Cognitive Psychology 132 (2022) 101453

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Motivated to learn: An account of explanatory satisfaction

Emily G. Liquin *, Tania Lombrozo

Department of Psychology, Princeton University, Peretsman Scully Hall, Princeton, NJ 08540, USA

ARTICLE INFO

Keywords:
Satisfaction
Explanation
Learning
Information search
Inquiry

ABSTRACT

Many explanations have a distinctive, positive phenomenology: receiving or generating these explanations feels *satisfying*. Accordingly, we might expect this feeling of explanatory satisfaction to reinforce and motivate inquiry. Across five studies, we investigate how explanatory satisfaction plays this role: by motivating and reinforcing inquiry quite generally ("brute motivation" account), or by selectively guiding inquiry to support useful learning about the target of explanation ("aligned motivation" account). In Studies 1–2, we find that satisfaction with an explanation is related to several measures of perceived useful learning, and that greater satisfaction in turn predicts stronger curiosity about questions related to the explanation. However, in Studies 2–4, we find only tenuous evidence that satisfaction is related to actual learning, measured objectively through multiple-choice or free recall tests. In Study 4, we additionally show that perceptions of learning fully explain one seemingly specious feature of explanatory preferences studied in prior research: the preference for uninformative "reductive" explanations. Finally, in Study 5, we find that perceived learning is (at least in part) causally responsible for feelings of satisfaction. Together, these results point to what we call the "imperfectly aligned motivation" account: explanatory satisfaction selectively motivates inquiry towards learning explanatory information, but primarily through fallible perceptions of learning. Thus, satisfaction is likely to guide individuals towards lines of inquiry that support perceptions of learning, whether or not individuals actually are learning.

1. Introduction

For many years the Egyptian pyramids presented a puzzle: How did the ancient Egyptians transport such enormous, multi-ton blocks of stone? And then, in a 2014 paper, a satisfying explanation was offered: the Egyptians could have transported the stone blocks on sleds, pouring water along the sand to reduce friction as they were pulled (Fall et al., 2014). The explanation was not only consistent with the results of experimental investigations testing the effects on friction of adding water to sand, but also with a wall painting from 1880 BCE found on the tomb of Djehutihotep, showing a figure pouring water from a jug at the front of a large sled.

This episode from the annals of Egyptology and physics illustrates a striking (if familiar) feature of human cognition: we are motivated to find explanations, and once found, explanations can be deeply satisfying. These phenomenological states—which we call explanation-seeking curiosity and explanatory satisfaction, respectively—seem to spur us on to seek information and to learn, both in science and in everyday life. Indeed, Gopnik (2000) compares explanatory satisfaction to the experience of orgasm: just as the latter motivates humans to engage in behaviors that support reproduction, so explanatory satisfaction motivates humans to engage in inquiry that supports learning about the causal structure of the world.

* Corresponding author at: Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA.
E-mail addresses: emily.liquin@nyu.edu (E.G. Liquin), lombrozo@princeton.edu (T. Lombrozo).

https://doi.org/10.1016/j.cogpsych.2021.101453
Received 19 March 2021; Received in revised form 11 August 2021; Accepted 24 November 2021
Available online 4 December 2021

While some connection between explanatory phenomenology and learning is widely accepted within psychology (see also Chater & Loewenstein, 2016; Vogl et al., 2021), there are two quite distinct forms this connection could assume. On one view, which we call the *aligned motivation* account, explanation-seeking curiosity and explanatory satisfaction are finely tuned to support epistemic success: we experience curiosity about the explanations that are most likely to provide useful knowledge, and we find them more satisfying when they do so. Put differently, insofar as curiosity and satisfaction are selective—and they surely are—they are selective in ways that *align* with the goal of effective learning. If we find Fall et al.'s (2014) explanation for the pyramids satisfying, for example, it is at least in part because we have gained new (and perhaps useful) explanatory knowledge. On another view, which we call the *brute motivation* account, explanation-seeking curiosity and explanatory satisfaction support learning simply because they motivate us to seek explanations. But insofar as these phenomenological states are selective, their selectivity is not specifically aligned with epistemic success. On this view, the satisfaction we gain from an explanation for the pyramids might be valuable insofar as it motivates and reinforces inquiry, but not due to the epistemic merits of the explanation itself.

As an analogy, consider two possible views about craving and "gustatory satisfaction," where the analogue for epistemic success is nutritional value. On one view—analogous to aligned motivation—the selectivity of craving and gustatory satisfaction is tuned to our nutritional needs: we will tend to crave and be satisfied by those foods with the highest nutritional value. On another view—analogous to brute motivation—the selectivity of craving and gustatory satisfaction is largely independent of our nutritional needs: we might crave and be satisfied by foods with suboptimal nutritional value. But because craving and gustatory satisfaction motivate us to eat, they will (in the right environment) nonetheless result in the fulfillment of our nutritional needs.

For curiosity, and for explanation-seeking curiosity in particular, prior work supports an aligned motivation account over a brute motivation account: we experience greater curiosity when we expect to learn useful information (Abir et al., 2020; Dubey et al., 2019; Dubey & Griffiths, 2020; Liquin et al., 2020; Liquin & Lombrozo, 2020a). For instance, Liquin and Lombrozo (2020a) found that two of the strongest predictors of a participant's curiosity about the answer to a "why" question were the extent to which the participant expected the answer to support learning and to be useful in the future. In addition, explanation-seeking curiosity was related to explanation-seeking behavior: the more curiosity that a participant reported about the answer to a question, the more likely they were to choose to reveal the answer to that question over others. These findings suggest that explanation-seeking curiosity motivates inquiry selectively, and towards explanations that are expected to be epistemically valuable.

When it comes to explanatory satisfaction, however, the picture is much less clear. Research suggests that explanatory satisfaction is selective in the sense that some explanations are found more satisfying than others (e.g., Johnson et al., 2019; Khemlani et al., 2011; Lim & Oppenheimer, 2020; Lombrozo, 2007; Pacer & Lombrozo, 2017). However, it's not clear whether explanations that are "good" in the sense that they induce explanatory satisfaction are also "good" in the sense that they successfully reflect and motivate learning. In fact, explanatory satisfaction can at times point *away* from epistemic success (Giffin et al., 2017; Hopkins et al., 2016; Johnson et al., 2016; Khemlani et al., 2011; Trout, 2008; Weisberg et al., 2008, 2015; Williams et al., 2013). These findings suggest that insofar as the phenomenology of explanation supports epistemic success, it may be through brute motivation, not aligned motivation.

The primary goal of the present research is to test the aligned motivation account of explanatory satisfaction. Of course, the right account may well be in between aligned and brute motivation, with explanatory satisfaction tracking epistemic merits to some extent, but falling short of perfect correspondence. If so, then identifying the contours of this partial correspondence is nonetheless informative for understanding how we do (and do not) succeed in efficiently learning useful information about the world. Below we motivate three predictions that follow from the aligned motivation account: the actual learning prediction, the perceived learning prediction, and the selective reinforcement prediction. In the course of doing so we review relevant prior work, and we then offer an overview of the five experiments we go on to report.

1.1. Aligned motivation: actual learning

The first prediction of the aligned motivation account is that explanations that are "good" in the sense that they engender satisfaction should also be "good" in the sense that they support learning. More concretely, we ought to be satisfied by explanations to the extent they successfully teach us something relevant to the posed query that is useful and new. The most direct evidence for this prediction—what we call the *actual learning prediction*— would be an association between the experience of explanatory satisfaction and some objective measure of explanatorily relevant learning. For example, if individuals were presented with the explanation for the Egyptian pyramids offered in Fall et al. (2014), we would expect their satisfaction with the explanation to track gains in their knowledge of how the pyramids were constructed. The brute motivation account instead suggests that satisfaction need not be reliably linked to learning.

Unfortunately, there are two important reasons to doubt that the actual learning prediction would find empirical support. First, people are typically poor at assessing their own learning: learners' judgments of comprehension after reading a passage are often unrelated or weakly related to their performance on a test assessing actual learning (Glenberg & Epstein, 1985; Lin & Zabrucky, 1998; Maki, 1998). Moreover, judgments of learning are formed at least in part using heuristic cues, such as processing fluency, that fall short of perfectly indexing actual comprehension or learning (Begg et al., 1989; Hertzog et al., 2003; Koriat, 1997; Reber & Greifeneder, 2017). Likewise, judgments of confidence are often miscalibrated because people fail to consider unknowns (Walters et al., 2017). In addition, individuals often believe that their explanatory knowledge is much more complete than it actually is (Rozenblit & Keil, 2002). If judgments of learning and assessments of explanatory understanding are unreliable, it seems unlikely that satisfaction could be a more reliable guide to actual learning.

Second, there is evidence that explanatory satisfaction often tracks superficial and potentially irrelevant features of an explanation. For example, people sometimes judge explanations that appeal to reductive information more satisfying than those that do not (e.g., an

explanation for a psychological phenomenon that includes neuroscience information; Hopkins et al., 2016; Weisberg et al., 2008). This has been found even when the reductive information is arguably uninformative with regard to the target question, at least according to experts. Other features of explanations—such as length (Weisberg et al., 2015) and the inclusion of categorical labels (Giffin et al., 2017)—have also been shown to boost the perceived quality of an explanation, even when the added value of the additional information is highly questionable. To the extent these influences on explanatory satisfaction point away from actual learning (at least in some circumstances), these findings challenge the actual learning prediction.

Despite these considerable reasons for pessimism, there is some evidence consistent with the actual learning prediction. For example, memory for the answers to trivia questions is related to the extent to which satisfaction exceeds initial curiosity (Marvin & Shohamy, 2016). In addition, children are better able to recall explanations over non-explanations (Frazier et al., 2016). These findings provide tentative evidence that when one learns more (or perhaps more than expected), there may be a corresponding boost to explanatory satisfaction.

