate explanations depending on their own and the orator's confidence. In particular, they could react differently to richness, which summarizes a broad range of explanation features.

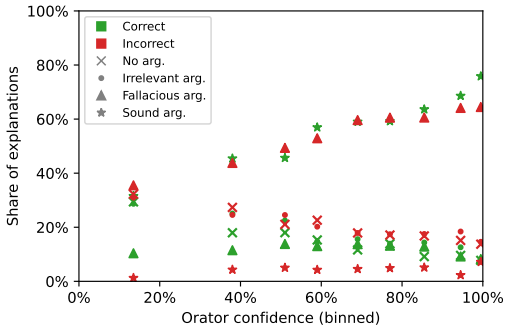
Table G2 repeats our main decomposition, from Columns (1) and (7) in Table 1, separately for confident and unconfident receivers. Columns (1) and (3) reflect our earlier finding that explanations have a learning effect for both, but that it is about two times larger for confident receivers. Columns (2) and (4) show that the imitation of richness is approximately the same for both, around 10 p.p., with an insignificant difference ($p = 0.10$). Since the richness gap between correct and incorrect explanations experienced by receivers is the same in both groups, richness therefore drives the same amount of learning asymmetry, of approximately $0.1 \times 0.744 = 7.4$ p.p. This means it effectively explains the entire learning asymmetry for unconfident receivers, but only half of it for confident receivers. Put differently, confident receivers receive a benefit



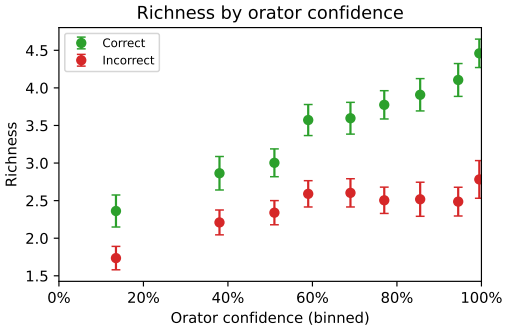
(a) High certainty markers by orator confidence



(b) Low certainty markers by orator confidence



(c) Argument quality by orator confidence



(d) Richness by orator confidence

Figure G4: Explanation content differences by orator confidence in learning and unlearning situations. *Notes:* Sample is the Orator experiment (466 orators, 6,910 obs.). For (c), see argument categorization in Section 4.3. Whiskers show 95% CIs with clustering at the orator level. In Panel (c), CIs are not shown for legibility, but all SEs are smaller than 4.4%.

from explanations that is not explained by differential imitation of richness.

Turning to orators, we repeat the same decomposition separately for confident and unconfident orators in Table G3. Columns (1) and (3) again reflect our finding that explanations have a roughly two times larger learning effect when given by confident orators. Columns (2) and (4) establish that the imitation of richness is approximately the same in both samples in this case too, around 10 p.p., also with an insignificant difference ($p = 0.32$). Since we have found that the richness gap is 0.48 SD among unconfident orators but two times larger at 0.91 SD among confident ones, richness drives a two times larger learning effect among the latter. In turn, the residual not explained by richness is also roughly twice as large. In conclusion, the difference in explanation richness accounts for about half of the higher imitation of confident correct orators.

In conclusion, we have found that explanations have a learning benefit for confident and unconfident receivers, that this benefit is larger for more confident receivers, but that this effect is not driven by differential imitation of richness. It therefore reflects either content differences other than richness or a greater propensity to change their answer based on the arguments contained in explanations. In parallel, we have documented that explanations also have a larger learning benefit when they are delivered by confident orators, and that about half of this effect is driven by the fact that the richness gap between correct and incorrect orators is larger when they are more confident, which leads listeners to imitate correct and confident orators more.

Table G2: Richness imitation by prior confidence

	Dependent variable: Imitation in Learning & Unlearning			
	Receiver prior confidence ≤ 73%		Receiver prior confidence > 73%	
	(1)	(2)	(3)	(4)
Intercept	0.347***	0.339***	0.118***	0.122***
	(0.017)	(0.019)	(0.011)	(0.012)
Explanation	-0.008	0.055**	0.007	0.048***
	(0.022)	(0.023)	(0.014)	(0.016)
Learning	0.165***	0.175***	0.169***	0.163***
	(0.023)	(0.025)	(0.024)	(0.025)
Explanation × Learning	0.098***	0.013	0.196***	0.126***
	(0.030)	(0.031)	(0.030)	(0.031)
Richness		-0.016		0.008
		(0.013)		(0.012)
Explanation × Richness		0.122***		0.085***
		(0.016)		(0.017)
Observations	4849	4849	3940	3940
R^2	0.054	0.077	0.133	0.155

Notes: See notes for Table 1. 73% is the median pooled orator and receiver prior confidence. We drop the 0.5% of observations with missing receiver prior confidence from all regressions. Standard errors are clustered at the orator and receiver levels.

Table G3: Richness imitation by orator confidence

	Dependent variable: Imitation in Learning & Unlearning			
	Orator confidence ≤ 73%		Orator confidence > 73%	
	(1)	(2)	(3)	(4)
Intercept	0.244***	0.226***	0.208***	0.210***
	(0.013)	(0.016)	(0.016)	(0.017)
Explanation	-0.029*	0.035*	0.036*	0.072***
	(0.017)	(0.020)	(0.022)	(0.022)
Learning	0.176***	0.192***	0.225***	0.221***
	(0.024)	(0.024)	(0.023)	(0.025)
Explanation × Learning	0.086***	0.030	0.156***	0.077**
	(0.029)	(0.030)	(0.030)	(0.032)
Richness		-0.033**		0.005
		(0.014)		(0.013)
Explanation × Richness		0.113***		0.089***
		(0.018)		(0.016)
Observations	4583	4583	4206	4206
R^2	0.061	0.073	0.124	0.144

Notes: See notes for Table G2.

G.6 Changes in confidence

We also examine how explanations affect receivers' posterior confidence. Figure G5 plots the change in confidence between the posterior and prior answer, separately by situation and treatment. Confidence increases more in *Explanation* than in *Choice Only* in all four situations. When both receiver and orator are correct, confidence rises by 9.8 p.p. in *Explanation*, compared to 6.6 p.p. in *Choice Only*. When both are wrong, the corresponding increase is 9.1 p.p. vs. 7.3 p.p. In learning situations, confidence is essentially unchanged in *Choice Only* (-0.5 p.p.) but rises by 3.0 p.p. in *Explanation*. Finally, in unlearning situations, confidence falls by 1.9 p.p. in *Choice Only* but is approximately unchanged in *Explanation* (+0.2 p.p.). Thus, explanations do not merely change choices: they also tend to increase receivers' confidence in their posterior answers.

The confidence effect is particularly pronounced in learning situations, where the *Explanation* treatment raises confidence by 3.5 p.p. relative to *Choice Only*. This likely reflects the large increase in posterior optimality documented above, but may also indicate that receivers who learn from an explanation become more confident than receivers who learn from observing a choice alone. In unlearning situations, the 2.1 p.p. relative increase in confidence is more ambiguous: explanations do not reduce unlearning, but they appear to make receivers slightly less

uncertain after exposure to conflicting advice.

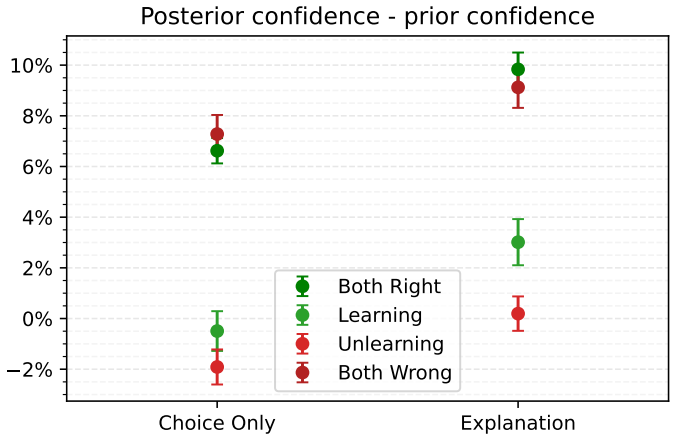


Figure G5: Difference between prior and posterior confidence by treatment. *Notes:* Sample is the main Receiver experiment for *Explanation* and all collections for *Choice Only*. Whiskers show 95% CIs with clustering at the orator and receiver levels.

To study the role of argument quality in these confidence movements, Figure G6 repeats the analysis separately by argument category. The resulting estimates are necessarily noisier, and clear monotonic patterns are hard to identify. In *Choice Only*, confidence movements vary only modestly by argument quality, as expected, since receivers do not observe the arguments. In *Explanation*, the effect of arguments appears most visible in learning situations. Relative to *Choice Only*, explanations raise confidence by 5.3 p.p. when the orator provides no argument, 5.6 p.p. when the explanation contains a fallacious argument, and 3.2 p.p. when it contains a sound argument, but only 2.2 p.p. when it contains an irrelevant argument. In unlearning situations, the corresponding effects are 3.6 p.p. for no argument, 0.0 p.p. for irrelevant arguments, 2.2 p.p. for fallacious arguments, and 4.1 p.p. for sound arguments.

These patterns should be interpreted cautiously. Changes in confidence combine two distinct margins: whether receivers change their answer, and how confident they are conditional on the answer they choose. As a result, confidence movements are not a pure measure of perceived argument quality. Still, the evidence is consistent with the broader interpretation that explanations affect not only imitation and posterior optimality, but also the confidence with which receivers hold their posterior beliefs.

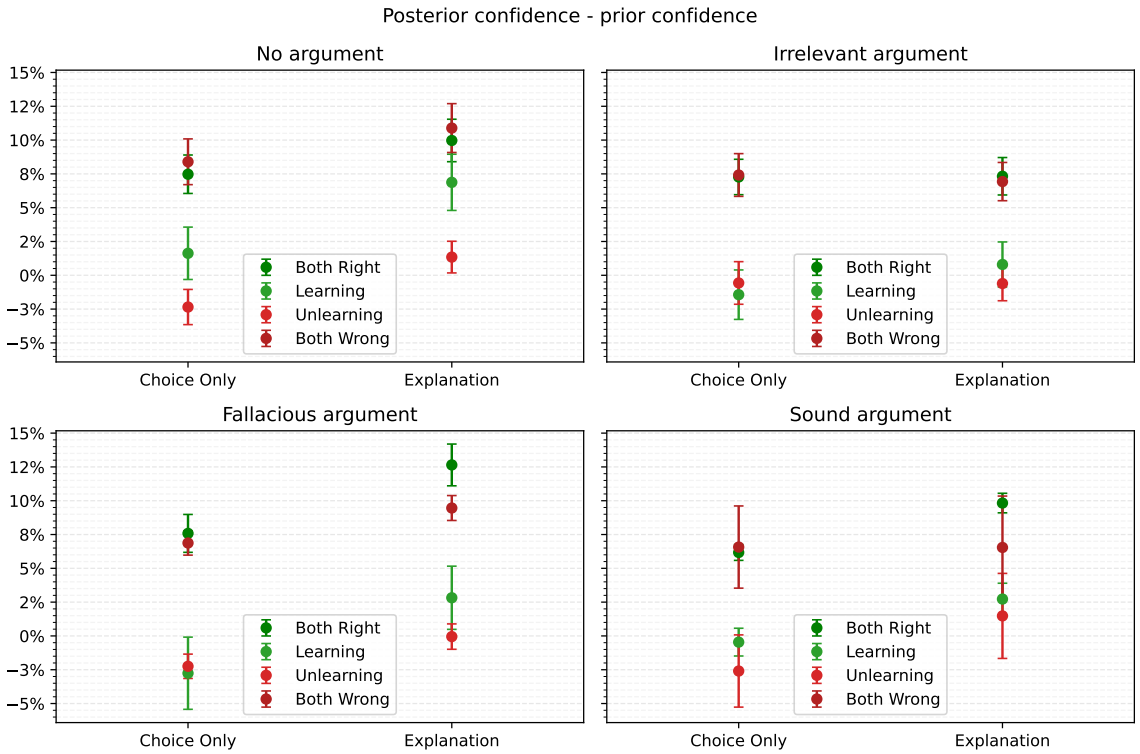


Figure G6: Difference between prior and posterior confidence by treatment and presence of arguments. *Notes:* Sample is the main Receiver experiment for *Explanation* and all collections for *Choice Only*. Whiskers show 95% CIs with clustering at the orator and receiver levels.

H Additional results on orator & receiver characteristics

We examine how orator and receiver characteristics across learning and unlearning situations, as well as differential imitation based, e.g., on sociodemographics, may contribute to the asymmetric treatment effect of explanations.

We first document the heterogeneity of participants on observable characteristics. We then study which *orator* characteristics predict imitation, and whether this helps explain the asymmetric effect above and beyond the language differences associated with different groups of orators. We then test whether the characteristics of *receivers* predict their responsiveness to explanations, and to what extent this contributes to the asymmetric effect in the aggregate.

H.1 Participant characteristics in learning and unlearning

Because receivers and orators are drawn from the same population and because learning and unlearning are determined by the same endogenous variable—receiver prior accuracy—, orators in learning and receivers in unlearning situations *should* have the same characteristics on average; similarly, orators in unlearning and receivers in learning situations should be similar. Figure H7 shows the full set of observable characteristics we elicit in the study, for orators in the left panel and receivers in the right panel, separately for learning and unlearning situations.

Among orators, we find significant differences across three of the eight characteristics we examine between those in learning and unlearning situations. The first six characteristics capture participant sociodemographics. Orators with a correct answer have more education, are more likely to be male, less likely to be Black, and similarly likely to be Republican, to be older than 35 (the approximate median in our dataset) and to be working. The remaining two features characterize participants within the context of our study: orators with correct answers have a substantially higher prior accuracy rate in the 14 remaining tasks (61% vs. 52%) and a higher confidence on the present task (72% vs. 62%). Among receivers, we find very similar patterns for those with correct and incorrect priors, with significant differences in five out of eight features.