1.2. Aligned motivation: perceived learning

A more modest prediction of the aligned motivation account is the *perceived learning prediction*. According to the perceived learning prediction, explanations ought to be satisfying in proportion to how strongly they are *perceived* to support useful learning relevant to the posed query. Given that metacognitive assessments of learning are typically so poor, the perceived learning prediction is antecedently much more plausible than the actual learning prediction. If explanatory satisfaction tracks perceived learning quite closely—but actual learning rather poorly—this could support a version of the aligned motivation account according to which explanatory satisfaction is "as aligned as possible" given the regrettable limitations of our metacognitive capabilities.

Prior work relevant to the perceived learning prediction is scarce. In one study, Zemla et al. (2017) related judgments of explanation quality ("This is a good explanation") to what they called novelty ("I learned something new from this explanation"), but found that novelty did not reliably predict explanation quality after correcting for multiple comparisons. Most prior research has instead tested whether explanation quality or satisfaction is influenced by the presence of "explanatory virtues." In particular, explanations are typically valued to the extent they are broad and generalizable (Johnston et al., 2018; Kim & Keil, 2003; Read & Marcus-Newhall, 1993), and under some conditions, to the extent they are simple (Blanchard, Lombrozo, et al., 2018; Lombrozo, 2007; Pacer & Lombrozo, 2017; but see Johnson et al., 2019; Zemla et al., 2017, 2020; Lim & Oppenheimer, 2020). It has sometimes been proposed that these explanatory virtues inform explanation quality precisely because they are cues to useful learning: for example, simplicity and breadth could indicate that an explanation is likely to support generalization beyond the specific case being explained (Blanchard, Vasilyeva, et al., 2018; Friedman, 1974; Kitcher, 1989; Lombrozo, 2016; Strevens, 2004; Thagard, 1978), which might contribute to perceptions of learning or the utility of information. However, the link between explanatory virtues and perceptions of learning has not (to our knowledge) been tested empirically, and other accounts of when explanatory simplicity is preferred (and when it is not) have been proposed (Johnson et al., 2019; Lim & Oppenheimer, 2020). Therefore, it is unclear whether individuals' sensitivity to explanatory virtues provides evidence for the perceived learning prediction.

1.3. Aligned motivation: selective reinforcement

A final prediction of the aligned motivation account is that receiving a satisfying explanation should be related to subsequent inquiry. This prediction is shared by the brute motivation account, which also treats satisfaction as an internal "reward" that reinforces inquiry. However, the aligned motivation account makes more specific predictions about the selectivity of this reinforcement. On the aligned motivation account, further inquiry should be guided by expectations about learning, which are constrained by the explanation that has just been received. If that explanation has satisfactorily answered the initial query (and is therefore satisfying), then the learner should no longer expect to learn as much by pursuing an answer to that question, and inquiry concerning the initial query should halt. However, the explanation might lead to the recognition of *new* questions of interest (see Murayama et al., 2019), and signal the value of inquiry within the general domain. Thus, we might expect a satisfying answer to increase curiosity about the answers to questions that follow from the explanation. But because the answer to a given question shouldn't change expectations about learning from the answers to *unrelated* questions, we would not expect a satisfying explanation to be related to increased curiosity about new questions indiscriminately: the inquiry that is reinforced should be related to the given explanation. In sum, the aligned motivation account makes the following three predictions concerning *selective reinforcement*: the extent to which an explanation is satisfying should be related to (a) a decrease in curiosity concerning the original query and (b) higher curiosity about the answers to related questions, but (c) should be unrelated to curiosity about the answers to unrelated questions.[1] While the brute motivation account shares a commitment to the idea that a satisfying explanation should reinforce inquiry, it does not clearly specify the form this reinforcement should assume across the three kinds of cases considered above.

Prior work offers some support for prediction (a): Children are less likely to re-ask a question after receiving an explanatory response as opposed to a non-explanatory response (Frazier et al., 2009, 2016; Kurkul & Corriveau, 2018), indicating that the receipt of

[1] Note that these predictions are neutral with respect to causal mechanism. It is possible that satisfaction is causally responsible for subsequent curiosity, mediated by expectations about learning. Alternatively, it is possible that the amount of learning from an explanation (perceived or actual) independently determines both satisfaction and expectations about learning, the latter of which determines subsequent curiosity. In the present research, we only test the correlation between satisfaction and subsequent curiosity, and we leave questions of causal mechanism to future research.

an explanation halts children's inquiry on that question. Similarly, children are less likely to request additional information in response to a target question when a mechanistic explanation is provided as opposed to a circular explanation (Mills et al., 2019). In contrast, and consistent with (b), children ask more follow-up questions when an experimenter provides informative responses as opposed to uninformative responses (Ünlütabak et al., 2019), indicating that the receipt of informative responses may be reinforcing and lead to further questions. However, to our knowledge, there has been no systematic investigation of whether and how a satisfying explanation motivates further inquiry in adults, and no systematic comparison of inquiry on related versus unrelated follow-up questions.

1.4. Overview of experiments

In the present research, we test the aligned motivation account of explanatory satisfaction. The aligned motivation account predicts that satisfaction is determined by actual and perceived learning (the actual learning prediction and the perceived learning prediction), while the brute motivation account predicts that satisfaction is unrelated to actual and perceived learning. The aligned motivation account also predicts that a satisfying explanation should be related to less curiosity about the initial question, greater curiosity about related questions, and no change in curiosity about unrelated questions (selective reinforcement prediction). The brute motivation account is similarly consistent with a role for explanatory satisfaction in reinforcing inquiry but does not offer a clear basis for predicting selectivity in such effects (see Table 1).

Using self-report ratings of naturalistic stimuli (i.e., real-world questions and answers from question-and-answer books and textbooks) combined with multiple assessments of learning, we test the actual learning prediction (Studies 2–4), the perceived learning prediction (Studies 1-2 and 4-5), and the selective reinforcement prediction (Studies 1–2). To preview our results, we find that objective assessments of learning are not reliably associated with explanatory satisfaction, but perceived learning is robustly related to satisfaction and is at least in part causally responsible for satisfaction. In addition, we find some evidence for the selective reinforcement prediction: satisfaction with an explanation is more strongly associated with curiosity about related versus unrelated follow-up questions, but satisfaction is not related to decreased curiosity about the initial question. Finally, in Study 4, we use our account to explain prior work that has cast explanatory satisfaction as sensitive to "irrational" explanatory vices. We find that individuals' explanatory satisfaction in the presence of superficial reductive information that does not *objectively* provide relevant explanatory content can be fully explained by participants' *subjective* judgment that it does.

Together, these studies take important steps towards a comprehensive characterization of explanatory satisfaction as it relates to epistemic success. This research illuminates one possible reason humans experience explanatory phenomenology: to motivate the process of explanatory inquiry, guiding learners to seek explanations and—sometimes—to achieve a better understanding of the world.

2. Study 1

In Study 1, we test the perceived learning prediction, as well as one component of the selective reinforcement prediction. We present participants with explanation-seeking questions (e.g., "Why do some stars explode?"), and later with their corresponding answers. Participants indicate how satisfying they find each answer, and also report perceived learning through several measures of learning and utility (e.g., "To what extent has the answer to this question taught you something new?").

We also test the association between satisfaction and several explanatory virtues (see Table 2 for all measures). Measuring these explanatory virtues allows us to determine whether perceived useful learning explains variance in satisfaction above and beyond these additional cues to satisfaction, which have been the focus of prior research. We include a measure of simplicity, a measure of breadth, and a measure of the extent to which the explanation identifies a general pattern or regularity (referred to as "regularity").

We also include a measure of expertise ("Do you think that answering this question required special expertise in some domain?"). Prior research has emphasized that explanations are frequently attained and justified by deference to domain experts, and children and adults are adept at identifying which experts to consult given a particular question (Bromme & Thomm, 2016; Danovitch & Keil, 2004;

Table 1
Predictions of the aligned motivation account in comparison to the brute motivation account, along with evidence for/against these predictions across our five studies.

Prediction	Aligned Motivation Account	Brute Motivation Account	Evidence
Actual Learning Prediction	Explanations should be more satisfying to the extent they result in useful learning.	No connection between satisfaction and actual learning.	Studies 2–4: some evidence for a connection between satisfaction and actual learning (free recall, not multiple-choice performance).
Perceived Learning Prediction	Explanations should be more satisfying to the extent they induce perceptions of useful learning.	No connection between satisfaction and perceived learning.	Studies 1–2, 4–5: evidence that perceived useful learning is (causally) related to satisfaction; perceived learning explains preference for reductive explanations.
Selective Reinforcement Prediction	Satisfaction increases perceived value of subsequent inquiry on related questions, decreases perceived value of subsequent inquiry on initial question, and no change in perceived value of inquiry on unrelated questions.	Satisfaction reinforces the perceived value of further inquiry. While this reinforcement may be selective (such that more satisfying explanations are more reinforcing), it should not track expectations about learning.	Studies 1–2: association between satisfaction and subsequent curiosity is stronger for related vs. unrelated questions, but no evidence for an association between satisfaction and reduced curiosity about initial question.

Table 2
Items rated (each on a seven-point scale) in response to explanations in Studies 1–2.

Dimension	Item (Scale Anchors)
Phenomenology of explanatory inquiry	
Satisfaction	How satisfying do you find the answer to this question? (Not at all satisfying – Very satisfying)
Perceived useful learning	
Learning	To what extent has the answer to this question taught you something new? (Definitely nothing new – Definitely a lot new)
Information Content	Do you think there is something to be learned from the answer to this question (even if you yourself already knew the answer)? (Definitely nothing to be learned – Definitely something to be learned)
Future Utility	To what extent will the answer to this question be useful to you in the future? (Not at all useful – Very useful)
Explanatory virtues	
Simplicity	Do you think the answer to this question is simple or complex? (Very simple – Very complex)
Breadth	Do you think the answer to this question is narrow (only applies to what is being explained) or broad (also applies to other similar cases)? (Very narrow application – Very broad application)
Regularity	Do you think the answer to this question helps reveal a genuine pattern, structure, or regularity? (Definitely does not – Definitely does)
Expertise	Do you think that answering this question required special expertise in some domain? (Definitely did not – Definitely did)

Keil et al., 2008; Lutz & Keil, 2002) and at asking for expert help specifically when it is needed (Kominsky et al., 2018; Vredenburgh & Kushnir, 2016). Furthermore, adults are more likely to defer to others who they judge to "understand why" a relevant event happened (Wilkenfeld et al., 2016), and an association between judgments of an explanation's quality and judgments that it was written by experts has been reported in prior research (Zemla et al., 2017). Thus, we include perceptions of expertise among other explanatory virtues, and as a feature of explanations that is plausibly linked to perceived learning.

As a final feature of our task, we ask participants to indicate how curious they are about questions that follow-up on each provided explanation. This allows us to test one component of the selective reinforcement prediction: that more satisfying explanations will be associated with greater curiosity about related follow-up questions.

2.1. Method

2.1.1. Participants

Participants were 184 adults (106 male and 78 female, ages 21–69) recruited from Amazon Mechanical Turk (MTurk). Sixteen additional participants completed Study 1 but were excluded from analysis because they did not pass two attention checks (see below). The sample size was planned to match that of an initial study, reported in the supplementary material (Study S1). For all studies conducted through MTurk (reported here and in the supplementary material), participation was restricted to MTurk workers in the United States who had completed at least 1000 prior tasks with a minimum approval rating of 99%, and participation was restricted so that participants could not participate in more than one study in this series.

2.1.2. Materials

Twenty explanation-seeking questions and answers were selected from the book *1000 Questions & Answers Factfile* (Kerrod et al., 2006). For example, the question "Why do some stars explode?" was answered with the following explanation: "Massive stars explode when they come to the end of their lives. They swell up into huge supergiants. Supergiants are unstable, so they collapse and blast into pieces in an explosion called a supernova. Supernovae are the most intense explosions in the universe, as bright as billions of suns put together." Explanations ranged in length from 34 to 91 words.

Each question–answer pair was classified into a "topic," based loosely on the chapter and page topics in the *1000 Questions & Answers Factfile* book. The 20 questions fell into 14 distinct topics (e.g., "Dinosaurs," "Stars," "Ancient Egypt"). Additionally, to obtain follow-up questions associated with each question–answer pair, an independent sample of 48 participants on Amazon Mechanical Turk read random samples of five question–answer pairs (from the larger set of 20) and wrote between 3 and 10 follow-up questions in response to each answer (see supplementary material Study S3 for further information). From this set of follow-up questions, we selected 10 for each question–answer pair. These follow-up questions were edited lightly for grammar and readability. The full set of materials can be found at https://osf.io/hf3cj/.

2.1.3. Procedure

The study had five phases, which we describe in turn below. First, in the *interest/knowledge phase*, participants were presented with the 14 topics described above and rated their interest in each topic ("How interested are you in the following topics?") and how knowledgeable they were about each topic ("How much do you know about the following topics?"), each on a seven-point scale. We included these measures as controls: if an individual is particularly interested in or knows a lot about a topic (e.g., dinosaurs), they might find any explanations pertaining to that topic satisfying or be especially curious about related follow-up questions.

Next, in the *question phase*, participants saw four questions (without their corresponding answers) randomly selected from the 20 questions described above, each on a separate page. For each question, participants indicated their curiosity about the answer ("How curious are you about the answer to this question?") and responded to seven items assessing their expectations about the answer, each on a seven-point scale. These measures were drawn from those used by Liquin and Lombrozo (2020a), and they were included to assess an additional prediction of the aligned motivation account: that participants should be well-calibrated in their expectations about learning. We report corresponding analyses in the supplementary material.[2]

The question phase was followed by a short *distractor phase*, in which participants completed a distractor task that doubled as an attention check. Participants answered seven arithmetic problems, where one of the arguments of each problem was the solution from the previous problem. A correct answer was determined based on a participant's previous solution, so that an incorrect solution on one problem was not counted against the participant for subsequent problems. Those who did not correctly respond to at least five items were excluded from all analyses.[3]

In the *answer phase*, participants saw the same questions, but this time with their corresponding answers, each on a separate page. For each explanation, participants rated explanatory satisfaction ("How satisfying do you find the answer to this question?"), three items assessing perceived learning, and four items assessing explanatory virtues (see Table 2), each on a seven-point scale. The order of these items (including satisfaction) was randomized for each participant.

In the *follow-up phase*, participants saw each question–answer pair again, along with five randomly selected related follow-up questions. For each question–answer pair, they responded to the following prompt: "After reading the answer above, how curious are you about the answers to the following questions?" Curiosity about each follow-up question was indicated on a seven-point scale (1 = Not at all curious; 7 = Very curious).

Finally, participants provided their age and selected one of: male, female, other, prefer not to specify. They also completed an additional attention check that required them to select the four explanation-seeking questions they had seen in the previous phases from a list that also included four distractor questions. Participants were given one point for each correct response (hit or correct rejection), and those who scored fewer than six points were excluded.

2.2. Results

2.2.1. Analytic approach

We take the following approach to analyses: all continuous ratings were z-scored before analysis. Correspondingly, we report standardized regression coefficients (β) for all analyses. For all analyses, we fit mixed-effects regression models using the lme4 package (Bates et al., 2015) in R. The reported models include random intercepts for all relevant clustering variables (i.e., participant and question–answer pair). In all cases where the significance of a predictor is tested, we compare the full regression model with a reduced model excluding that predictor using a likelihood ratio test.

2.2.2. Testing the perceived learning prediction

First, we investigate the associations between explanatory satisfaction, perceived learning, and our four explanatory virtues (see Fig. 1). Ratings on each learning and explanatory virtue measure were entered into a simultaneous regression model predicting explanatory satisfaction. We also included topic-level ratings of interest and knowledge as fixed effects to control for these as possible confounds. Together, the learning and explanatory virtue measures explained significant variance in satisfaction beyond interest and knowledge, $\chi^2(7) = 271.17$, $p < .001$. In particular, unique variance in explanatory satisfaction was explained by perceived learning, $\beta = 0.13$, 95% CI [0.05, 0.20], $\chi^2(1) = 11.58$, $p < .001$, information content, $\beta = 0.31$, 95% CI [0.23, 0.39], $\chi^2(1) = 59.20$, $p < .001$, future utility, $\beta = 0.12$, 95% CI [0.04, 0.19], $\chi^2(1) = 9.05$, $p = .003$, regularity, $\beta = 0.15$, 95% CI [0.08, 0.22], $\chi^2(1) = 17.26$, $p < .001$, and expertise, $\beta = 0.09$, 95% CI [0.01, 0.17], $\chi^2(1) = 5.02$, $p = .03$. In contrast, simplicity, $\beta = 0.02$, 95% CI [-0.05, 0.10], $\chi^2(1) = 0.43$, $p = .51$, and breadth, $\beta = 0.01$, 95% CI [-0.06, 0.07], $\chi^2(1) = 0.04$, $p = .85$, did not explain unique variance in satisfaction. These results suggest that explanatory satisfaction is related to perceptions of learning above and beyond explanatory virtues—most notably, direct measures of how much has been learned, how much information an explanation contains, and how useful that information is. In addition, satisfaction is related to perceptions of the explanation's generality and expertise. Furthermore, as we report in the supplementary material and as depicted in Fig. 1, all learning measures and explanatory virtue measures were related to explanatory satisfaction when tested in isolation.

2.2.3. Testing the selective reinforcement prediction

We next tested whether explanatory satisfaction motivates further inquiry on related questions, which we assessed through

[2] In the interest of brevity, we omit a full analysis of this "calibration prediction" in Studies 1–2, but we direct interested readers to supplementary material for a full account. In brief, we find fairly high levels of calibration: when participants expect an explanation to support learning, or to be simple, they generally report that it is. Moreover, this correspondence is not an artifact of a within-subjects design, as we also find such correspondence when one group of participants indicates expectations about explanations, and an independent group evaluates the corresponding explanations (supplementary material Study S2).

[3] While we intended this attention check to assess whether participants were reading instructions and paying attention, it may also assess numeracy, leading to potentially unnecessary exclusions. We repeated all analyses (in all studies) including participants who failed this attention check; the pattern of results remained unchanged.