H.2 The role of orator characteristics

Heterogeneous treatment effects by orator characteristics. We ask whether orator characteristics drive imitation patterns. We begin by examining the raw relationship between orator characteristics and the treatment effect of explanations based on the *Choice Only* and *Explanation* conditions. In the left panel of Figure H8, we show the effect of orator characteristics on imita-

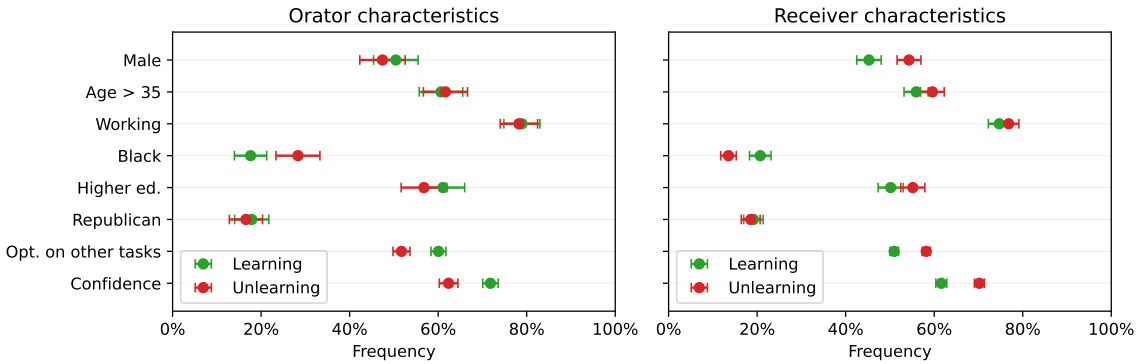


Figure H7: Characteristics of orators and receivers in learning and unlearning situations. *Notes: Optimality on other tasks* is the average optimality in the 14 other tasks. Confidence is rescaled from $[0, 100]$ to $[0, 1]$. 35 is the approximate median age in our data. Sample for both panels is the main Receiver experiment, matched with the Orator experiment for the left panel. Whiskers show 95% CIs with clustering at the orator and receiver levels.

tion, controlling for the orator's and receiver's accuracy. We find that imitation is 2.2 p.p. lower for orators aged 35 or above ($p = 0.04$), 3.6 p.p. lower for African American orators ($p = 0.01$), and 2.4 p.p. higher for orators with higher education ($p = 0.02$). Point estimates further suggest that men are imitated more, but this is not significant. No receiver characteristics significantly predict imitation.

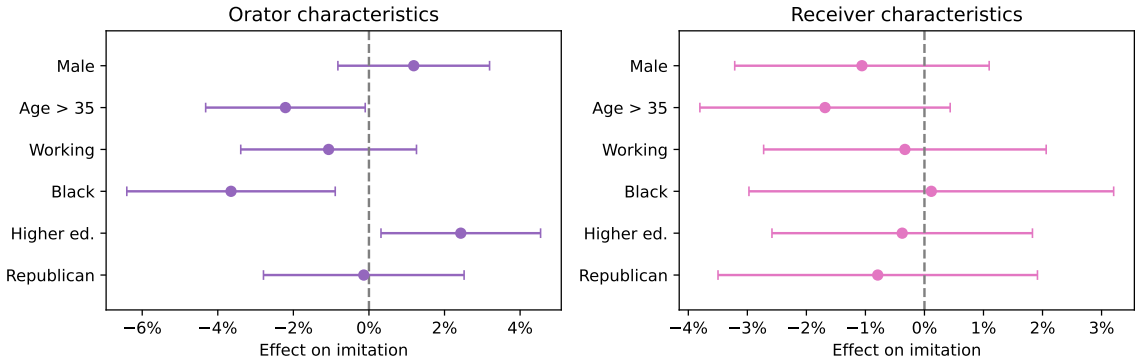


Figure H8: Effect of orator and receiver characteristics on imitation in *Explanation*. *Notes:* Left panel shows coefficients on *Explanation* interacted with orator characteristics, in a linear regression of imitation on orator and receiver optimality, *Explanation*, orator characteristics, and *Explanation* interacted with orator characteristics. Right panel shows the same for receiver characteristics. Sample for *Explanation* is the main Receiver experiment, *Choice Only* is pooled from all collections. Whiskers show 95% CIs with clustering at the orator and receiver levels.

Do orator characteristics operate via content differences? Orator characteristics could drive imitation via their content, via their oral delivery, or because the orator's voice signals sociodemographic features that could influence imitation. To separate the role of content difference, we can focus on the transcript of each explanation and repeat the previous analysis while contrasting *Choice Only* with the *Transcript* treatment. In Figure H9, we can see that the effects of orator characteristics in *Transcript* are very close to those in *Explanation*, and that none of the differences between effects are statistically significant. This suggests that the effect of orator characteristics arises not from the orator's identity as revealed by speech, but from the different kinds of explanations different demographics use.

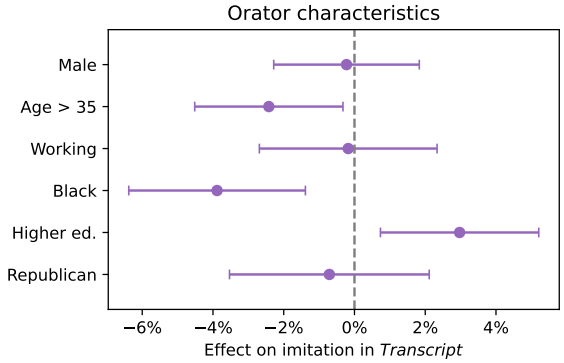


Figure H9: Effect of orator characteristics on imitation in *Transcript* treatment. *Notes:* See notes in Figure H8. *Transcript* sample is the corresponding Receiver experiment, *Choice Only* is pooled from all collections. Whiskers show 95% CIs with clustering at the orator and receiver levels.

Do orator characteristics explain away the asymmetric effect? Intuitively, orators in learning situations may signal characteristics through their voice, e.g., gender, which could make them more or less likely to be imitated. However, they also deliver different explanatory content. The preceding estimates should be interpreted as encapsulating *every* feature correlated with orator characteristics, both those contained in words and in their oral delivery.

Column (2) of Table 1 shows that orator characteristics explain away 17% of the differential effects, far below the 60% explained by richness. Moreover, columns (6) and (7) show that, once richness is controlled for, adding controls for orator characteristics does not further decrease the size of the unexplained learning effect. Figure H10 examines which orator characteristics explain away the differential effect. It shows that prior confidence and accuracy in the other tasks are most predictive, consistent with our finding that the effects of orator characteristics primarily operate through the content of explanations, though these effects are again very weak.

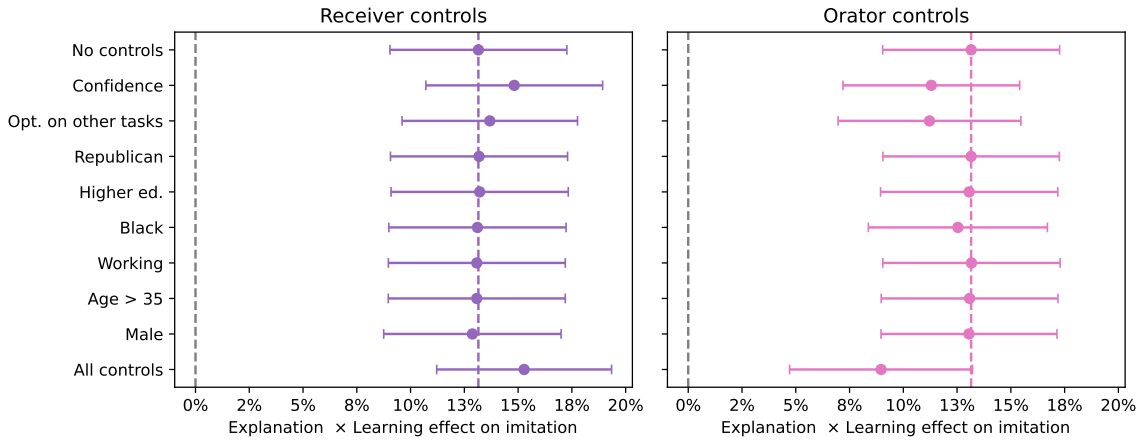


Figure H10: Learning asymmetry after controlling for orator or receiver characteristics. *Notes:* Coefficient on *Explanation* × *Learning* in a regression of imitation on *Explanation*, *Learning*, *Explanation* × *Learning*, *Control* and *Explanation* × *Control*. *Control* is a receiver control in the left panel and an orator control on the right. All regressions except 'No controls' and 'All controls' contain a single control. *Explanation* sample is the main Receiver experiment, *Choice Only* is pooled from all collections, both restricted to learning and unlearning situations. Whiskers show 95% CIs with clustering at the orator and receiver levels.

H.3 The role of receiver characteristics

We plot the effect of receiver characteristics on imitation in the left panel of Figure H8. For example, after controlling for orator and receiver accuracy, a male receiver is 1.1 p.p. less likely to imitate. Again, it should be stressed that these effects are relatively weak since none of them are significant. Unlike in our analogous study of the orator side, the receiver analysis is not affected by content differences, since the content is controlled by orators, who are randomly assigned to receivers.

H.4 The role of interacted orator and receiver characteristics

We have studied orator and receiver characteristics separately, and found relatively mild effects, but it could be that their interaction produces strong effects. For example, men might be more likely to imitate other men, women might be less likely to imitate men, etc. This example shows that, for binary characteristics like ours, there are four relevant configurations, depending on whether the orator or receiver are in the in- or out-group.

For example, a male receiver is 2.5 p.p. less likely to imitate a female orator, while a woman is 1.9 p.p. more likely to imitate a man and a man is 0.1 p.p. more likely to imitate a man. All

of these effects are relative to a woman listening to a woman, contrast *Explanation* with *Choice Only*, and control for orator and receiver accuracy. However, much like before, none of these effects appear strong enough to be significant.

Figure H12 repeats the same analysis but contrasts *Transcript* with *Choice Only*. Estimates are very similar and not statistically different from the previous ones. This suggests that the effects of orator-receiver characteristics operate, if at all, mainly through the content of explanations rather than signaling of socioeconomic characteristics.

In conclusion, although online speech-based surveys are promising to study bias in social learning or even discrimination, we do not find strong effects of orator or receiver characteristics when discussing financial reasoning tasks.

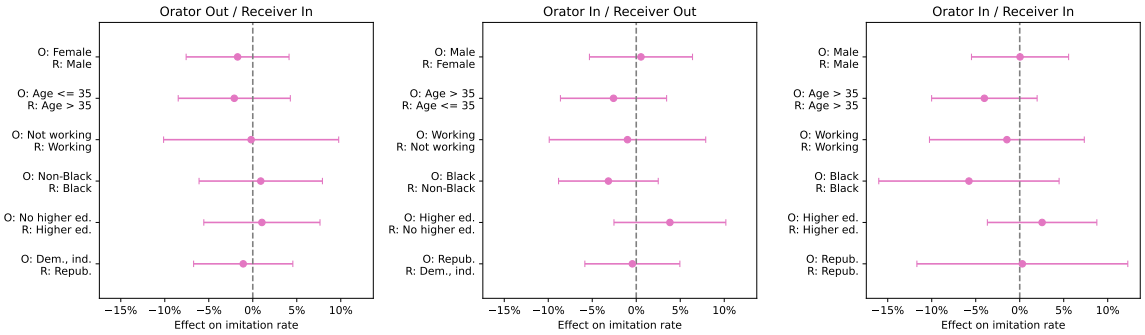


Figure H11: Effect of orator-receiver characteristics on imitation in *Explanation* treatment. *Notes:* Coefficients on *Explanation* interacted with orator-receiver characteristics, in a linear regression of imitation on orator and receiver optimality, *Explanation*, orator-receiver characteristics and *Explanation* interacted with orator-receiver characteristics. *Explanation* sample is the main Receiver experiment, *Choice Only* is pooled from all collections. Orator-receiver characteristics are an *Orator Out / Receiver In* dummy equal to 1 if the receiver has the characteristic but not the Orator; *Orator In / Receiver Out* and *Orator In / Receiver In* are analogously defined; *Orator Out / Receiver Out* is left out and serves as reference level. Whiskers show 95% CIs with clustering at the orator and receiver levels.

H.5 The effect of orator and receiver characteristics on imitation of richness

To better understand why respondents tend to imitate richness, it is interesting to explore whether characteristics of orators or receivers predict the imitation of richness. We therefore focus on learning and unlearning situations, and regress imitation on a dummy for learning situations and richness as well as richness interacted with orator characteristics, all interacted with

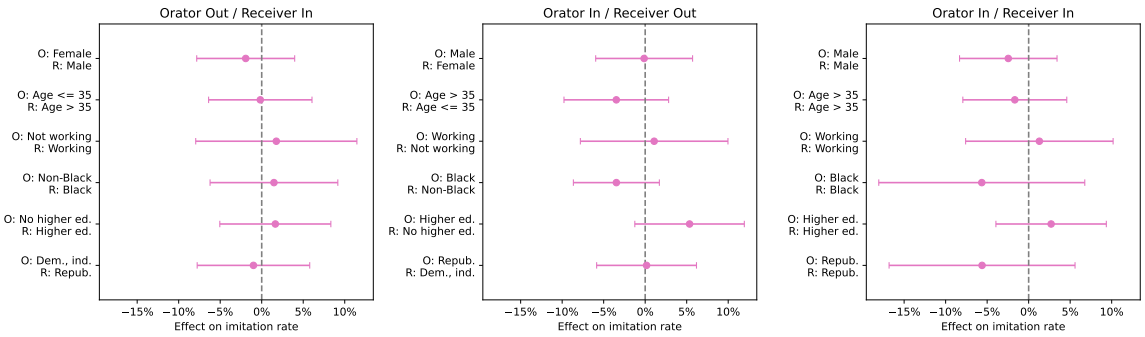


Figure H12: Effect of orator-receiver characteristics on imitation in *Transcript* treatment. *Notes:* See notes in Figure H11. *Transcript* sample is the corresponding Receiver experiment, *Choice Only* is pooled from all collections. Whiskers show 95% CIs with clustering at the orator and receiver levels.

the *Explanation* treatment (put differently, we augment column (6) of Table 1 with controls for orator characteristics, their interactions with richness, and their triple interactions with richness and *Explanation*). Results are shown in the left panel of Figure H13.

First, the Figure shows that the richness effect stays large and significant, with a 1 SD increase in richness raising imitation by 11.8 p.p. ($p = 0.02$). However, it also shows that orator characteristics do not strongly predict the imitation of richness. All interaction effects are insignificant at the 5% level, with the largest point estimate for Republican orators, for whom the richness effect is 5.0 p.p. lower ($p = 0.06$). Effects are also insignificant for orator confidence, optimality on other tasks, education, race, employment status, age, and gender.

The right panel shows corresponding results for receiver characteristics. The richness effect is again large and significant, while receiver characteristics also do not appear to significantly predict richness imitation. In particular, the interactions with receiver prior confidence and optimality on other tasks are small and insignificant. This suggests that observable characteristics, including financial reasoning ability as measured by the share of correct answers on other tasks, have only moderate effects on the tendency to anchor on richness. However, it should be stressed that this correlational analysis is not very well-powered to conclusively answer this question.

To leverage exogenous variation, Section I.8 explores whether these receiver characteristics predict imitation in the richness manipulation experiment of Section 4.4.2. It shows that lower confidence, and to a weaker extent lower financial reasoning ability, predict stronger anchoring on linguistic richness when arguments are held constant.