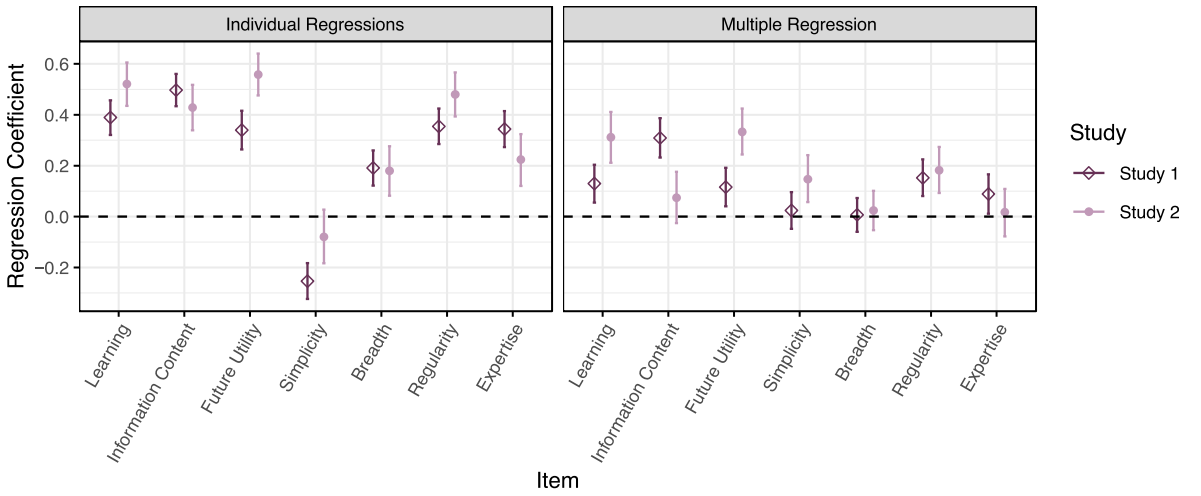

Fig. 1. Explanatory Satisfaction and Perceived Learning. Association between each measure and rated explanatory satisfaction, in Studies 1 and 2. For individual regressions (left panel), each measure was entered into a separate regression model predicting satisfaction, with random intercepts for participant (in Study 1) and question–answer pair (in Studies 1–2). For multiple regressions (right panel), each measure was entered as a fixed effect in a simultaneous regression model, with random intercepts for participant (in Study 1) and question–answer pair (in Studies 1–2). Topic-level interest and knowledge were included as control measures in both individual and multiple regressions.

curiosity ratings. The five follow-up curiosity ratings were averaged to create a "follow-up curiosity score" (Cronbach's $\alpha = 0.84$). We fit a mixed-effects regression model predicting follow-up curiosity score, with explanatory satisfaction, interest, and knowledge as fixed effects. As predicted, explanatory satisfaction was significantly associated with follow-up curiosity, $\beta = 0.22$, 95% CI [0.16, 0.28], $\chi^2(1) = 48.38$, $p < .001$. The more satisfied a given participant was with a given explanation, the more curious they were about related follow-up questions.

2.3. Discussion

In Study 1, we found preliminary evidence supporting two core predictions of the aligned motivation account. First, consistent with the perceived learning prediction, we found that the satisfaction derived from an explanation was associated with several measures of perceived useful learning—not only how much was learned from the explanation, but also how much information it was judged to contain and how useful it was expected to be. Several explanatory virtues also explained unique variance in satisfaction, and all measures were related to satisfaction when considered in isolation.[4] Second, consistent with the selective motivation prediction, we found that explanatory satisfaction was positively related to subsequent curiosity about related follow-up questions. For example, if an explanation for "Why did dinosaurs swallow stones?" was found particularly satisfying by a given participant, that participant was likely to report more curiosity about the question "What percentage of the stomach space did the stones take up?".

These results are consistent with the aligned motivation account, but they also raise a number of questions. First, while we found support for the perceived learning prediction, there is still reason to doubt the actual learning prediction (e.g., Glenberg & Epstein, 1985; Lin & Zabrucky, 1998; Maki, 1998). Does explanatory satisfaction track actual learning of explanatory information, or only *perceived* learning? To address this question, the following study includes an objective measure of learning. Second, the association between satisfaction and subsequent curiosity about follow-up questions supported only one element of the selective reinforcement prediction: that a satisfying explanation should foster curiosity concerning *related* questions. This finding is also consistent with the brute motivation account. A more diagnostic finding would show that reinforcement is selective, such that curiosity about *unrelated* questions is not fostered to the same extent, and that curiosity about the initial question is decreased. We test these additional predictions in Study 2.

[4] Interestingly, ratings of an explanation's simplicity were negatively related to explanatory satisfaction. In prior work, paradigms that operationalize simplicity by participant report tend to find preferences for more complex explanations (Liquin & Lombrozo, 2020a; Zemla et al., 2017), while paradigms that operationalize simplicity using objective metrics (e.g., the causal structure invoked in an explanation) tend to find preferences for simpler explanations (Blanchard, Lombrozo, et al., 2018; Bonawitz & Lombrozo, 2012; Lombrozo, 2007; Pacer & Lombrozo, 2017). Given that simplicity/complexity was not a significant predictor of satisfaction when controlling for other measures of perceived learning, it is possible that the folk conception of "simplicity" or "complexity" has more to do with the amount of information contained in an explanation than the formal structure of that explanation, and that preferences for simplicity only emerge when unconfounded from perceived learning.

3. Study 2

Study 2 had four aims. First, we attempt to replicate evidence for the perceived learning prediction from Study 1, but using a new set of questions and answers. Second, we test the actual learning prediction: are explanations that are "good" in the sense that they are satisfying also "good" in the sense that they support actual learning? To do so, we assess multiple-choice performance on a post-test, under the assumption that learning that occurs from reading an explanation (and that might generate a sense of explanatory satisfaction at this moment of learning) will be reflected in accuracy on the subsequent post-test. Third, we explore whether satisfaction (to the extent that it is related to actual learning at all) is specifically related to *relevant* learning (i.e., performance on multiple-choice questions that assess core explanatory content), as opposed to learning of any content contained within an answer to a question (i.e., performance on multiple-choice questions that assess superficial or irrelevant content).

Finally, we follow up on the results of Study 1 to further test the selective reinforcement prediction. We test whether satisfaction is related to *decreased* curiosity about the original question, and we explore whether the association between satisfaction and subsequent curiosity is selective to related follow-up questions, or whether it extends to unrelated questions, as well. Study 2 was preregistered (https://aspredicted.org/qp76w.pdf); any departures from our preregistered analysis plan are detailed in the Results section.

3.1. Method

3.1.1. Participants

Participants were 396 adults (221 male and 175 female, ages 20–78) recruited from Amazon Mechanical Turk. Twenty-five additional participants completed the study but were excluded from analysis because they failed to pass two attention checks. The target sample size of 340 was preregistered,[5] and was based on a power analysis using the R package simr (P. Green & MacLeod, 2016) based on pilot data.

3.1.2. Materials

Twenty explanations were selected from fifteen introductory-level textbooks, with no more than two explanations drawn from a single textbook. Explanations were edited slightly so that they could stand alone without additional textbook content. For each explanation, an explanation-seeking question was constructed. For example, the question "Why does the angle of the sun affect the seasons on Earth?" was answered with the explanation "Earth passes through different seasons as it circles the Sun. As Earth travels around the Sun, in June the Northern Hemisphere 'leans into' the Sun and is more directly illuminated. When a hemisphere leans into the Sun, sunlight hits that hemisphere at a more direct angle and is more effective at heating Earth's surface." Explanations ranged in length from 40 to 150 words.

As in Study 1, each question–answer pair was classified into a "topic." The 20 questions fell into 19 distinct topics (e.g., "the science of consciousness," "emotion," "electric circuits"). For each question–answer pair, we also generated three related follow-up questions and four multiple-choice comprehension questions. Two of the multiple-choice questions for each question–answer pair concerned details that were explanatorily relevant (i.e., "relevant details"), while two questions concerned irrelevant details (e.g., the order of information in the explanation). With the exception of the follow-up questions and topic classifications, these materials were drawn from Aronowitz et al. (in prep), who also pre-tested and normed the comprehension questions. The four comprehension questions for each question–answer pair were normed so that participants who had not read the answer scored no more than 50% correct averaged over all four questions, and so that each question was answered correctly by no more than 75% of participants. The full set of questions, explanations, topics, follow-up questions, and multiple-choice questions can be found at https://osf.io/hf3cj/.

3.1.3. Procedure

Participants were randomly assigned to a single question–answer pair from the set of 20. First, participants completed the *interest/knowledge phase*, the *question phase*, the *distractor phase*, and the *answer phase*, which were identical to those in Study 1. Next, participants completed three sets of follow-up curiosity ratings, using a seven-point scale ranging from 1 (Not at all curious) to 7 (Very curious). For the first set of curiosity ratings, participants again rated their curiosity about the same question that had been answered, given the following prompt: "Next, you will see the same question again. Now that you've read the answer to this question, please rate your curiosity again (that is, we're interested in how curious you are *now* to know the answer to the question)." Then, participants rated their curiosity about the three follow-up questions related to the explanation they had read (related follow-up questions), as well as three follow-up questions related to an explanation they had not read (unrelated follow-up questions), with the two sets of three questions presented in a random order. Next, participants answered four multiple-choice questions assessing the extent to which they had learned from the explanation. Finally, participants provided their age and gender, and answered an attention check question that required them to identify the explanation-seeking question they had rated from a list of four distractors.

[5] Due to an initial error in determining the number of participants who failed attention checks, we recruited an additional set of 30 participants after our target sample size had already been reached. No analyses were conducted prior to this additional data collection. We report the results for the full dataset.

3.2. Results

3.2.1. Analytic approach

Our analytic approach was identical to that of Study 1. In Study 2, all regression analyses included random intercepts for question–answer pair.

3.2.2. Testing the perceived learning prediction

First, we replicated the results of Study 1, investigating the association between perceived learning and explanatory satisfaction for a new set of question–answer pairs (see Fig. 1). The seven measures of perceived useful learning and explanatory virtues together explained significant variance in satisfaction beyond interest and knowledge, $\chi^2(7) = 234.14$, $p < .001$. In particular, unique variance in explanatory satisfaction was predicted by perceived learning, $\beta = 0.31$, 95% CI [0.21, 0.41], $\chi^2(1) = 35.86$, $p < .001$, future utility, $\beta = 0.33$, 95% CI [0.24, 0.42], $\chi^2(1) = 49.29$, $p < .001$, simplicity, $\beta = 0.15$, 95% CI [0.06, 0.24], $\chi^2(1) = 9.98$, $p = .002$, and regularity, $\beta = 0.18$, 95% CI [0.09, 0.27], $\chi^2(1) = 15.49$, $p < .001$. Replicating Study 1, these results suggest that judgements of how much has been learned and how useful this information is expected to be are associated with explanatory satisfaction, as are judgments of how general the explanation is. In addition, like Study 1, there was no evidence that breadth explained unique variance in satisfaction, $\beta = 0.02$, 95% CI [-0.05, 0.10], $\chi^2(1) = 0.37$, $p = .54$. Diverging from Study 1, perceived information content, $\beta = 0.07$, 95% CI [-0.03, 0.18], $\chi^2(1) = 2.15$, $p = .14$, and expertise, $\beta = 0.02$, 95% CI [-0.08, 0.11], $\chi^2(1) = 0.11$, $p = .74$, did not explain unique variance in explanatory satisfaction, while perceived simplicity did. However, as reported in the supplementary material and as shown in Fig. 1, an exploratory (non-preregistered) analysis revealed that nearly all measures were again related to explanatory satisfaction in isolation. We also report in the supplementary material a preregistered analysis predicting satisfaction ratings using both explanation ratings and question ratings in a single regression model; this analysis suggests some phenomenological value to receiving an explanation that exceeds expectations about utility (see also Marvin & Shohamy, 2016). Together, these findings provide additional support for the perceived learning prediction: perceived useful learning explained unique variance in satisfaction, above and beyond the variance explained by explanatory virtues.

3.2.3. Testing the actual learning prediction

Next, we tested whether explanatory satisfaction was associated with actual learning, operationalized as multiple-choice performance. Controlling for topic-level interest and knowledge, there was no evidence that accuracy on multiple-choice questions predicted satisfaction ratings, $\beta = -0.04$, 95% CI [-0.14, 0.06], $\chi^2(1) = 0.74$, $p = .39$. Next, we decomposed multiple-choice performance into two components: performance on relevant questions and performance on irrelevant questions (each scored out of two). In a regression model including both scores as fixed effects and controlling for interest and knowledge, there was no evidence for an association between explanatory satisfaction and accurate responses concerning relevant details, $\beta = -0.10$, 95% CI [-0.21, 0.004], $\chi^2(1) = 3.58$, $p = .06$, or irrelevant details, $\beta = 0.04$, 95% CI [-0.06, 0.15], $\chi^2(1) = 0.66$, $p = .42$.

However, unsurprisingly in light of prior research (Glenberg & Epstein, 1985; Lin & Zabrucky, 1998; Maki, 1998), there was also no evidence for an association between multiple-choice performance and perceived learning. We fit a mixed-effects regression model predicting "learning" ratings (the most direct measure of perceived learning), with multiple-choice performance, interest, and knowledge as fixed effects and with random intercepts for question–answer pair. Multiple-choice performance was not a significant predictor of perceived learning, $\beta = -0.04$, 95% CI [-0.14, 0.06], $\chi^2(1) = 0.52$, $p = .47$. Moreover, multiple-choice performance did not explain additional variance in satisfaction when controlling for perceived learning (see supplementary material).

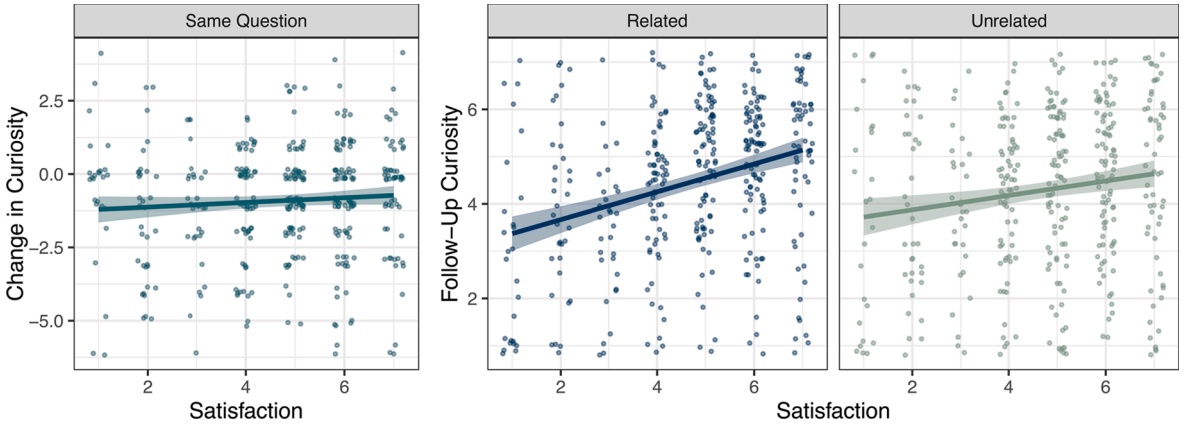

Fig. 2. Explanatory Satisfaction and Subsequent Curiosity in Study 2. The association between satisfaction and (1) participants' change in curiosity about the target question, (2) participants' curiosity about related follow-up questions, and (3) participants' curiosity about unrelated follow-up questions. Each jittered point corresponds to one participant's satisfaction rating for one explanation.

3.2.4. Testing the selective reinforcement prediction

We additionally tested whether explanatory satisfaction was related to decreased curiosity about the initial question after having read the explanation (see Fig. 2). A paired-samples *t*-test revealed that post-answer curiosity ratings were significantly lower than their initial levels, $t(395) = -9.47$, $p < .001$, $d = 0.48$. However, satisfaction was *positively* related to post-answer curiosity, $\beta = 0.19$, 95% CI [0.10, 0.29], $\chi^2(1) = 16.29$, $p < .001$, such that higher levels of explanatory satisfaction were associated with more curiosity about the question that had just been answered. We conducted an exploratory analysis predicting the change in curiosity ratings from pre-answer to post-answer, in a model with the same fixed and random effects. There was no evidence for an association between satisfaction and change in curiosity, $\beta = 0.06$, 95% CI [-0.04, 0.16], $\chi^2(1) = 1.37$, $p = .24$. These results do not support the hypothesis that satisfaction is associated with decreased curiosity about the initial question.

Next, we tested whether the association between satisfaction and subsequent curiosity is restricted to related questions, and thus does not extend to unrelated questions (see Fig. 2). As in Study 1, the three follow-up curiosity ratings related to the question–answer pair were averaged to create a "follow-up curiosity score" (Cronbach's $\alpha = 0.82$), as were the three unrelated follow-up curiosity ratings (Cronbach's $\alpha = 0.85$). We fit a regression model predicting follow-up curiosity, with explanatory satisfaction, follow-up question type (related/unrelated) and the interaction between these variables as fixed effects. Interest and knowledge were not included for this analysis, as separate topic-level interest/knowledge ratings would be relevant for predicting satisfaction for related versus unrelated follow-up questions. There was evidence for a significant interaction between satisfaction and question type, $\chi^2(1) = 7.48$, $p = .006$, which we further analyzed by fitting separate models predicting follow-up curiosity within each question type. For related questions, there was a significant, positive association between explanatory satisfaction and follow-up curiosity, $\beta = 0.32$, 95% CI [0.23, 0.41], $\chi^2(1) = 42.87$, $p < .001$. For unrelated questions, there was also a significant, positive association between explanatory satisfaction and follow-up curiosity, $\beta = 0.17$, 95% CI [0.08, 0.27], $\chi^2(1) = 11.91$, $p < .001$, but this association was weaker.

3.3. Discussion

In Study 2, we replicated the results of Study 1 in support of the perceived learning prediction: explanations that were perceived to support (useful) learning were more likely to be found satisfying, and perceived useful learning explained variance in satisfaction above and beyond explanatory virtues. However, there were some differences between Studies 1 and 2 in which specific measures explained unique variance in satisfaction. One possibility is that the different stimulus sets we used in the two studies—easy to understand explanations from a children's book versus more complicated explanations from textbooks—had differing variation along some dimensions. For example, most explanations in the Study 2 stimulus set were rated as quite high in information content, while those in Study 1 were more variable. We further discuss the need to study satisfaction across a broader set of explanations in the General Discussion.

The results of Study 2 also go beyond those of Study 1 in several ways. First, we tested the actual learning prediction, but we found no evidence for an association between explanatory satisfaction and an objective measure of learning. This result is perhaps not surprising given prior research on the calibration of comprehension (Glenberg & Epstein, 1985; Lin & Zabrucky, 1998; Maki, 1998). Indeed, we found no evidence for a significant association between perceived learning and multiple-choice performance, suggesting that limits in metacognition could explain the lack of association between satisfaction and actual learning. However, it is also possible that our measure of actual learning (a four-item multiple-choice test) has weaknesses in its ability to gauge the kind of learning that might be most relevant to explanatory understanding. While the multiple-choice questions were chosen so that a different sample of participants scored below 50% without having read the explanation, we did not include a pre-test for participants in our study. As a result, multiple-choice performance may have reflected variation in participants' prior knowledge, rather than variation in the extent to which they learned from the explanation. In addition, performance in our sample was fairly low: on average, participants answered only 60% of questions correctly (just over 2 of the 4 questions). Finally, the multiple-choice assessment only had four questions (and only two assessing relevant details), so scores might not be sufficiently fine-grained to capture variation in learning. Given these concerns, we adopt a different measure of actual learning in Study 3, and we return to the link between perceived and actual learning in Study 4.

A second way in which Study 2 went beyond Study 1 is that we tested all three components of the selective reinforcement prediction: that satisfaction should be (a) related to decreased curiosity concerning the initial question, (b) related to greater curiosity about related questions, and (c) independent of curiosity about unrelated questions. Challenging prediction (a), there was no evidence that satisfaction was related to a change in curiosity from pre-answer ratings to post-answer ratings. These results are perhaps puzzling given previous findings that children are less likely to pursue inquiry on the same question after receiving an explanatory answer (Frazier et al., 2009, 2016; Mills et al., 2019). However, these previous studies contrasted explanatory answers with non-explanatory answers (e.g., on-topic information without explanatory content, or circular explanations), while our studies include only explanatory answers that varied in the extent to which they elicited satisfaction. Additionally, while Mills et al. (2019) found that children's ratings of the quality of an answer were negatively related to subsequent inquiry on the same question, this association was not reported separately for explanatory vs. circular answers. Thus, it may be the case that any explanation judged to be explanatory decreases curiosity about the initial question relative to any question judged to be non-explanatory. Within the range of explanatory responses, however, the relation between satisfaction and ongoing curiosity about the initial question is less clear. We discuss these results further in the General Discussion.