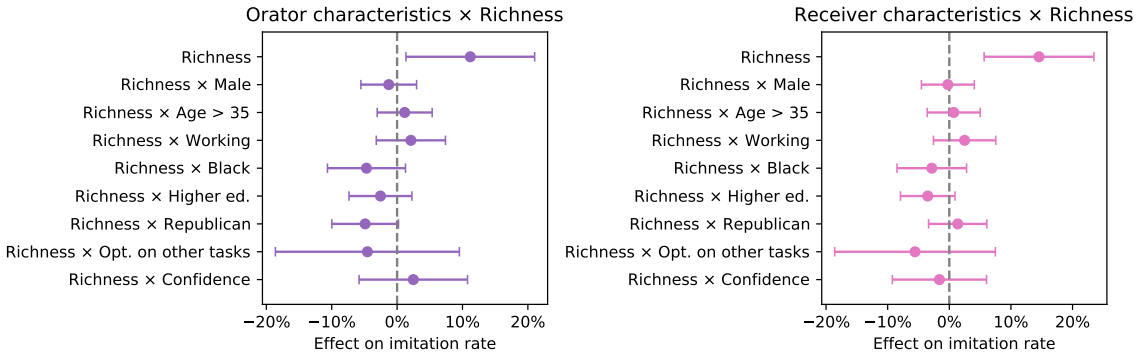


Figure H13: Effect of orator and receiver characteristics on richness imitation in *Explanation*. *Notes:* The figure reports the overall richness-imitation effect and its heterogeneity by orator characteristics (left panel) and receiver characteristics (right panel), estimated from regressions of imitation on *Explanation*, learning, richness, demographic characteristics, and all interactions between *Explanation*, richness, and the corresponding characteristics. Sample is restricted to learning and unlearning situations. Sample for *Explanation* is the main Receiver experiment, *Choice Only* is pooled from all collections. Whiskers show 95% confidence intervals with clustering at the orator and receiver levels.

I Richness manipulation experiment

This section summarizes additional details about the experiment exogenously manipulating linguistic richness.

I.1 Details on the generation of rich & sparse versions

Our experimental design aims to take each explanation given by a respondent and generate a rich and a sparse variant that (i) preserve the informational content, and (ii) remain as close as possible to the natural tone of respondents.

We address (i) by heavily emphasizing in our prompt that the informational content, and in particular "all arguments, facts or references" should be exactly preserved. An inspection of original explanations and generated versions shows this works quite well. The ex-post argument identification exercise described below confirms this finding.

Achieving (ii) is more challenging because LLMs are pre-trained on corpora of written text and fine-tuned to be maximally intelligent and helpful assistants. We therefore give the model examples, for each question, of 20 explanations randomly drawn from the 50 with the lowest richness rating (to avoid having only extremely sparse explanations), and the 5 with the highest richness rating (since these are longer).

We then prompt it to generate a maximally sparse and a maximally rich version by imitating the style of the examples while keeping the informational content constant. We instruct the LLM to pay close attention to specific features when imitating the sparse and rich examples. These features are informed by our analysis of the content of explanations and of the determinants of richness (cf. Section 4.4).

To generate the versions, we used GPT-4.1 (more specifically, `gpt-4.1-2025-04-14`), OpenAI's "flagship model for complex tasks" as of June 2025. Below, we reproduce the full prompt used to generate both versions at once:

> Survey respondents were asked the following financial literacy question: {question}
>
> They also recorded oral responses to explain their choice. A rich explanation is detailed, comprehensive, logically structured, nuanced, and tailored to the context. A sparse explanation is basic, narrow, unclear or disorganized, presents only surface-level understanding, lacks depth or specific details, and fails to relate clearly to the context. Richness is measured as an integer between 0 and 10, both inclusive. If the explanation provides no argument, it is

0. Richness should not depend on whether the respondent's answer or their premises are factually correct—only the quality of explanation matters.

To give some examples, here are twenty maximally sparse and five maximally rich explanations given by respondents:

{20 sparse examples}

{5 rich examples}

We now focus on a specific respondent, who chose the following answer: {answer}

The respondent also gave the following oral response to explain their choice: {speech}

Please produce two versions of the above explanation, one that is maximally sparse (richness score: 1/10) and one that is maximally rich (richness score: 9/10). Each version should exactly preserve the informational content of the original explanation. In particular, it should preserve all arguments, facts or references of the original explanation, and not add to them. Instead, your re-writings should focus on structure, coherence, presentation and wording of the explanation, and very closely imitate the corresponding examples given above.

Both versions should closely imitate the oral style of corresponding explanations. In particular, you should pay close attention to filler words ("uh", "um", "huh"), false starts ("So, uh, I think, maybe..."), repetitions and restarts ("I believe that, um, I believe that..."), rhetorical questions ("Why do they outperform?") and first-person framing ("I believe...") and meta-commentary ("This is an easy/difficult question", "To answer this question", "To maximize your chances of getting it right" etc.).

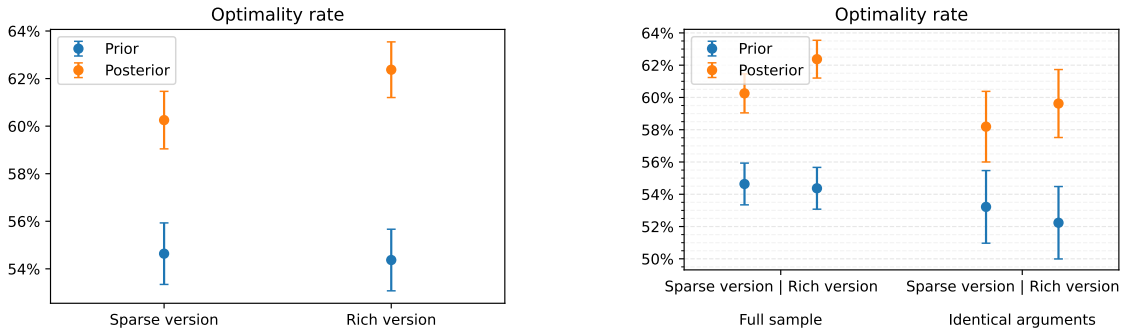
The maximally sparse version should not be a pure summarization of the explanation. When imitating the given examples, you should pay close attention to contradictions or self-corrections mid-sentence ("I think it's one—no wait, maybe two..."), vague or circular language, replacement of specific vocabulary with informal terms ("stuff", "thing"), fragmented or run-on sentences ("So I say one, because, like, I think..."), clauses connected without logical links ("I think, it has to be that, I mean clearly"), abrupt cutoffs or trailing thoughts ("So yeah, number one, because they..."), disconnected or illogical flow, introduction of arguments without relating to the overall context and over-reliance on the prompt.

The maximally rich version should not be a mere rewording of the explanation. When imitating the given examples, you should pay close attention to their clear logical sequence and explicit connectives ("because", "therefore", "for instance"), clarification of underlying assumptions and concepts, specific vocabulary and avoidance of vague words, illustrative examples and real-world analogies, contextual framing of arguments ("Let's say the interest rate is 2%..."), and structure designed to help someone else choose the correct answer rather

than just restating their own.

I.2 Effect of richness manipulation on optimality

Although it is not the main outcome of interest in this experiment, since both correct and incorrect explanations are rewritten, we examine the effects of the richness manipulation on optimality. Figure I14a plots the prior and posterior optimality rate by treatment. Posterior optimality increases by 2.1 p.p. ($p < 0.01$), an effect which is small but significant.



(a) Prior and posterior optimality

(b) Optimality with identical arguments

Figure I14: Effect of richness manipulation on optimality. *Notes:* See Notes for Figure 7. Whiskers show 95% CIs with clustering at the orator and receiver levels.

I.3 Richness of rich & sparse versions

We display the distribution of richness scores in Figure I15, which shows that the sparse versions are indeed sparser and the rich versions richer than the originals. Moreover, all distributions appear relatively well-behaved, with no extreme clustering at either end.[32]

We can characterize the effect of the richness manipulation depending on the accuracy of the orator's answer. When the orator is correct, original richness is 3.74 on average, against 3.38 for *Sparse version* ($p < 0.01$) and 6.74 for *Rich version* ($p < 0.01$). On the other hand, when the orator is incorrect, average richness is 2.40, 2.24 ($p < 0.01$) and 4.73 ($p < 0.01$), respectively.

Finally, we examine how each version's richness depends on the original's richness in Table I4. Column (1) shows that the sparse version dampens original richness but has a slight fixed boost, while column (3) shows the rich version amplifies original richness differences and also has a

[32]This well-behavedness is somewhat at odds with our prompt, which instructed the LLM to aim for a richness of 1/10 and 9/10, respectively. Experimentation shows LLMs are not very good at obtaining very specific richness scores, so that these indications were only meant to broadly increase richness differences.

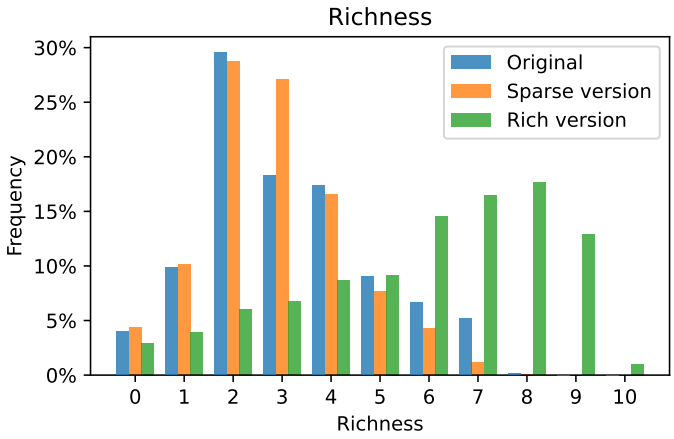


Figure I15: Distribution of richness of original, rich & sparse versions. *Notes:* Sample is the Orator experiment (466 orators, 6,910 obs.) and variants generated with an LLM.

much more sizable fixed richness boost. Columns 2 and 4 show that there is no strong effect of orator accuracy on this pattern for sparse versions, but that rich versions amplify richness heterogeneity less strongly for correct orators. Perhaps this is because correct explanations are often already comparatively rich, so that there is less room to increase it further.

Table I4: Effect of original richness on sparse and rich versions' richness

	Sparse version richness		Rich version richness	
	(1)	(2)	(3)	(4)
Intercept	0.657***	0.649***	2.212***	1.487***
	(0.024)	(0.033)	(0.067)	(0.070)
Orator correct		0.157***		1.571***
		(0.040)		(0.093)
Original richness	0.705***	0.662***	1.157***	1.353***
	(0.008)	(0.013)	(0.018)	(0.026)
Original richness × Orator correct		0.026*		-0.370***
		(0.014)		(0.025)
Observations	6910	6910	6910	6910
R^2	0.701	0.706	0.650	0.671

Notes: All richness measures are standardized using the mean and standard deviation of original explanations. Sample is the Orator experiment (466 orators, 6,910 obs.) and variants generated with an LLM. Standard errors are clustered at the orator level.

I.4 Text features of rich & sparse versions

To shed more light on what is driving this exogenous linguistic richness variation, we analyze the features of the generated versions in more depth. We therefore repeat the feature annotation exercise described in Section B.2 separately for each version, and display results in Figure I16.

We find that both rich and sparse versions contain more low-confidence markers than original explanations, meaning that even in rich versions the LLM communicates more extensively about the uncertainty contained in the initial explanation. Sparse explanations contain fewer high-confidence markers and even rich versions contain slightly fewer than the originals, perhaps because the LLM is slightly overcautious in wanting to preserve informational content.

Disfluencies are more common in sparse versions than in the original and rarer in rich versions, a pattern that applies to self-corrections, repetitions, false starts and especially to filled pauses. Indeed, originals contain 1.06 filled pauses (e.g., "uh" or "uhm") against only 0.17 in rich versions and 2.89 in sparse versions. Clearly, the LLM identifies these as prominent features of maximally sparse explanations. For most other features, we document that they are less frequent in sparse versions and more frequent in rich ones.

Taken together, these results show that the linguistic richness manipulation affects the presence of correlates of richness and of language markers that contribute to it. We therefore conclude that it achieved its aim in an effective, broad and balanced way.

I.5 Arguments of rich & sparse versions

As an additional verification, we repeat the human argument annotation experiment of Sections 4.4.3 and J with GPT-4. For each version, we ask it to identify which of our manually identified arguments are present in each version of each explanation.

Across all rounds and all arguments (including "Other" and "Irrelevant"), the correlation between arguments identified in the original explanation and in the sparse version is 0.68, with a Cohen's κ of 0.67 indicating "substantial agreement". Similarly, the correlation between arguments in the original and the rich version is 0.66, and a Cohen's κ of 0.66 again reflecting "substantial agreement". Agreement between the two versions, the condition which ensures the validity of our experimental design, is even higher with a correlation of 0.76, and a Cohen's κ at 0.76 showing we have "substantial agreement" and are very close to "almost perfect agreement".