In support of prediction (b), we replicated the positive association from Study 1 between satisfaction and curiosity about related follow-up questions. However, we found mixed support for prediction (c): satisfaction ratings were in fact positively related to curiosity about questions unrelated to the explanation, though this association was weaker than the association between satisfaction and

curiosity about related questions. When and why satisfaction may spark subsequent inquiry about unrelated questions is a topic we return to in the General Discussion.

4. Study 3

In Study 3, we sought to test the actual learning prediction with a more ecologically valid and comprehensive measure of learning. Specifically, we asked participants to recall the explanations as accurately as possible, and we evaluated the similarity between the source explanation and the explanation they recalled. This is a measure similar to that used in prior research (Frazier et al., 2016), in which children were better able to accurately reproduce explanations over non-explanations (as were adults in some cases, though recall in adults was quite high overall). Additionally, this measure has the advantage of mimicking more realistic forms of explanation transmission: explanation often involves individuals making arguments to persuade their peers (e.g., Mercier, 2016; Mercier & Sperber, 2011), or knowledgeable individuals teaching less-knowledgeable learners (e.g., Frazier et al., 2009). This transmission function of explanation (see Gwynne & Lombrozo, 2010) relies on the ability to reproduce the core content of an explanation from memory. Additionally, because explanation recall is a more constructive and demanding task, it is also likely to reflect the extent to which participants successfully understood the content of the explanation and integrated this content with their prior beliefs. Thus, in Study 3, we conduct an additional (and arguably better) test of the actual learning prediction: we attempt to replicate our results with the same multiple-choice measure of learning, but we also ask participants to reproduce the explanation from memory and assess the fidelity of their recall.

4.1. Method

4.1.1. Participants

Participants were 320 adults (158 male, 159 female, 2 other, and 1 prefer not to specify, ages 21–70) recruited from Amazon Mechanical Turk. Thirty-three additional participants completed the study but were excluded from analysis because they failed to pass two attention checks. The size of this sample was determined by doubling the sample size of an initial study, reported in the supplementary material (Study S4).

4.1.2. Materials

Study 3 used the same set of question–answer pairs as Study 2.

4.1.3. Procedure

First, participants completed the *interest/knowledge phase*, rating their interest in and knowledge of the 19 topics corresponding to the 20 question–answer pairs. Next, participants rated explanatory satisfaction for four explanations—two that were the same for all participants (the first and last rated), and two randomly selected from the remaining 18. After the *distractor phase* (which again doubled as an attention check, as in Studies 1–2), participants completed two measures of actual learning for the two randomly selected question–answer pairs. For the first measure of learning, participants completed a free recall task, for which they were provided with the question and prompted: "Please reproduce the answer you read as best you can, as if you were explaining it to another person. We're interested in what you remember about the answer, so please try to reproduce it to the best of your ability." Participants were instructed to type anything they could remember, like key words or phrases. However, if participants remembered absolutely nothing, they were instructed to check a box labelled "I remember absolutely nothing about the answer to the question," rather than typing in the provided text box. After completing the free recall task for both explanations, participants completed the second measure of learning: the same multiple-choice measure used in Study 2. Finally, participants provided their age and gender, and answered an attention check question that required them to identify one of the questions they had previously rated from a list of four distractors.

4.2. Results

4.2.1. Analytic approach

Our analytic approach was again identical to that of Study 1. Study 3 also includes mixed-effects models with categorical measures. Categorical measures were dummy coded, and unstandardized regression coefficients (*b*) are reported for these models (with any continuous measures z-scored). We indicate the reference group for each categorical measure below. All analyses include random intercepts for participant and question–answer pair.

4.2.2. Testing the actual learning prediction: multiple-choice performance

First, replicating Study 2, we tested whether multiple-choice performance predicted explanatory satisfaction, controlling for topic-level interest and knowledge. Overall multiple-choice performance (including performance on relevant and irrelevant questions) was not significantly related to explanatory satisfaction, $\beta = 0.06$, 95% CI [-0.02, 0.13], $\chi^2(1) = 2.10$, $p = .15$. In addition, there was no evidence for an association between explanatory satisfaction and learning of relevant details, $\beta = 0.04$, 95% CI [-0.04, 0.11], $\chi^2(1) = 0.99$, $p = .32$, or irrelevant details, $\beta = 0.03$, 95% CI [-0.04, 0.11], $\chi^2(1) = 0.80$, $p = .37$, controlling for interest and knowledge. As in Study 2, these results provide no evidence for an association between learning—as measured by multiple-choice performance—and explanatory satisfaction.

4.2.3. Testing the actual learning prediction: free recall

Next, we tested whether explanatory satisfaction was associated with a different metric of learning: the ability to reproduce an explanation from memory. For each reproduced explanation (excluding any text responses where participants checked the box indicating no memory for the provided explanation), we measured its similarity to the original explanation using text analysis techniques. Explanations were converted to sequences of word vectors using pre-trained fastText word embeddings (specifically, a set of 1 million word vectors trained on English Wikipedia, news datasets, and other corpora; Mikolov et al., 2018). Centroids were found for each set of word embeddings for a given explanation. The cosine similarity between the centroid of the original explanation and the centroid of the reproduced explanation was calculated for each recalled explanation. We use this measure to capture recall fidelity.

We validated this measure by having an independent coder rate (1) how similar each recalled explanation was to the original explanation, and (2) how well each recalled explanation demonstrated understanding of the original explanation, for a subset of 100 randomly sampled explanations. The coder was only provided with the original question–answer pair and the recalled answer, and was thus unaware of recall fidelity scores, satisfaction, and multiple-choice performance. Similarity and understanding were both rated on a seven-point scale, where 0 indicated no similarity or understanding, and 6 indicated complete similarity or understanding. The coder was encouraged to use the full range of the rating scale. Because the distribution of recall fidelity scores was skewed and many ties existed in the ranked human-coded data, we tested the association between human-coded similarity/understanding and recall fidelity using Kendall's rank correlation, a non-parametric measure of correlation that adjusts for tied ranks. As predicted, recall fidelity scores were significantly correlated with both similarity ratings, $\tau = 0.52$, $z = 7.23$, $p < .001$, and understanding ratings, $\tau = 0.41$, $z = 5.67$, $p < .001$. This suggests that the ranked ordering of explanations by recall fidelity is similar to the ranked ordering of explanations by human-coded similarity or understanding.

Having validated our measure of recall fidelity, we fit a regression model predicting explanatory satisfaction as a function of recall fidelity, controlling for interest and knowledge. Recall fidelity was positively related to explanatory satisfaction, $\beta = 0.12$, 95% CI [0.04, 0.21], $\chi^2(1) = 8.03$, $p = .005$ (see Fig. 3). Furthermore, explanatory satisfaction was significantly lower when no details of the explanation were successfully encoded (that is, when participants checked the box indicating they could remember absolutely nothing about the previously read explanation) relative to when participants generated a remembered explanation (i.e., reference group: box unchecked), controlling for interest and knowledge, $b = -0.20$, 95% CI [-0.37, -0.03], $\chi^2(1) = 5.20$, $p = .02$. These results suggest that explanatory satisfaction may be sensitive to actual learning, in particular to how well an explanation has been integrated with prior beliefs and can be reproduced in linguistic form.[6]

4.3. Discussion

In Study 3, we found evidence for an association between explanation recall and explanatory satisfaction. This provides some support for the actual learning prediction: explanations that are satisfying are also more likely to be accurately recalled, suggesting that more learning occurred at the time of encoding. However, like Study 2, Study 3 failed to find evidence for an association between multiple-choice performance and explanatory satisfaction. Moreover, the association between explanatory satisfaction and recall fidelity was relatively weak, and explanations that were not remembered at all were only less satisfying than recalled explanations by 0.56 scale points on average. One possibility is that recall fidelity, like multiple-choice performance, is not the most relevant measure of learning. For example, it is possible to memorize and recite an explanation without any understanding of explanatory content, and thus participants might have recalled only surface details. In conflict with this possibility, recall fidelity was significantly associated with the extent to which the recalled explanation demonstrated understanding as rated by an independent coder. In addition, participants scored similarly on multiple-choice questions assessing relevant details (53%) and irrelevant (i.e., surface-level) details (54%), suggesting that participants did not merely encode surface details of the explanation. However, it remains likely that other measures of learning—for example, how well one is able to use the explanation to make predictions, draw novel inferences, or design interventions—are better suited to assessing the learning relevant for predicting satisfaction. We further discuss the nuances of "actual learning" in the General Discussion.

Another possibility is that the weak association between explanatory satisfaction and actual learning reflects the restricted range of explanations used in the present study: all explanations were drawn from textbooks, and thus were designed to be reasonably satisfying and to promote learning. As a result, we may not have captured the full range of satisfaction and recall fidelity. In the following study, we attempt to replicate the association between explanatory satisfaction and actual learning using a broader range of good and bad explanations.

[6] This pattern of results could also arise if individuals who are better at reproducing explanations are more likely to feel satisfied with any explanation. This possibility is unlikely for two reasons. First, these analyses included random intercepts for participant, and thus controlled for baseline variation in satisfaction across participants. Second, we directly compared the association between recall fidelity and explanatory satisfaction for the same explanation to the association between recall fidelity for one explanation and explanatory satisfaction for the other explanation rated by the same participant. There was a significant interaction, $\chi^2(1) = 12.96$, $p < .001$, and the association between recall fidelity for one explanation and satisfaction for the other explanation was significant but negative, $b = -0.09$, 95% CI [-0.18, −0.01], $\chi^2(1) = 4.45$, $p = .03$. Together, this suggests that actual learning is associated with satisfaction selectively for the learned-from explanation.

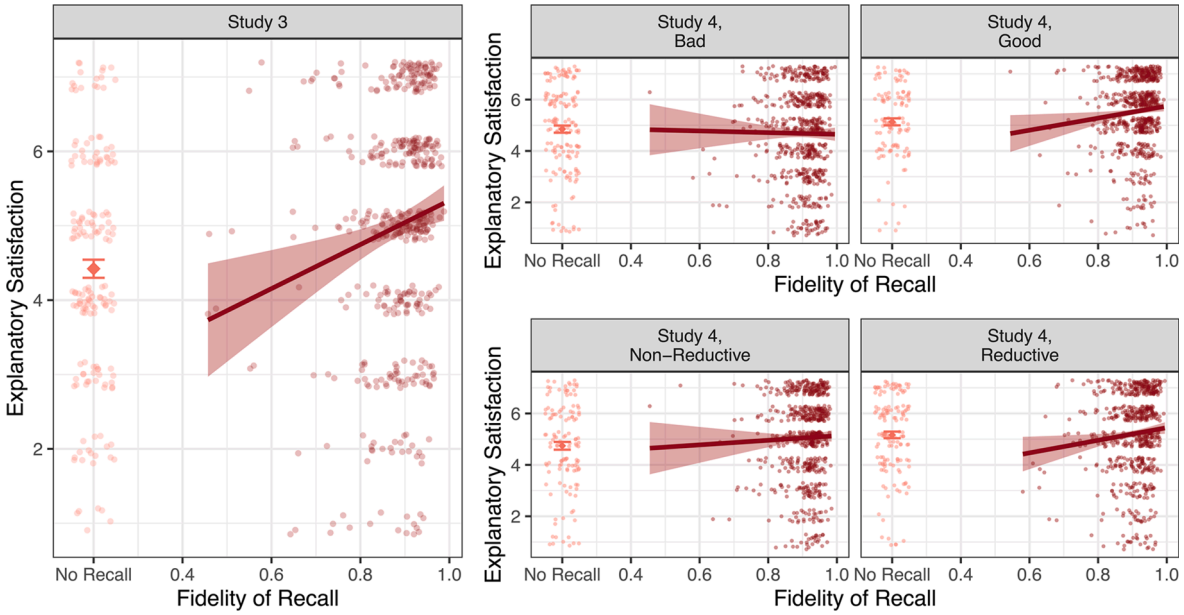

Fig. 3. Explanatory Satisfaction and Actual Learning. Association between recall fidelity (either "no recall" or cosine similarity between recalled explanation and original explanation) and explanatory satisfaction in Study 3 (textbook explanations) and Study 4 (good and bad scientific explanations; with or without reductive details).

5. Study 4

In Study 4, we had three main aims. First, we attempt to replicate the evidence found in Study 3 for the actual learning prediction. In particular, the association between satisfaction and recall fidelity was fairly weak, but recall fidelity and satisfaction were reasonably high in general. This could be because all the explanations used in Study 3 were from the same reputable sources and all contained explanatorily relevant information. In Study 4, we attempt to replicate the association between actual learning and satisfaction using a range of good (i.e., containing explanatorily relevant information) and bad (e.g., circular) explanations.

Second, we tease apart two causal interpretations of the association between satisfaction and actual learning. If, as we suggest, actual learning in part determines satisfaction, then an independent manipulation of satisfaction that provides no new explanatory content should have little or no effect on actual learning. If, on the other hand, satisfaction has a causal influence on later recall (e.g., because satisfying explanations are encoded with heightened attention), then a manipulation of satisfaction should also affect actual learning. To test this, we take advantage of prior work on reductive explanations (Hopkins et al., 2016), which shows that people prefer explanations that contain explanatorily irrelevant reductive information (e.g., neuroscience details in an explanation of a psychological phenomenon) over explanations that do not. If satisfaction has a causal influence on actual learning, we would expect the addition of reductive details to inflate both satisfaction and explanation recall. If, in contrast, satisfaction has no influence on actual learning (perhaps because actual learning instead determines satisfaction, as the aligned motivation account predicts), then the addition of reductive details should affect satisfaction but not recall.

Finally, we explore one possible implication of the aligned motivation account. From the point of view of the scientist, individuals' inflated satisfaction for explanations containing reductive information (Hopkins et al., 2016; Weisberg et al., 2008) challenges the aligned motivation account: if reductive information is not explanatorily relevant information, it should not drive satisfaction (Trout, 2008). But from the point of view of the learner, explanations with reductive information could be *perceived* as providing relevant information, and this perception could explain effects on satisfaction. For example, an explanation of a psychological phenomenon that appeals to neural mechanisms might lead learners to judge they have learned more relevant information, even if this is not objectively the case, and this judgment of perceived learning might mediate the effect of the reductive information on satisfaction. Further, this should be the case specifically for perceived learning of relevant details, not just general perceptions of learning. This hypothesis is compatible with one mechanism proposed by Hopkins et al. (2016): individuals might prefer explanations containing irrelevant reductive details because they overgeneralize from prior experiences with *informative* reductive explanations. However, to our knowledge, this hypothesis has not been tested.

Study 4 was preregistered (https://aspredicted.org/tz6gz.pdf); any departures from our preregistered analysis plan are detailed in the Results section.

5.1. Method

5.1.1. Participants

Participants were 525 adults (217 male, 300 female, 6 other, and 2 prefer not to specify, ages 18–73) recruited from Prolific. Participants were required to reside in the United States and to have completed at least 100 prior studies on Prolific with a minimum 95% approval rate. Twenty-nine additional participants completed the study but were excluded from analysis because they failed to pass two attention checks. The target sample size of 520 was determined by power analysis based on Study 3 effect sizes (for the association between recall fidelity and satisfaction), using the R package simr (P. Green & MacLeod, 2016).

5.1.2. Materials

We used a new set of question–answer pairs, drawn from Hopkins et al. (2016). Hopkins et al. developed a set of explanations in response to explanation-seeking questions about four scientific phenomena in six domains (physics, chemistry, biology, neuroscience, psychology, and social science). Each question–answer pair was preceded by a passage describing the scientific phenomenon. Each question had four explanations, which varied on two dimensions: their epistemic quality, which was either good (answered the question with explanatory information) or bad (did not answer the question, e.g., was circular); and the presence or absence of reductive information, which refers to fundamental component parts of the target phenomenon (e.g., using information from neuroscience to explain a psychological phenomenon). Critically, the explanations were designed so that the presence of reductive information added no new explanatory information to the good or circular explanations, as verified by experts in the relevant field. We used a subset of these items, drawn from three psychology phenomena and three physics phenomena. The full set of materials can be found at https://osf.io/hf3cj/.

5.1.3. Procedure

Participants were initially introduced to the task using instructions adapted from Hopkins et al. (2016): "On the following pages, you will read several passages of text, which describe various scientific findings. All the findings come from solid, replicable research; they are the kind of material you would encounter in a textbook. You will also read a question and answer about each finding. Unlike the findings themselves, the answers may range in quality." Participants were then presented with six question–answer pairs, each on a separate screen. Because several of the target phenomena were likely unfamiliar to participants (e.g., "Why does the attentional blink effect happen?"), each question–answer was preceded by a brief passage describing the phenomenon. This departs from our method in Studies 1–3, where participants were solely presented with question–answer pairs. To ensure that participants' ratings reflected their evaluation of the explanation and not of the preceding passage, all answers were presented in blue text, and each item referred to the "text in blue" as the target of evaluation. For each question–answer pair, participants rated three items on seven-point scales, presented in a random order for each participant: *relevant learning* ("To what extent has the answer to this question (the text in blue) taught you something new about [explanandum, e.g., why microwaves work faster than conventional ovens]?"), *general learning* ("To what extent has the answer to this question (the text in blue) taught you something new in general?"), and *satisfaction* ("How satisfying do you find the answer to this question (the text in blue)?").

The first two phenomena, with one drawn from physics and one drawn from psychology, were explained with a good/reductive explanation (which received the highest ratings in prior research) and a bad/non-reductive explanation (which received the lowest ratings in prior research). These were presented first to anchor participants to relatively good/bad explanations in the experimental context; ratings in response to these items were not analyzed. The subsequent four phenomena included two from psychology and two from physics, each randomly assigned to one explanation type (good/reductive, good/non-reductive, bad/reductive, bad/non-reductive) and presented in a random order.

After rating the six explanations, participants completed a shortened version of the *distractor phase* used in the prior studies (five math problems, with participants excluded from analyses with fewer than four correct answers). Subsequently, participants completed the recall task from Study 3 for three explanations: the third, fourth, and fifth previously read, presented in the same order (i.e., the first three explanations of interest). We did not ask participants to recall the final explanation to avoid recency effects.

Finally, participants completed a second attention check, identifying a question they had rated and recalled from a list of three distractors, then provided their age and gender.

5.2. Results

5.2.1. Analytic approach

We followed the same analytic approach as in Studies 1–3. For all models, we fit random intercepts for participant and for scientific phenomenon, nested within domain. In several cases, these models did not converge or resulted in singular fit; in these cases, we modeled scientific phenomenon with random intercepts without accounting for domain. This modification was not preregistered, but it produced comparable estimates of all fixed effects. Categorical measures were dummy coded; the reference group for each measure is indicated below.