When an argument is not identified in the rich version, it is also not identified in the sparse version in 98% of cases. However, when an argument is identified in the rich version, it is also

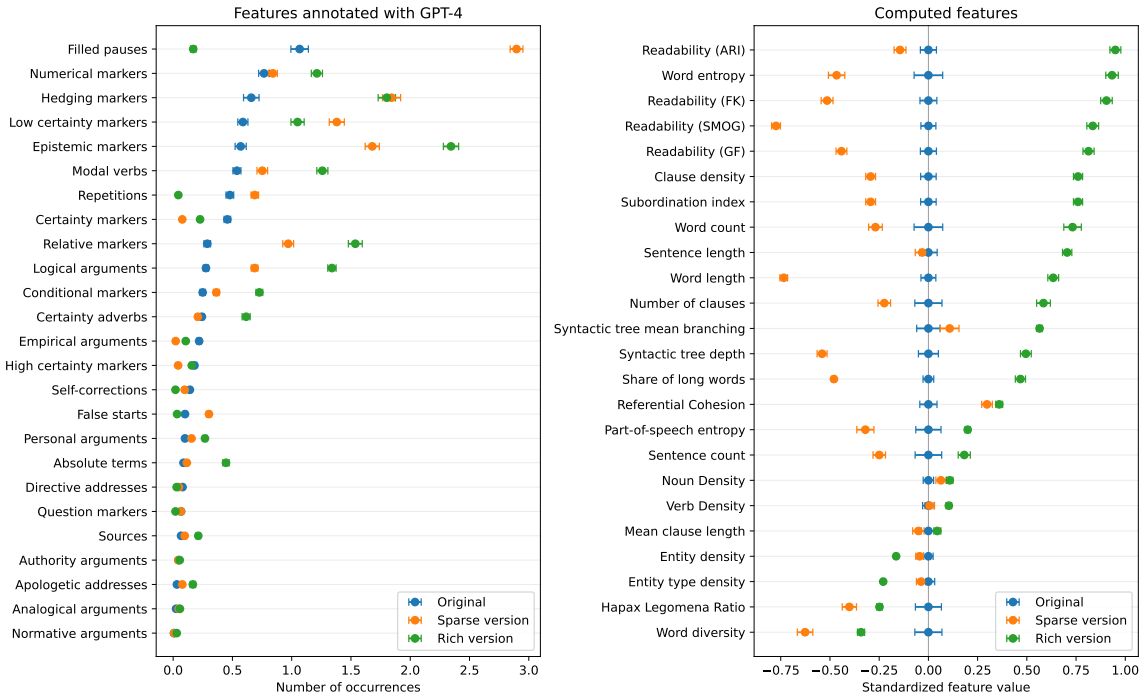


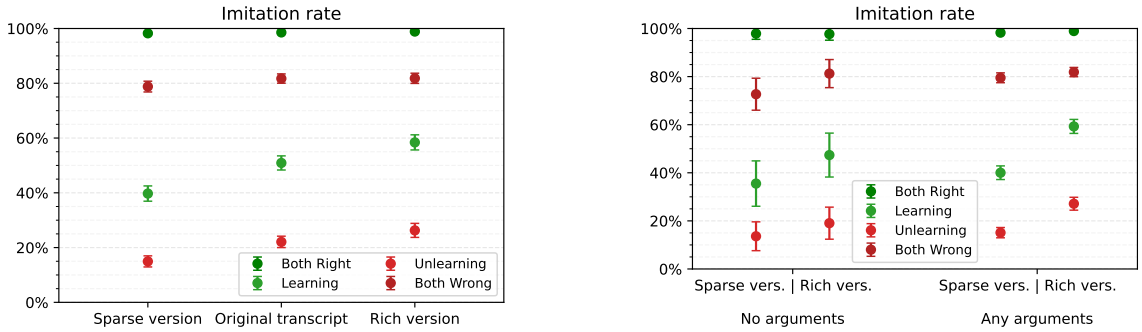
Figure I16: Text features of original, rich & sparse versions. *Notes:* See Table B1 for list of features. Sample is the Orator experiment (466 orators, 6,910 obs.) and variants generated with an LLM.

identified in the sparse version 73% of the time. Indeed, sparse versions contain 1.09 arguments on average, slightly less than the rich versions which contain 1.32, a comparatively small but significant difference ($p < 0.01$).

In conclusion, the richness manipulation appears very effective at preserving the informational content of explanations, though it is not perfect. It should be noted that assessing the presence of specific arguments is inherently noisy, so that even preserving informational content perfectly is likely to lead to some disagreement in the ex-post argument identification. Moreover, since rich versions contain more words and LLMs work on tokenized words, there is a higher chance that the LLM picks up on an instance of an argument in richer transcripts, so that the small differences we do observe could reflect this algorithmic bias.

Nonetheless, as a conservative robustness check, as we do with human annotators in Section 4.4.3, we repeat our main analysis on the explanations for which there is complete agreement on the informational content of both versions. For 62.6% of the sample, GPT-4 exactly agrees on the presence or absence of each of the arguments, including “Other” and “Irrelevant”, in both versions. This fraction is quite high given the noisy nature of argument tagging, with

another substantial fraction of explanations agreeing on all but one or two arguments. Among this sample with perfect agreement, imitation is 7.1 p.p. higher in *Rich version* than in *Sparse version* ($p < 0.01$). Moreover, it is 18.6 p.p. higher in learning situations ($p < 0.01$), 8.6 p.p. higher in unlearning situations ($p < 0.01$), 2.9 p.p. higher when both are wrong ($p = 0.05$) and 0.6 p.p. higher when both are right ($p = 0.22$). Since these effects are significant and very close to those in the full sample, we conclude that our experimental findings are not driven by changes in the informational content of explanations.



(a) Imitation of original, rich & sparse versions

(b) Imitation by argument presence

Figure I17: Effect of richness manipulation on imitation, relative to *Transcript* and by presence of arguments. *Notes:* Sample is the Richness receiver survey (972 receivers, 14,580 obs.). Whiskers show 95% CIs with clustering at the orator and receiver levels.

I.6 Heterogeneity by presence of arguments

Another check we can implement is to test whether the richness manipulation has effects that differ based on the presence of any argument. Large differences in the effect of the treatment between explanations with and without arguments suggest that variation in informational content plays some role in the results. For explanations without any arguments, imitation is 57.0% and 62.1% in *Sparse version* and *Rich version*, respectively, against 65.0% and 72.8% for explanations with arguments. This 2.7 p.p. difference in effect sizes is not very large and not statistically significant ($p = 0.37$). Figure I17b plots the imitation rates for each condition separately for each configuration, which also all have small and insignificant differences in effect sizes. However, it should be noted that this test is not very powerful because explanations without any arguments are quite rare, making up only 7.9% of the sample.

I.7 Comparison with original *Transcript* treatment

Although results should be interpreted more carefully since they are not part of the same experimental design, it is nonetheless interesting to compare imitation of the edited versions with that of the original explanations in the *Transcript* condition. Overall, it sits between the sparse and rich versions, since its imitation rate is 69.5%, above the 64.3% of *Sparse version* ($p < 0.01$) and below the 71.9% of *Rich version* ($p < 0.01$). Figure I17a documents that, also within each situation, original explanations are in-between the sparse and rich versions. Since the richness of originals also sits in-between that of the versions, these findings further point toward an effect of richness on imitation.

I.8 Richness manipulation and receiver characteristics

Section H.5 documented that, when using cross-sectional variation across tasks and respondents, receiver characteristics did not appear to predict imitation of richness. We can now ask the same question while leveraging the exogenous variation of this experiment. Figure I18 plots the results of a regression of imitation on the interaction between the *Rich version* treatment and receiver characteristics. The effect of the treatment itself remains large and significant at 37.0 p.p. ($p < 0.01$). Lower confidence has a strong negative effect: going from 0% to 100% receiver prior confidence decreases imitation by 14.4 p.p. Similarly, going from 0% to 100% correct prior answers on other tasks reduces imitation by 9.5 p.p., though this is not significant ($p = 0.11$). Other demographic characteristics do not appear to have a significant effect. Hence, less confident, and perhaps less financially able, respondents tend to more strongly anchor on linguistic richness. This suggests that, at least to some extent, richness serves as a shortcut to evaluate the correctness of an argument, as we discuss at the end of Section 4.4.3.

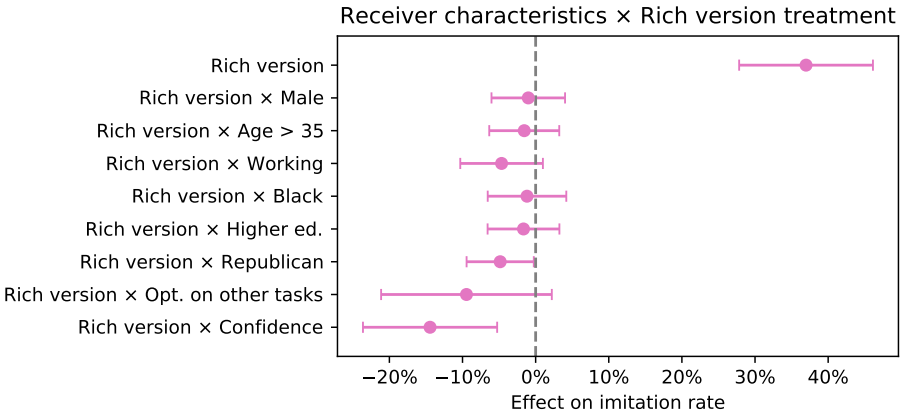


Figure I18: Effect of receiver characteristics on imitation in Richness manipulation experiment. *Notes:* Overall richness-imitation effect and its heterogeneity by receiver characteristics, estimated from a regression of imitation on learning, *Rich version*, receiver characteristics and their interactions with *Rich version*. Sample is the Richness receiver survey (972 receivers, 14,580 obs.), restricted to learning and unlearning situations. Whiskers show 95% CIs with clustering at the orator and receiver levels.

J Argument annotation experiment

J.1 Design & logistics

In the instructions, participants were told that previous respondents had answered financial reasoning questions and recorded explanations for future participants. These explanations had been transcribed and edited. The task was to read 10 such explanations for a single task and, for each one, select all arguments from a provided list that were present in the explanation. Participants were explicitly instructed that an explanation could contain multiple arguments, one argument, or none of the listed arguments.

The instructions emphasized that arguments should be identified based on their substance rather than their wording. Participants were told to select an argument when the explanation expressed the same reasoning, even if it used different words, and not to select an argument merely because it contained similar vocabulary. The list of arguments in each task is as in Table F1, with the description being presented to participants. Figure M30 shows the annotation screen.

Participants received a base reward of $5. To incentivize careful coding, receivers had a 10% chance of being eligible for a $10 bonus with "beauty contest" incentives. If the receiver was eligible for the bonus, one of their annotated explanations and one of its arguments were randomly selected, and the participant entered a lottery with winning probability equal to the share of other participants whose classification matched theirs.

The experiment was run in May 2026 using a U.S. sample on Prolific. Median completion time was 16 minutes. Out of 1,990 respondents, 1,967 gave valid data. The quota system distributed explanations across annotators as uniformly as possible, while ensuring that no respondent saw both the rich and the sparse version of the same explanation. In line with our pre-registration, we discard the first two out of ten annotations so annotators could get used to the argument set. We then select exactly one annotation per speech-version, and subset to speeches for which each version has an annotation. The final sample hence contains one annotation for the rich and sparse versions of 6,883 explanations, representing 99.6% of our original sample of 6,910.[33]

[33]The discarded annotations, coming notably from the first two rounds, yield multiple annotations per version-pair, which were used to implement the "beauty contest" incentives.

J.2 Optimality effect in subsample with identical arguments

As in I.2, we analyze the effect of the richness manipulation on optimality in the subset with identical arguments. Since both correct and incorrect explanations are edited into rich and sparse versions, one should not expect the richness manipulation to increase optimality. In the full sample, it indeed did not produce a significant effect. In the subsample with exactly identical arguments, the effect on optimality is again small at 1.4 p.p. and insignificant ($p = 0.26$). Figure I14b shows the corresponding prior and posterior optimality rates.

J.3 Comparison of human and LLM annotation

We investigate agreement between the human annotation in this experiment and the annotation performed with GPT-4 in Section I.5. Inter-rater reliability is high. If GPT-4 identifies a specific argument, there is a 66% chance the human annotator does so as well. If GPT-4 does not identify an argument, there is a 92% chance the human does not either. Cohen's κ between the two is 0.59, indicating "moderate agreement" (and very close to 0.6 i.e. "substantial agreement"). Inter-rater reliability is again even higher for argument categories, with positive and negative agreement rates of 88% and 46% for "any argument", 83% and 88% for "any sound argument", 77% and 83% for "any fallacious argument", 18% and 98% for "any irrelevant argument". The latter is a notable exception, and reflects the fact that humans classify 13.5% of explanations as containing an irrelevant argument, while GPT-4 is much more generous at only 3.9%. We view these benchmarks as confirming the overall quality of the respondents' annotation. We find no evidence of a difference in treatment effects between the human and GPT-4 identical-argument subsamples ($p = 0.26$).

Finally, we investigate how richness affects the argument annotations of humans and GPT-4. Table J5 regresses the number of arguments identified by humans and GPT-4 on the version. Column (1) shows human annotators from Prolific identify 0.997 arguments on average in sparse versions, and 0.275 arguments more in rich versions ($p < 0.01$). Conversely, column (2) shows GPT-4 identifies 0.955 arguments in sparse versions and 0.188 ($p < 0.01$) arguments more in rich versions. Therefore, as is shown in column (3) where the outcome is the difference in the number of arguments identified by respondents and GPT-4, humans identify 0.042 arguments more in sparse versions ($p < 0.01$), and an additional 0.087 more in rich versions ($p < 0.01$). Humans are therefore more strongly influenced by richness than GPT-4 when identifying arguments.

Table J5: Number of arguments identified by humans and GPT-4 in rich and sparse versions

	# arguments (humans) (1)	# arguments (GPT-4) (2)	Δ # arguments (3)
Intercept	0.997***	0.955***	0.042***
	(0.018)	(0.017)	(0.014)
Rich version	0.275***	0.188***	0.087***
	(0.013)	(0.007)	(0.014)
Observations	13766	13766	13766
R^2	0.021	0.013	0.003

Notes: Δ is the difference in the number of arguments annotated by humans and GPT-4. Sample is a subset of the richness manipulation experiment (1,963 receivers, 13,766 obs.). Standard errors are clustered at the orator (i.e., explanation) and annotator levels.

J.4 Channels of richness imitation

Richness could increase imitation through two channels. First, it could increase the perceived quality and number of arguments in an explanation. Second, it could lead receivers to imitate more even holding the arguments fixed, perhaps because linguistic richness acts as a signal about the speaker's competence.

We can cautiously approach this question by investigating how the effect of the linguistic richness manipulation evolves as we add finer and finer controls for explanations' argumentative content. Column (1) of Table J6 regresses imitation in learning and unlearning situations on the rich version, showing our main 15.4 p.p. treatment effect ($p < 0.01$). Column (2) adds a control for the two versions having different perceived arguments in the annotation experiment, recovering the 13.4 p.p. effect in the agreement subset ($p < 0.01$), which represents 87% of the total effect. Column (3) controls for the number of arguments identified, showing that the rich version treatment still induces a 14.2 p.p. effect ($p < 0.01$), i.e. 92% of the total. Finally, column (4) adds a fixed effect for each of the identified arguments, after which the richness manipulation nonetheless increases imitation by 13.7 p.p. ($p < 0.01$), making up 89% of the total effect.

These findings are compatible with the view that richness affects both the perceived informational content of explanations and imitation conditional on informational content, with a large share of its total effect probably driven by the latter. However, it should be stressed that this is only very imperfect evidence, as it is very difficult to proxy the informational content of explana-

tions as perceived by each listener. Moreover, here, variation in perceived argumentative content is driven by residual differences across versions written by the LLM, which is presumably very different from variations in natural settings.

In conclusion, the annotation experiment documents that richness robustly affects imitation both via the perceived informational content of explanations and via linguistic richness conditional on the informational content, both of which drive social learning.

Table J6: Effect of richness manipulation conditional on argument content of explanations

	Dependent variable: Imitation in Learning & Unlearning			
	(1)	(2)	(3)	(4)
Intercept	0.279***	0.292***	0.224***	0.268***
	(0.011)	(0.016)	(0.012)	(0.017)
Rich version	0.154***	0.134***	0.142***	0.137***
	(0.013)	(0.021)	(0.013)	(0.013)
Different arguments		-0.020		
		(0.017)		
Rich version × Different arguments		0.030		
		(0.025)		
# arguments			0.056***	
			(0.007)	
Argument FEs				✓
Observations	6036	6036	6036	6036
R^2	0.026	0.026	0.037	0.131

Notes: Rich version is a dummy for the rich version treatment. *Different arguments* is a dummy for whether the rich and sparse versions contain different arguments as annotated by humans. Sample is a subset of the richness manipulation experiment (1,963 receivers, 13,766 obs.) restricted to learning and unlearning situations. Standard errors are clustered at the orator and receiver levels.

K Binary Balls-and-urns experiment

K.1 Design and logistics

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

Respondents in both studies received a base reward of $2 and could win a bonus of $10 following the same incentive structure as in the main experiment. Median completion times were 11 and 10 minutes in the Orator and Receiver experiments respectively. Both experiments were run in May 2026 using a U.S. sample on Prolific. The Orator experiment was run for a total of 552 respondents, of whom 483 gave valid data. The Receiver experiment was run for 2,006 respondents, of whom 1,914 gave valid data.

In line with our pre-registration, we drop respondents who self-reported searching for answers online or using external tools. To curb the use of AI assistants,[34] we asked receivers whether the orator used an automatic voice generator or seemed to be reading a script. As specified in our pre-registration, we then exclude all orators for whom a majority of receivers agree: these cases represent 4.4% of valid recordings.

Due to the binary elicitation, we expected the optimality rate to be high. This implies learning and unlearning situations should be very rare under random matching. To increase power, we therefore stratify the orator assignment based on the receiver's prior answer. If a receiver's prior answer is correct, we randomly assign them to a correct or incorrect orator with equal probability. If it is incorrect, we assign them an orator without stratification. In our raw sample, receiver and orator are both right in 63.2% and both wrong in 9.0% of observations, while learning and unlearning situations occur 9.5% and 18.3% of the time. We then de-stratify our sample using

[34] In all experiments, the question text is not selectable and therefore difficult to copy-and-paste into a search engine or AI assistant. However, between the financial reasoning experiments in December 2023 and additional experiments in April 2025 and May 2026, multimodal AI assistants that can analyze screenshots have become much more prevalent. We believe this explains the noticeable rise in orators that appear to read a script from an AI assistant, and motivates our introduction of additional checks.

expected frequencies under random matching, with an optimality rate for each configuration estimated by pooling orators and receiver priors. All summary statistics, regressions and figures in the main body and in the following sections are re-weighted to this population frequency.

K.2 Optimality and imitation

Answers are correct if they say which bag is more likely in line with the Bayesian posterior. Figure 8 displays optimality rates and imitation rates for the balls-and-urns task. Prior optimality stands at 83.6% and 80.3% in *Choice Only* and *Explanation*, respectively ($p = 0.03$), which, as expected, is relatively high. This reflects the binary elicitation scheme, which considerably simplifies the task. In particular, when the ball draw signal goes in the same direction as the prior, the task becomes essentially trivial and optimality rates are very close to 100%. As in our main experiment, both treatments help listeners select correct answers and significantly increase their optimality rates by 1.9 p.p. ($p = 0.01$) and 3.7 p.p. ($p < 0.01$) respectively. *Explanation* hence increases the posterior optimality rate by 1.8 p.p. more ($p = 0.09$) than *Choice Only*.

Figure 8 also plots imitation in learning and unlearning situations. The share of receivers imitating the orator is 5.5% and 5.2% for *Choice Only* and *Explanation* in unlearning opportunities ($p = 0.93$). In learning opportunities, it jumps from 27.0% to 51.6%, a highly significant treatment effect ($p < 0.01$). As in our main experiments, explanations therefore increase imitation in learning but not unlearning situations.

K.3 Imitation of arguments and richness

We now analyze the substantive content of explanations in this experiment. Our binary elicitation implies that, while under *Choice Only* receivers only see which ball was deemed more likely by the orator, under *Explanation* the orator could also state a numerical posterior belief. As pre-registered, we use GPT-4 to identify which explanations state a numerical posterior. We specifically instruct it to only retain posterior belief statements, not pure repetition of numbers from the question. 10.9% of explanations explicitly state a numerical posterior belief, a relatively low share. Moreover, 43.1% and 5.9% of these state a posterior belief corresponding to the diagnosticity or the prior respectively. Our experiment therefore documents that, in balls-and-urns tasks with binary elicitation, subjects rarely communicate using explicit posterior beliefs.

As in our main experiment, we identify the main arguments and use GPT-4 to annotate them in individual explanations. Figure K19 plots the frequency of each argument in correct and in-

correct explanations in its left panels. Base-rate neglect is by far the most common argument, appearing in 74.4% of incorrect and 49.6% of correct explanations. In these explanations, orators focus primarily on the ball composition of the bags while ignoring the card-deck prior. Signal neglect, i.e. reliance on the card-deck prior while ignoring the ball signal, is less common but still more frequent among incorrect explanations (11.6% versus 8.1%). This reasoning-based evidence confirms previous research on the topic (Bordalo et al., 2025). We call the argument that points out when both prior and signal favor the same bag without explicit calculations *Intuitive Confirming*: it is naturally almost absent among incorrect explanations (1.2%) but common among correct ones (23.7%). Other intuitive but non-calculative arguments appear at moderate rates in both groups (5.8% and 9.6%). Explicit probabilistic calculations are rare overall. Correct mathematical reasoning appears in only 4.8% of correct explanations and never among incorrect ones, whereas wrong mathematical reasoning occurs in 4.7% and 4.3%, respectively. Despite our large data collection, the right panel shows we are not sufficiently powered to analyze the effects of individual arguments on imitation. Point estimates suggest correct mathematical reasoning has a strong positive effect on imitation, while base-rate neglect has a negative one.

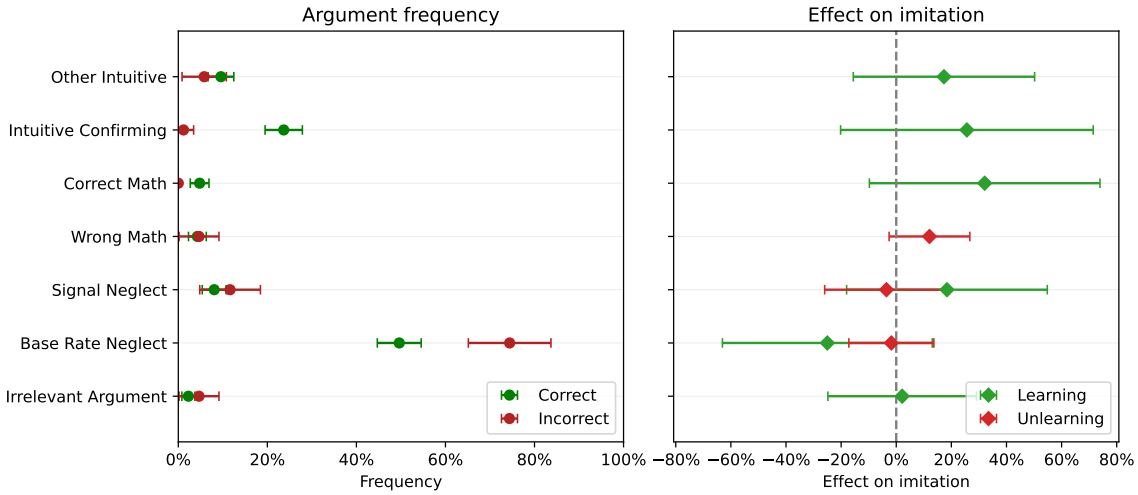


Figure K19: Argument frequencies and effects on imitation in the binary Balls-and-urns experiment. *Notes:* See notes for Figure 4. Sample is the Balls-and-urns receiver experiment (1,914 receivers and obs.). Whiskers show 95% CIs with clustering at the orator level.

Finally, we annotate the richness of explanations on a scale of 0-10. Correct explanations are 0.54 SD richer than incorrect explanations ($p < 0.01$): we thus document a significant richness gap in line with the main experiment. Table K7 then analyzes the effect of richness on imitation

in learning and unlearning situations. Column (1) shows that there is a significant effect of *Explanation* on imitation only in learning situations, reflecting the main reduced-form finding. Column (4) shows that a 1 SD increase in richness increases imitation by 7.4 p.p. ($p < 0.01$). In turn, this means richness explains 22.0% of the learning asymmetry in the binary Balls-and-urns experiment. Other columns confirm this pattern weakens somewhat but remains qualitatively similar when controlling for the arguments used in each explanation, as well as receiver and orator characteristics. In conclusion, as in the main experiment, in the binary Balls-and-urns task respondents tend to imitate richer explanations more without imitating sparse explanations less, and correct explanations are richer on average, which drives their aggregate learning effect.

Table K7: Contribution of richness to learning in binary Balls-and-urns experiment

	Dependent variable: Imitation in Learning & Unlearning				
	(1)	(2)	(3)	(4)	(5)
Explanation	0.024	-0.043	0.029	0.053	0.004
	(0.032)	(0.100)	(0.034)	(0.047)	(0.105)
Learning	0.222***	0.233***	0.228***	0.224***	0.230***
	(0.058)	(0.060)	(0.059)	(0.061)	(0.064)
Explanation × Learning	0.196**	0.193**	0.190**	0.153*	0.152*
	(0.082)	(0.085)	(0.082)	(0.091)	(0.092)
Orator confidence			-0.027		-0.034
			(0.025)		(0.026)
Explanation × Orator confidence			0.033		0.036
			(0.044)		(0.042)
Richness				-0.002	0.018
				(0.021)	(0.029)
Explanation × Richness				0.074**	0.047
				(0.034)	(0.039)
Orator controls		✓			✓
Correct vs incorrect gap			0.30**	0.54***	
Gap × Imitation			0.010	0.040	
Share explained	0%	2%	3%	22%	22%
Observations	532	532	532	532	532
R^2	0.207	0.233	0.208	0.227	0.247

Notes: See notes for Table 1. As there was only one task in this experiment, orator controls do not include optimality on other tasks. Sample is the Balls-and-urns receiver experiment (1,914 receivers and obs.). Standard errors are clustered at the orator level.

L Continuous Balls-and-urns experiment

L.1 Logistics

The continuous balls-and-urns experiment resembles the binary one. We use priors of either 40% or 60%, and a diagnosticity held constant at 80%. However, the participant's task is to state the *likelihood* with which they think the blue-majority bag was selected. Beliefs are elicited using a text box. The continuous elicitation follows the behavioral literature on this task (Edwards, 1968; Bordalo et al., 2025).

Respondents in both studies received a base reward of $2 for completing the study. Participants could win a bonus of $10 following the same incentive structure as in the main experiment, defining an answer as correct if it is within ± 3 p.p. of the Bayesian posterior. Median completion times were 13 minutes in both experiments. Both experiments were run in April 2025 using a U.S. sample on Prolific. The Orator experiment was run for a total of 554 respondents, of whom 474 gave valid data. The Receiver experiment was run for 1,975 respondents, of whom 1,826 gave valid data. In line with our pre-registration, we drop respondents who self-reported using external tools. We further drop orators that a majority of receivers believed to have used an AI assistant: these cases represent 8.3% of valid recordings.

In light of prior research, we expected the share of respondents selecting a correct response to be low (Bordalo et al., 2025). This implies learning and unlearning situations should be very rare. As in the binary experiment, we therefore stratify the orator assignment based on the receiver's prior answer. If a receiver's prior answer is incorrect, we randomly assign them to a correct or incorrect orator with equal probability. If it is correct, we assign them an orator without stratification. In our raw sample, receiver and orator are both wrong in 40.6% and both right in 3.0% of observations, while learning and unlearning situations occur 43.0% and 13.5% of the time. We then de-stratify our sample using expected frequencies under random matching with an optimality rate of 16.7%, estimated by pooling orators and receiver priors. All summary statistics, regressions and figures are re-weighted to this population frequency.

L.2 Optimality

In line with the instructions and incentives given to survey participants, we define a correct answer as a belief within ± 3 p.p. of the Bayesian posterior. Figure L20 displays optimality rates for the balls-and-urns task. Prior optimality stands at 16.2% and 16.9% in *Choice Only* and *Ex-*

planation, respectively ($p = 0.70$), which, as expected, is considerably lower than the optimality rates around 50% seen in the main experiment. As before, both treatments help listeners select correct answers and significantly increase their optimality rates by 3.9 p.p. ($p < 0.01$) and 3.3 p.p. ($p = 0.01$) respectively. However, the difference in posterior optimality between the two conditions is very small ($p = 0.97$).

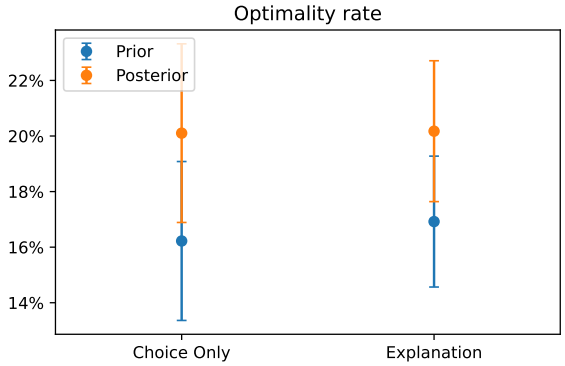


Figure L20: Effect of *Explanation* on optimality in Balls-and-urns experiment. *Notes:* Share of correct receivers before and after exposure to the orator's choice or explanation. An answer is defined as correct if it is within ±3 p.p. of the Bayesian posterior. Sample is the Balls-and-urns receiver experiment (1,826 receivers and obs.). Whiskers show 95% CIs with clustering at the orator level.

L.3 Imitation

In this abstract balls-and-urns setting, choices are continuous probabilities between 0% and 100%, so that imitation is harder to define and quantify. For completeness, we have therefore pre-registered four relevant measures of imitation:

1. *No imitation*: the receiver reports the same posterior belief as their prior belief
2. *Full imitation*: the receiver reports the same posterior belief as the orator
3. *Continuous imitation in [-1,2]*: to measure imitation continuously, we leverage a Grether (1980) decomposition to represent beliefs as log-odds in $\mathbb{R}$,[35] and then define continuous imitation as the interpolation weight placed on the orator's belief:

$$\text{Continuous imitation} = \frac{\text{logit(Receiver posterior belief)} - \text{logit(Receiver prior belief)}}{\text{logit(Orator belief)} - \text{logit(Receiver prior belief)}}$$

[35]Observations with beliefs of 0% or 100% are dropped from all analyses using continuous imitation.

No imitation yields a continuous imitation of 0, and full imitation yields 1. It can be negative if the receiver updates in the opposite direction, and larger than 1 if they go beyond the orator's belief. To limit the influence of extreme outliers due to beliefs close to 0% or 100%, we winsorize continuous imitation to $[-1, 2]$.

4. *Continuous imitation in [0,1]*: for completeness, we also report results for continuous imitation winsorized to $[0, 1]$.

Across all situations, the share of receivers who do not imitate at all is 50.9%, while 25.5% imitate fully. These two sets have a small overlap, as in 16.5% of cases the receiver has the exact same prior belief as the orator. These confirming signals are typically one of the salient modes of the problem, e.g., the prior or diagnosticity. On the other hand, 38.3% of receivers pick a posterior belief that is neither their own prior belief nor the orator's, with two thirds of these picking answers that are strictly between the two.

Figure L21 shows imitation in learning and unlearning situations. The share of receivers not imitating is close to 40% in learning for both treatments ($p = 0.89$) and closer to 50% for unlearning ($p = 0.90$). Full imitation rates are much lower, around 10%; they increase from 8.4% in *Choice Only* to 13.3% in *Explanation* in learning ($p = 0.03$), and from 7.4% to 11.9% in unlearning ($p = 0.25$). Continuous imitation winsorized to $[-1, 2]$ is about 40% in learning and 20% in unlearning situations, with insignificant differences between treatments ($p = 0.18$, $p = 0.78$). Findings are qualitatively similar for winsorization to $[0, 1]$ ($p = 0.10$, $p = 0.56$).

We further examine the full distribution of continuous imitation, as is done in Figure L22. The left panel shows the distribution pooling treatments, which displays strong modes at 0 and 1, corresponding to no and full imitation. There is a weaker mode around 0.5, suggesting that many receivers update somewhere roughly in the middle of their own prior belief and the orator's belief.[36] Imitation outside of these bounds is rare, as 12.7% of observations are outside of [0,1], and only 4.5% outside of [-1,2]. Splitting by treatments, the right panel reveals that the difference between *Choice Only* and *Explanation* appears small and unlikely to be statistically meaningful.

In conclusion, we only find very limited evidence of a treatment effect of *Explanation* relative to *Choice Only* in this continuous Balls-and-urns task. However, despite our oversampling strategy designed to maximize power, standard errors remain relatively large, so that this is not a very

[36]Note that, with integer percentage beliefs, it is impossible to achieve an imitation of exactly 0.5 or a similar round value in log-odds space.

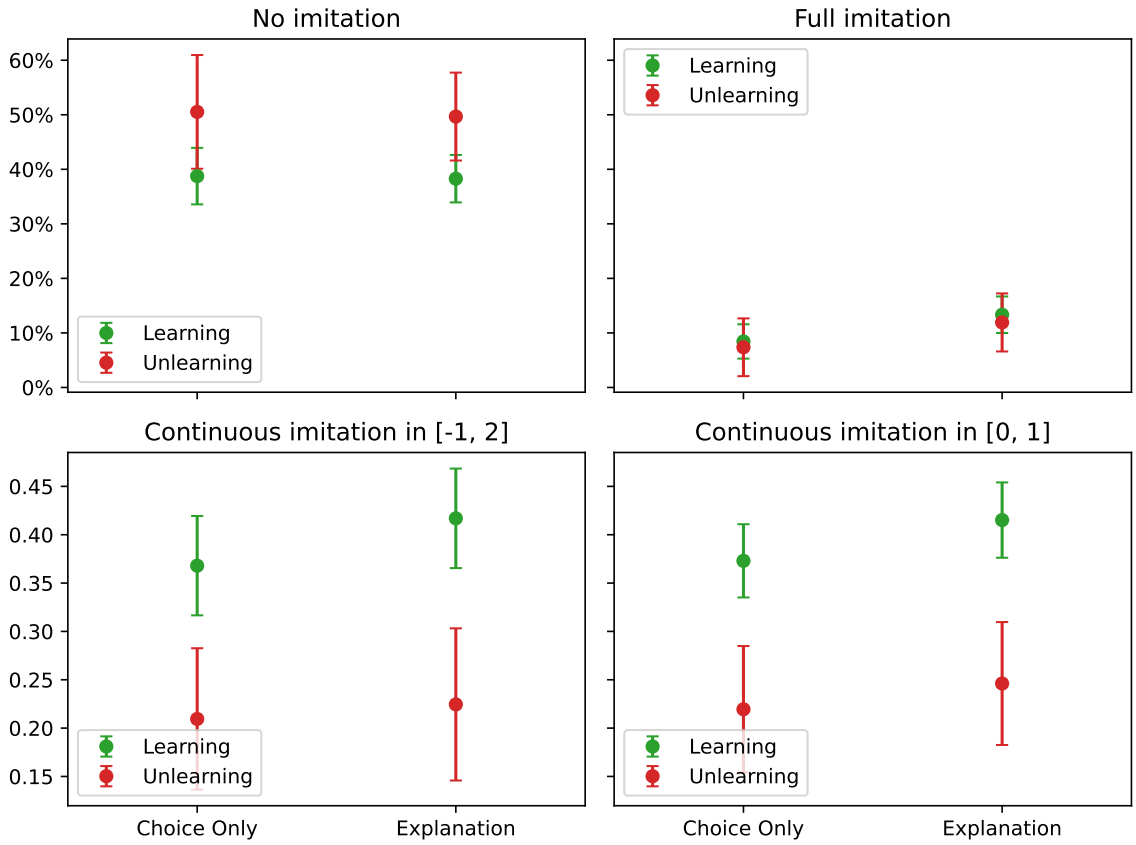


Figure L21: Effect of *Explanation* on imitation in Balls-and-urns experiment. *Notes:* See main text for definitions of imitation measures. Sample is the Balls-and-urns receiver experiment (1,826 receivers and obs.). Whiskers show 95% CIs with clustering at the orator level.

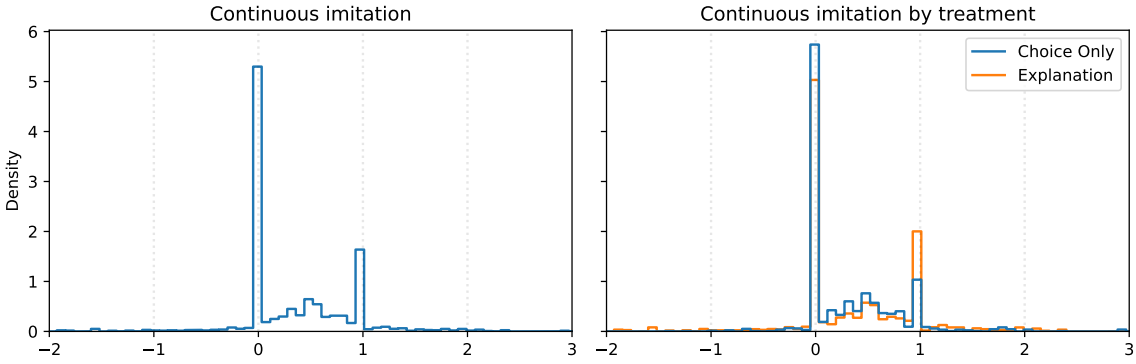


Figure L22: Distribution of continuous imitation in Balls-and-urns experiment. *Notes:* Left panel shows histogram of continuous imitation across treatments. Right panel shows histogram of continuous imitation in *Choice Only* and *Explanation*. The x-axis is cut to [-2,3], leaving 2.4% of observations outside of axis limits. Sample is the Balls-and-urns receiver experiment. Whiskers show 95% CIs with clustering at the orator level.

precise null. The difficulty of the task, the complexity of social learning in this context, and the difficulty in defining optimality and imitation make aggregate improvement and learning hard to measure.

L.4 Imitation of richness

Annotating the richness of explanations for this experiment reveals a 0.75 SD difference ($p < 0.01$) between correct and incorrect explanations.

We examine whether richness predicts imitation. Table L8 replicates Column (6) of Table 1. We find that richness has a statistically insignificant effect on no imitation ($p = 0.21$). However, it has a statistically significant effect on full imitation, at +3.6 p.p. ($p = 0.05$), and on continuous imitation winsorized to $[-1, 2]$ and $[0, 1]$, at +0.057 ($p = 0.07$) and +0.043 ($p = 0.06$) respectively. We do find evidence that our main hypothesized mechanism, the effect of richness on imitation, holds here. Crucially however, the magnitude of this effect is considerably smaller than for financial reasoning tasks.

Table L8: Imitation of richness in Balls-and-urns experiment

	No imitation (1)	Full imitation (2)	Continuous imitation (3)	Cont. in [0,1] (4)
Explanation	-0.012	0.050	0.021	0.031
	(0.066)	(0.039)	(0.054)	(0.046)
Learning	-0.110*	0.013	0.172***	0.155***
	(0.062)	(0.033)	(0.048)	(0.039)
Explanation × Learning	0.032	-0.024	-0.009	-0.017
	(0.078)	(0.047)	(0.070)	(0.054)
Richness	-0.011	-0.003	-0.019	-0.003
	(0.023)	(0.010)	(0.020)	(0.014)
Explanation × Richness	-0.035	0.036*	0.057*	0.043*
	(0.028)	(0.019)	(0.031)	(0.023)
Observations	1031	1031	1026	1026
R^2	0.020	0.014	0.036	0.053

Notes: Sample is the Balls-and-urns receiver experiment, restricted to learning and unlearning situations. *Explanation* is a dummy for the *Explanation* treatment, *Learning* is a dummy for learning situations. *Continuous imitation* is winsorized to [-1,2], *Cont. in [0,1]* to [0,1]. See main text for definition of imitation measures. Standard errors are clustered at the orator level.

Section A shows how the contribution of richness to the learning effect of explanations depends on the frequency of (un)learning situations, the richness gap and the imitation of richness. In the continuous balls-and-urns task, the richness gap between correct and incorrect expla-

nations is 0.75 SD ($p < 0.01$) while a 1 SD increase in richness raises imitation by 3.6 p.p. (cf. column (2) of Table L8). This implies a predicted imitation effect of $0.75 \times 0.036 = 2.7$ p.p. Since the optimality rate stands at 16.7%, learning and unlearning situations occur in $2 \times 0.167 \times (1 - 0.167) = 27.8\%$ of matches. Hence, the implied effect on optimality is just $0.75 \times 0.036 \times 0.278 = 0.8$ p.p., well below the ±2.4 p.p. confidence interval in Figure L20.

It is also instructive to compare this with the other tasks. On the supply side, the richness gap is comparable with the 0.54 SD and 0.77 SD gaps identified in binary balls-and-urns and financial reasoning tasks respectively. However, on the demand side, a 1 SD increase in richness raises imitation by a much larger 7.4 p.p. (cf. column (4) of Table K7) and 10.6 p.p. (cf. column (6) of Table 1) respectively. This suggests that, in this instance, social learning fails because receivers do not strongly imitate richer explanations, which could reflect the difficulty of the task.

The null results in the balls-and-urns experiment are therefore arguably a consequence of the environment rather than a failure of explanations. Correct explanations are richer in all experiments. When the task is made sufficiently tractable, richness becomes just as predictive of imitation as in financial reasoning tasks.

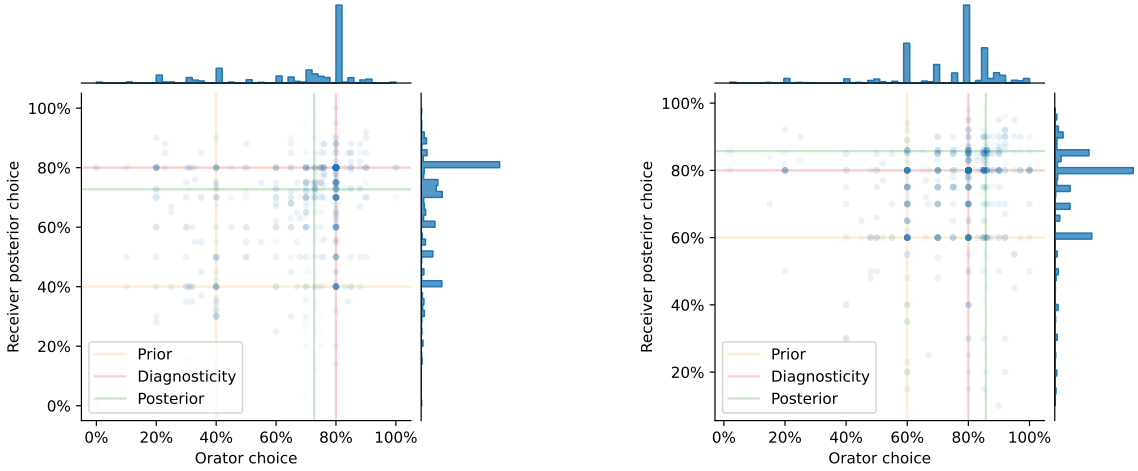
L.5 Imitation of salient modes

Previous research has shown that, when faced with abstract numerical problems like the balls-and-urns task, respondents rely on salient features to form an answer (Bordalo et al., 2025). In particular, salience causes neglect of other relevant features, leading to biased answers that cluster around specific modes. For a balls-and-urns inference task similar to ours, Bordalo et al. (2025) find that respondents very often answer the prior or 100%-prior, the diagnosticity or 100%-diagnosticity, or fall back to an ignorance default of 50%-50%. We obtain very similar modes in our data.

Our experiment allows us to study whether social learning about numerical problems also relies on salient features. We examine whether other modes, besides the posterior, are more likely to be imitated. We plot orator beliefs and receiver posterior beliefs in Figure L23. To increase power, we have grouped symmetric configurations based on the color of the ball draw, so that there are only two groups: configurations with a prior of 40% on the color of the ball draw, and configurations with a prior of 60%. The left panel shows the first case, where the posterior of 72.73% is below the diagnosticity of 80%, and the right panel shows the second case, where the posterior of 85.71% is above. The marginal distributions of beliefs exhibit strong modes at the

prior and the diagnosticity, and a weaker, more diffuse mode around the Bayesian posterior. The modes around 100%-prior, 100%-diagnosticity and 50%-50% are comparatively weaker.

The scatter plots, where the opacity of a dot reflects the number of overlapping observations, make clear that receiver posteriors also cluster on modes. Social learning is therefore largely based on switching from one mode to another, rather than smooth updating between likelihoods. For example, the dark blue dots along the "Diagnosticity" vertical line indicate a tendency by receivers to switch to this mode when faced with an orator that supports it. Salient features of a numerical problem therefore determine how subjects form responses to it, but also how they learn about it from others.



(a) Ball prior = 40%, Diagnosticity = 80%

(b) Ball prior = 60%, Diagnosticity = 80%

Figure L23: Distribution of orator and receiver beliefs in the continuous Balls-and-urns experiment. *Notes:* Symmetric configurations are grouped based on the color of the ball draw. Left panel shows configurations with a corresponding prior of 40%, right panel shows configurations with a corresponding prior of 60%. Histograms show marginal distribution, scatter plots show joint distribution, with opacity reflecting the number of observations. Sample is the continuous Balls-and-urns receiver experiment, pooling *Choice Only* and *Explanation* treatments.

In line with our pre-registration, we identify reasoning errors directly from estimated likelihoods. Among orators, 12.5% respond with the prior, 36.9% answer the diagnosticity inverted according to the signal (e.g., if the diagnosticity is 80% and the ball draw is blue, the belief that the ball is red would be 20%), 16.7% give an answer within ± 3 p.p. of the Bayesian posterior, and 34.0% state some other belief. Orators responding with 50%, the non-signal-flipped diagnosticity or the flipped prior are relatively rare. Due to our oversampling strategy, we have limited power to study these rarer modes, and therefore group them into the "Other" category.

Receiver prior belief modes are very similar. Importantly, receiver posterior belief modes are also very similar. In particular, the share of answers not falling in one of the three modes is 35.3% and 35.1% respectively ($p = 0.89$), confirming that social learning heavily relies on modes.

Table L9: Imitation of modes in Balls-and-urns experiment

	No imitation (1)	Full imitation (2)	Continuous imitation (3)	Cont. in [0,1] (4)
Intercept	0.454***	0.211***	0.410***	0.382***
	(0.038)	(0.036)	(0.041)	(0.032)
Diagnosticity · Signal	0.150***	0.260***	-0.048	-0.008
	(0.045)	(0.043)	(0.052)	(0.041)
Posterior ±3p.p.	-0.047	-0.061	-0.006	0.012
	(0.042)	(0.039)	(0.047)	(0.036)
Other	0.023	-0.125***	-0.136***	-0.096**
	(0.046)	(0.039)	(0.050)	(0.038)
Observations	1826	1826	1620	1620
R^2	0.023	0.153	0.011	0.014

Notes: Variables indicate orator belief mode: *Diagnosticity · Signal* is the diagnosticity inverted according to the signal draw, *Posterior ± 3 p.p.* are beliefs within ± 3 p.p. of the Bayesian posterior, *Other* are other answers. *Prior* is left-out. *Continuous imitation* is winsorized to [-1,2], *Cont. in [0,1]* to [0,1]. See main text for definition of imitation measures. Sample is the continuous Balls-and-urns receiver experiment, pooling *Choice Only* and *Explanation* treatments. Standard errors are clustered at the orator level.

Table L9 regresses imitation on the orator's choice mode, with "Prior" being the held-out mode as it represents the simplest heuristic. It shows that "Diagnosticity · Signal" is significantly more likely to be fully imitated, with a 26.0 p.p. increase, but also that it is more likely not to be imitated at all, with a somewhat smaller 15.0 p.p. increase. Correspondingly, there is no effect on continuous imitation. Presumably, this is because this mode reflects an (erroneous) mathematical understanding of the problem, and therefore calls for complete agreement or (slightly more rarely) disagreement, but not smooth updating somewhere in-between, so that the full- and no-imitation effects cancel out.

We can further observe that "Posterior ± 3 p.p." has no significant effect on imitation, indicating it is not much more or much less convincing than "Prior", which is an interesting finding. Finally, "Other" has a strong negative effect on full and continuous imitation, indicating receivers are less prone to update fully or even partially based on orators that are not clearly on one of the dominant modes.

So far, we have pooled *Choice Only* and *Explanation* to maximize power. For completeness,

Table L10 repeats the same analysis but distinguishes between *Choice Only* and *Explanation* by adding an interaction term for the latter. It finds little systematic evidence that *Explanation* changes the imitation of specific modes, in line with our previous findings for correct answers. However, it should be noted that our experiment is not very well-powered to study imitation of modes other than "Posterior ± 3 p.p." because of our stratification strategy, so that we can only detect relatively large effects.

Table L10: Effect of *Explanation* on imitation of salient modes in Balls-and-urns experiment

	No imitation (1)	Full imitation (2)	Continuous imitation (3)	Cont. in [0,1] (4)
Intercept	0.579***	0.246***	0.336***	0.313***
	(0.069)	(0.057)	(0.071)	(0.059)
Explanation	-0.201**	-0.056	0.113	0.106
	(0.093)	(0.073)	(0.102)	(0.083)
Diagnosticity · Signal	0.046	0.206***	-0.002	0.014
	(0.079)	(0.069)	(0.083)	(0.070)
Explanation × Diagnosticity · Signal	0.166	0.088	-0.069	-0.029
	(0.105)	(0.089)	(0.118)	(0.097)
Posterior ±3p.p.	-0.149**	-0.102	0.024	0.052
	(0.074)	(0.063)	(0.076)	(0.063)
Explanation × Posterior ±3p.p.	0.163	0.066	-0.040	-0.057
	(0.100)	(0.080)	(0.112)	(0.089)
Other	-0.062	-0.219***	-0.075	-0.069
	(0.080)	(0.059)	(0.081)	(0.066)
Explanation × Other	0.137	0.151*	-0.092	-0.038
	(0.108)	(0.077)	(0.118)	(0.093)
Observations	1826	1826	1620	1620
R^2	0.030	0.157	0.013	0.022

Notes: See notes for Table L9. *Explanation* is a dummy for the *Explanation* treatment. Sample is the continuous Balls-and-urns receiver experiment. Standard errors are clustered at the orator level.

Finally, we can examine the effect of the receiver's prior mode on imitation. Figure L24 plots the full and no-imitation rates by orator and receiver prior belief modes. Since two thirds of receivers either imitate fully or not at all, it reflects transitions for an important fraction of the sample. Receivers who answered the prior and saw an orator who answered the prior maintained their answer in 69% of cases, a figure that goes up to 94% for diagnosticity, indicating that these are socially stable modes. On the other hand, only 33% of receivers who gave a correct answer fully imitate the orator, though this effect is partly mechanical because this mode covers a range of ± 3 p.p. Indeed, a receiver who holds their belief while matched with an orator whose belief also falls within this range—but differs numerically—will not be classified as fully imitating,

even if no belief updating occurs. The corresponding no-imitation rate is relatively high at 53%.

Transitions between modes are lower but still relatively high, typically around 10%–15%, and these smaller differences are unlikely to be significant. No-imitation rates for these tasks are correspondingly lower, around 50%.

On the other hand, orators giving an answer in the "Other" category see full-imitation closer to 5–10%, lower than the other modes, and no-imitation around 60%, much higher than the others. Moreover, receivers whose prior belief was "Other" have no-imitation around 20%–30%, again considerably lower than for the other modes. They also appear more likely to imitate fully; in particular, they fully imitate the "Diagnosticity · Signal" mode in 28% of cases. In conclusion, non-modal answers are more likely to be deviated from, and less likely to be imitated.

Summing up, the continuous Balls-and-urns task appears much harder, reflected in a very low optimality rate, so that explanations do not increase aggregate optimality and show no robust effect across imitation measures. Analyzing explanations and modal answers shows reasoning errors are pervasive, making social learning effects smaller and harder to detect. A comparison with the binary Balls-and-urns experiment, where the elicitation made the task easier and explanations did improve learning, suggests that explanations lose their power when tasks become too abstract or too complex. Bordalo et al. (2025) show that changing the framing of this task, for example by adopting the more naturalistic taxicab format (Kahneman and Tversky, 1972), changes the distribution of modal answers. It is therefore an open question, beyond the scope of this paper, whether framing complex tasks more naturally can lead to more social learning.

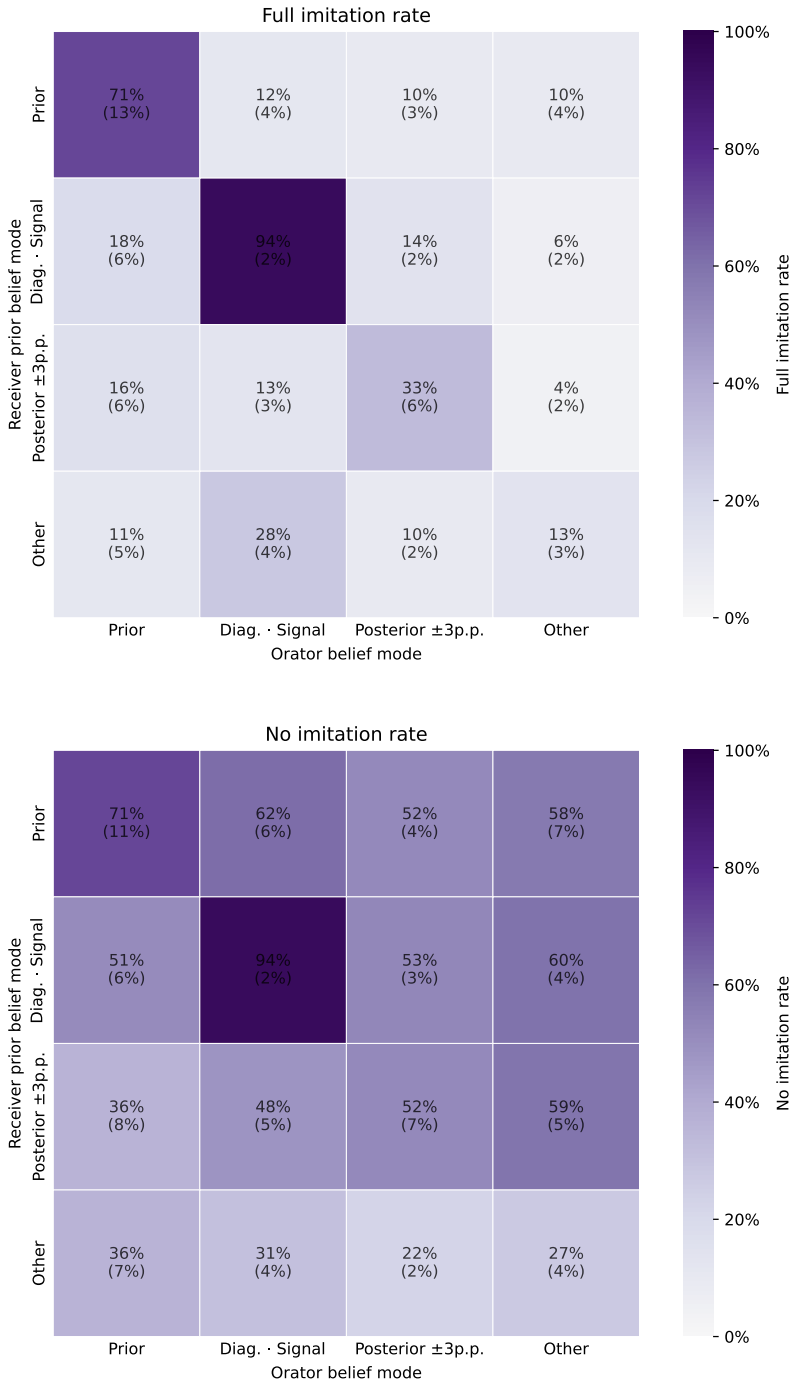


Figure L24: Imitation by receiver prior belief mode and orator belief mode in Balls-and-urns experiment. *Notes:* Top panel shows share of receivers stating the same belief as the orator, bottom panel shows share of receivers maintaining the same belief as their prior belief. Sample is the continuous Balls-and-urns receiver experiment.

M Survey screens

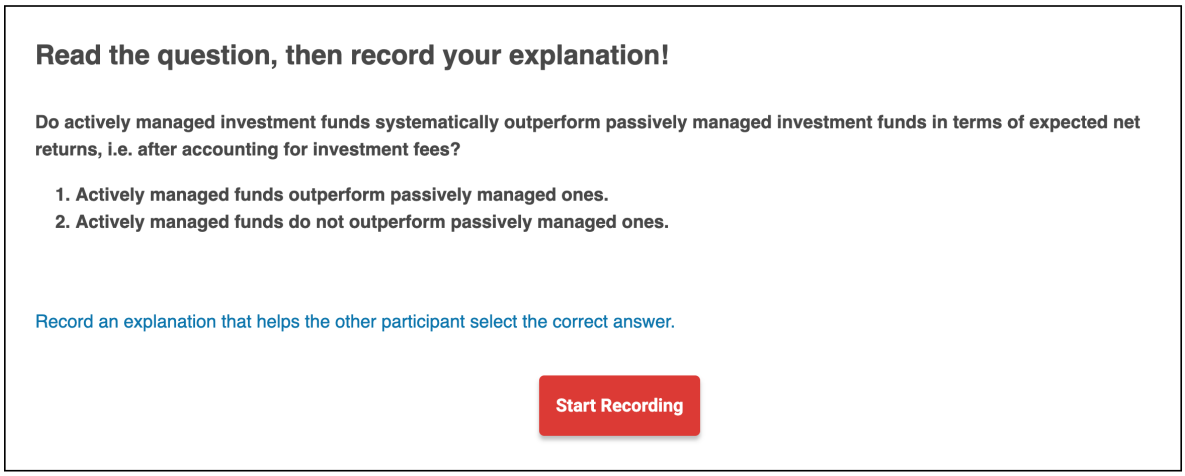


Figure M25: Recording screen from the Orator experiment.

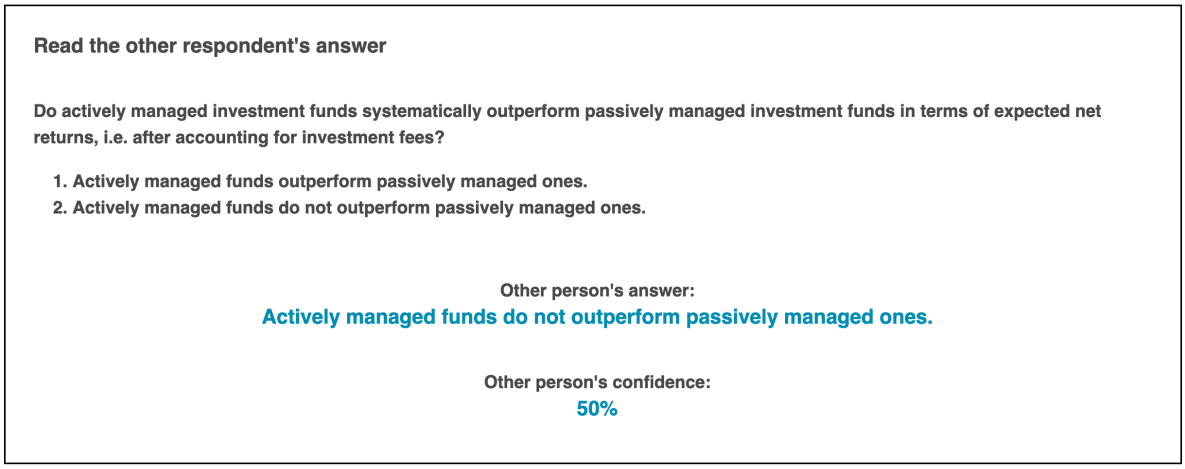


Figure M26: *Choice & Confidence* treatment screen from the Receiver experiment in Section 4.1.

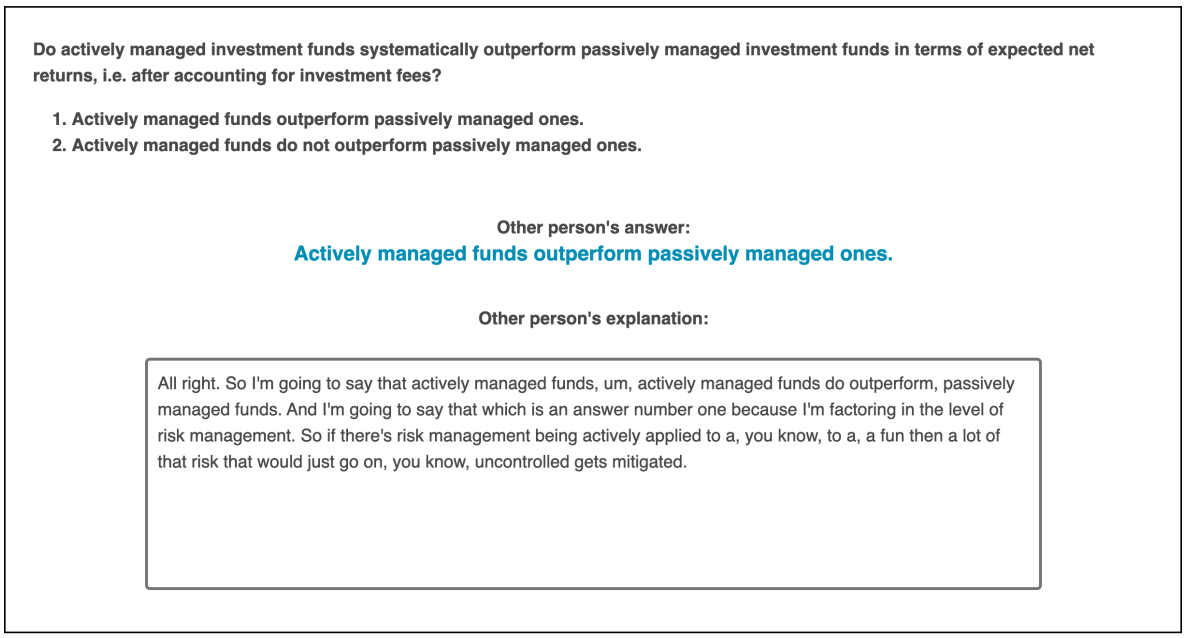


Figure M27: *Transcript* treatment screen from the Receiver experiment in Section 4.2.

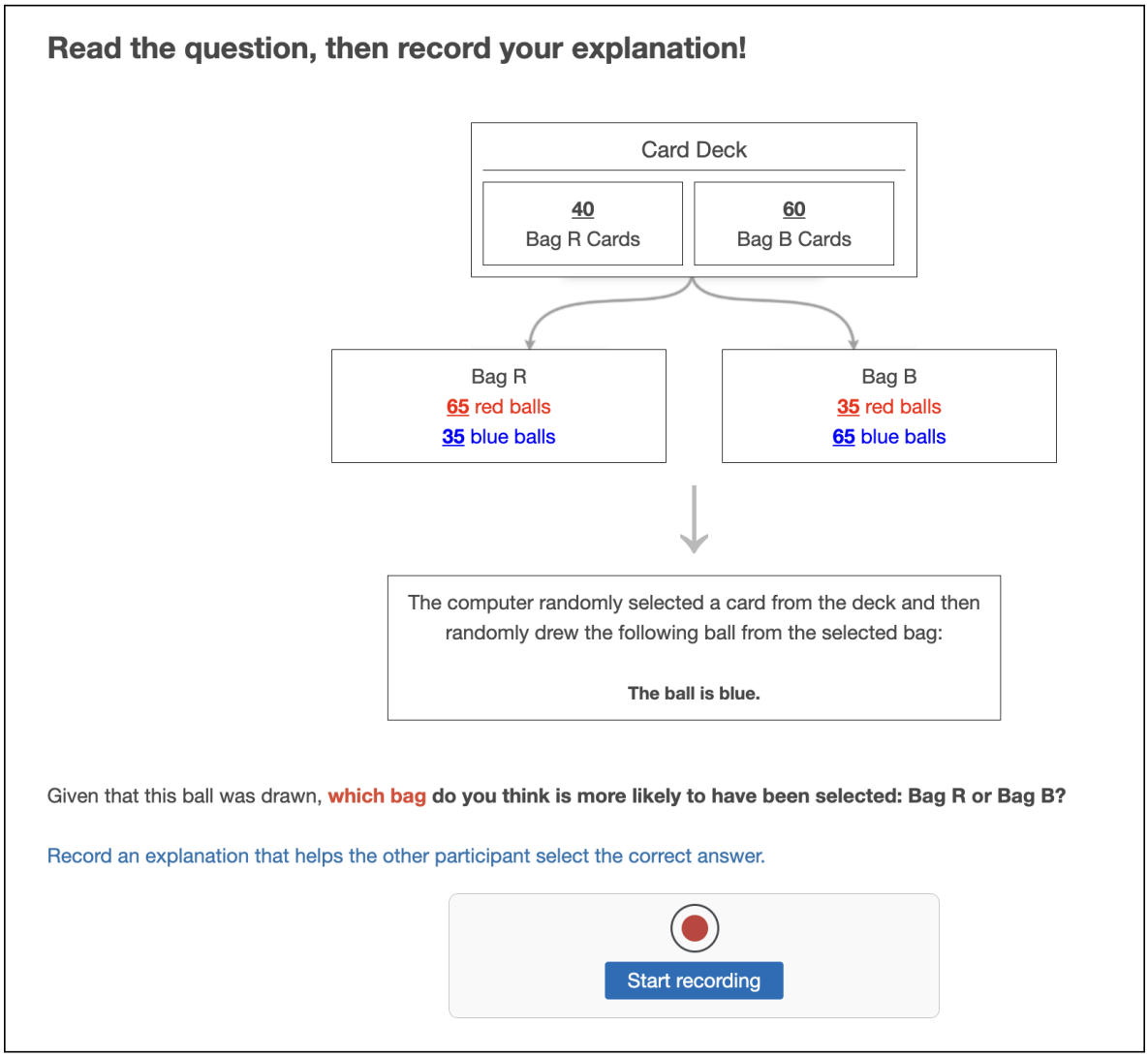


Figure M28: Recording screen from the binary Balls-and-urns orator experiment.

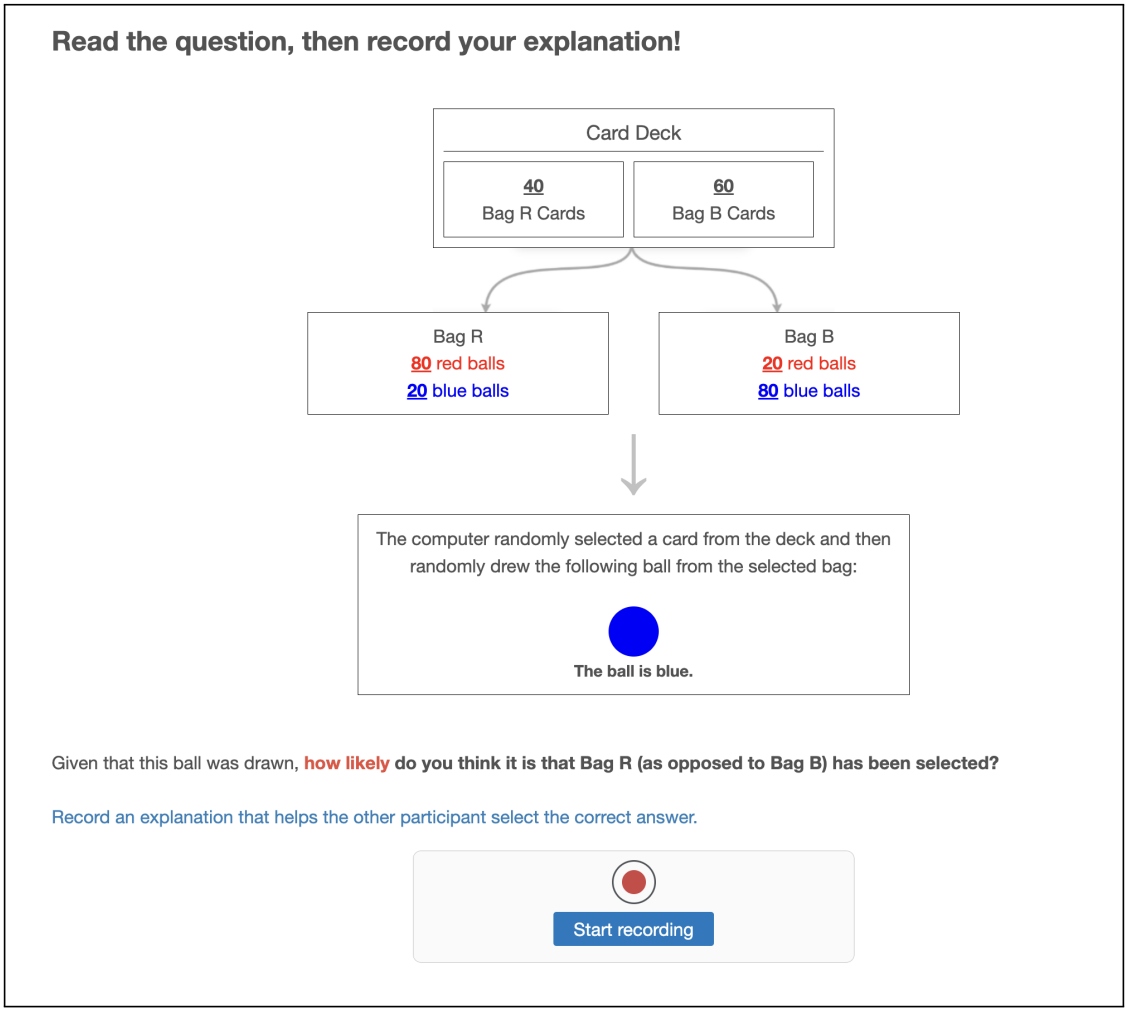


Figure M29: Recording screen from the continuous Balls-and-urns receiver experiment.

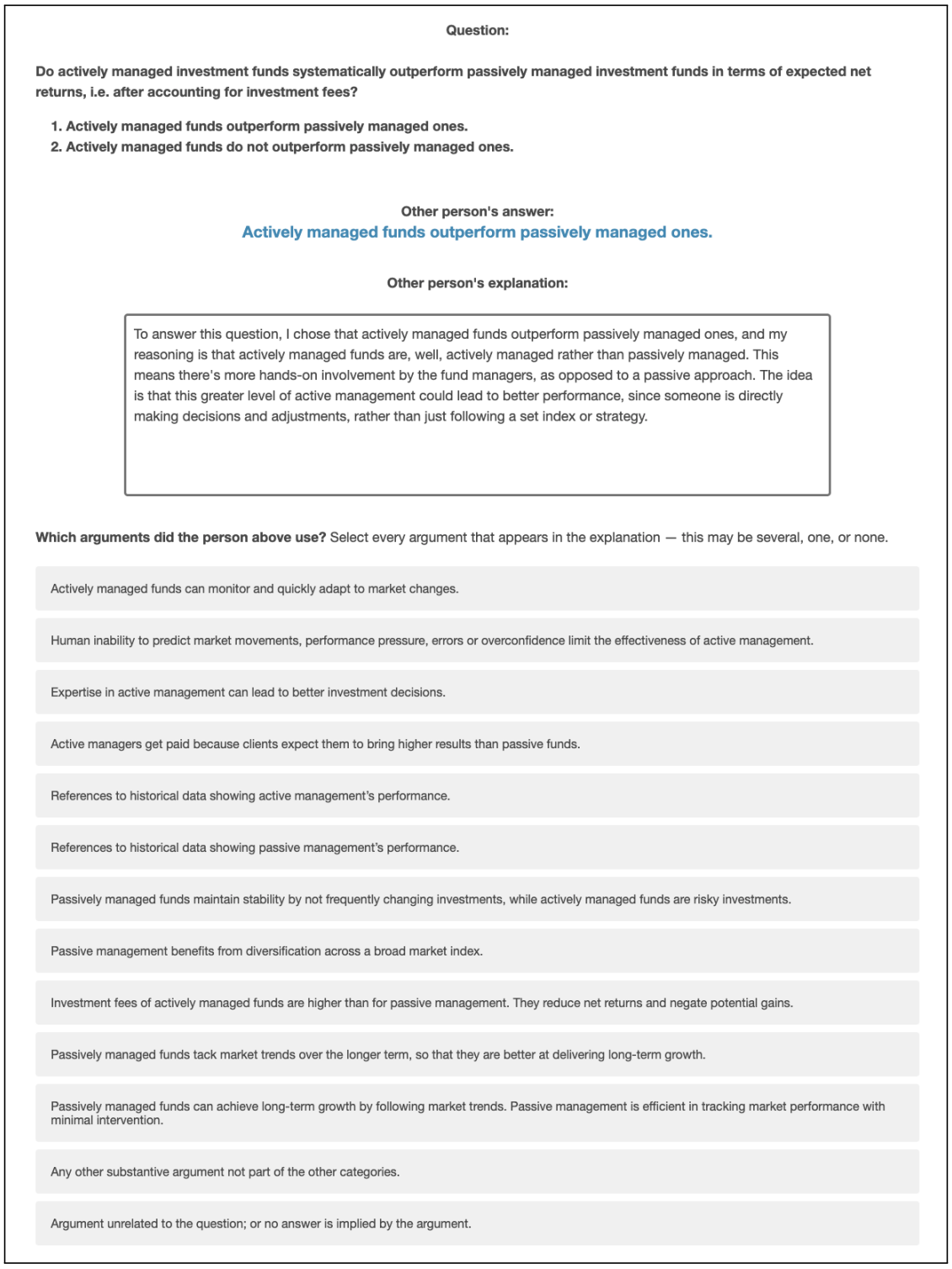


Figure M30: Screen from the argument annotation experiment.