Recall fidelity was calculated as in Study 3. For explanations including reductive details, we calculated recall fidelity twice: once where the "original" explanation was the explanation participants had read, and once where the "original" explanation was that explanation but excluding reductive details. This was to ensure that participants were not penalized for their inability to remember technical words, nor were recall fidelity scores inflated when participants succeeded in remembering those words. Both measures of recall fidelity produced comparable results, so we report recall fidelity scores calculated from the full explanation below.

5.2.2. Manipulation checks

First, we tested whether our manipulations of explanation quality and reductive detail shifted explanatory satisfaction as intended. A model predicting satisfaction as a function of explanation quality (reference group: bad), inclusion of reductive details (reference group: not included), and the interaction between these variables succeeded in finding a significant effect of reductive information on satisfaction.[7] Specifically, the best fitting model included explanation quality, $b = 0.46$, 95% CI [0.39, 0.53], $\chi^2(1) = 162.13$, $p < .001$, and inclusion of reductive details, $b = 0.10$, 95% CI [0.03, 0.16], $\chi^2(1) = 7.39$, $p = .007$, as fixed effects, with no interaction, $\chi^2(1) = 0.02$, $p = .88$ (see Fig. 4). We thus moved on to test our new predictions.

5.2.3. Testing the actual learning prediction

We first attempted to replicate the association between recall fidelity and satisfaction for explanations that did not contain reductive information, as it was unclear whether reductive information would influence the predicted association. There was no evidence for an association between recall fidelity and satisfaction within non-reductive explanations, $\beta = 0.04$, 95% CI [-0.04, 0.11], $\chi^2(1) = 0.85$, $p = .36$ (see Fig. 3). Next, we tested whether the association between recall and satisfaction was moderated by explanation quality or the inclusion of reductive details, fitting a model including explanation quality, reductive details, and recall, with an interaction between each explanation manipulation and recall. There was no evidence that the association between recall fidelity and satisfaction differed as a function of reductive details, $\chi^2(1) = 0.98$, $p = .32$. However, the interaction between recall fidelity and explanation quality was significant, $\chi^2(1) = 4.56$, $p = .03$. Within good explanations (reductive or non-reductive), recall fidelity was a significant predictor of satisfaction, $\beta = 0.08$, 95% CI [0.003, 0.16], $\chi^2(1) = 4.13$, $p = .04$. However, within bad explanations (reductive or non-reductive), recall fidelity was unrelated to satisfaction, $\beta = -0.03$, 95% CI [-0.11, 0.05], $\chi^2(1) = 0.61$, $p = .43$ (see Fig. 3).

We repeated these analyses using a different measure of recall: whether participants reported they could not remember any details of the explanation. We erroneously preregistered that this analysis would be conducted on the full dataset (excluding moderation by reductive details and explanation type). Our intention was to repeat the analyses above, and we report those results here. There was no evidence for an association between satisfaction with non-reductive explanations and participants' indication that they remembered nothing about the explanation (reference group: box unchecked), $b = -0.17$, 95% CI [-0.37, 0.02], $\chi^2(1) = 3.25$, $p = .07$ (see Fig. 3). In addition, the association between satisfaction and no recall was not moderated by inclusion of reductive details, $\chi^2(1) = 2.00$, $p = .16$. However, this association was moderated by explanation quality, $\chi^2(1) = 9.73$, $p = .002$. Within good explanations (reductive or non-reductive), satisfaction was significantly lower for non-recalled explanations than for recalled explanations, $b = -0.23$, 95% CI [-0.43, -0.03], $\chi^2(1) = 5.29$, $p = .02$. However, within bad explanations, there was no evidence for a difference in satisfaction based on successful recall, $b = 0.15$, 95% CI [-0.03, 0.32], $\chi^2(1) = 2.58$, $p = .11$. Together, these results provide some support for the actual learning prediction, but only within higher-quality explanations.[8]

What could explain these results? As in Study 2, we tested whether the null results could reflect limits in metacognition (Glenberg & Epstein, 1985; Lin & Zabrucky, 1998; Maki, 1998). We fit an exploratory mixed-effects regression model predicting perceived relevant learning with recall fidelity as a fixed effect. There was no evidence for a significant association between recall fidelity and perceived learning across all explanations, $\beta = 0.01$, 95% CI [-0.04, 0.07], $\chi^2(1) = 0.22$, $p = .64$. However, in an additional exploratory regression model including recall fidelity, explanation quality, and their interaction, there was evidence for a significant interaction, $\chi^2(1) = 8.23$, $p = .004$. Among good explanations, there was a significant positive association between recall fidelity and perceived learning, $\beta = 0.08$, 95% CI [0.01, 0.16], $\chi^2(1) = 4.66$, $p = .03$. Among bad explanations, there no evidence for an association between recall fidelity and perceived learning, $\beta = -0.08$, 95% CI [-0.16, 0.002], $\chi^2(1) = 3.64$, $p = .06$. These results closely mirror the associations found between recall and satisfaction and could suggest that satisfaction only tracks actual learning when judgments of learning are calibrated. Supporting this possibility, the interaction between recall fidelity and explanation quality was no longer a significant predictor of satisfaction when perceived learning was also included as a fixed effect, $\chi^2(1) = 0.31$, $p = .58$. Instead, perceived learning was a significant predictor of satisfaction above and beyond recall fidelity, explanation quality, and their interaction, $b = 0.55$, 95% CI [0.50, 0.59], $\chi^2(1) = 475.86$, $p < .001$. This analysis was exploratory.

5.2.4. Explanatory vices and actual learning

Next, we tested whether explanation quality or the presence of reductive information influenced actual learning. If actual learning

[7] Our findings only partially replicate Hopkins et al.'s (2016). Specifically, we failed to find a significant interaction between explanation quality and the inclusion of reductive details, as found in the Amazon Mechanical Turk sample (but not the undergraduate sample) in Hopkins et al. (2016) and as found by Weisberg et al. (2008). It remains an open question whether reductive details selectively salvage bad explanations or improve the quality of all explanations. However, our failure to detect this interaction has little bearing on the subsequently reported results.

[8] One potential concern is that participants could have recalled details from the descriptive passage presented prior to the question–answer pair, rather than the answer itself. In fact, exploratory analyses suggested this was not the case. We computed the similarity between the recalled text and the initial descriptive passage using the same method described for recall fidelity, and we compared this measure of "description recall fidelity" against "explanation recall fidelity." We fit a model with score type (description/explanation) as a fixed effect predicting the recall fidelity score. Description recall fidelity was significantly lower than explanation recall fidelity, $b = -0.01$, 95% CI [-0.02, −0.01], $\chi2(1) = 74.58$, $p < .001$, suggesting that participants' recalled explanations were more similar to the explanations than to the descriptive passages. In addition, among good explanations (where explanation recall fidelity was significantly associated with satisfaction), there was no evidence that description recall fidelity predicted satisfaction, $\beta = 0.04$, 95% CI [-0.04, 0.12], $\chi^2(1) = 0.94$, $p = .33$. This suggests that participants recalled the provided explanations rather than the initial descriptive passages, and this validates that our measure of recall fidelity is sensitive to fairly fine-grained distinctions in participants' recall.

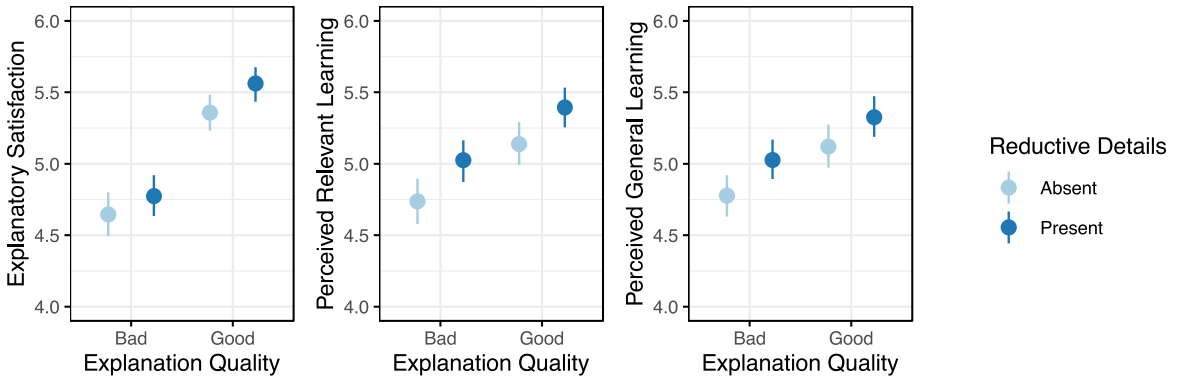

Fig. 4. Manipulating Perceived Learning and Satisfaction. Study 4 effect of explanation quality and reductive details on explanatory satisfaction, perceived relevant learning, and perceived general learning. Points indicate the average rating on scale ranging from 1 to 7; error bars indicate 95% confidence intervals.

partly determines satisfaction, as the aligned motivation account predicts, then an independent manipulation of satisfaction that does not add genuine explanatory information (e.g., the inclusion of reductive details) should not influence actual learning. If, on the other hand, the causal relationship is reversed—from satisfaction to actual learning—then manipulations of satisfaction should also influence actual learning (e.g., learning from reductive explanations should be better than learning from non-reductive explanations). In a mixed-effects regression model predicting recall fidelity, there was no evidence for an interaction between explanation quality and reductive details, $\chi^2(1) = 1.18$, $p = .28$. In a model excluding the interaction term, there was evidence for an effect of explanation quality, $b = 0.23$, 95% CI [0.14, 0.32], $\chi^2(1) = 26.47$, $p < .001$, with participants recalling good explanations with higher fidelity than bad explanations. However, the effect of reductive details above and beyond explanation quality was not significant (and opposite the predicted direction), b = -0.08, 95% CI [-0.17, 0.01], $\chi^2(1) = 3.22$, $p = .07$. To quantify the strength of evidence against the null hypothesis that reductive details have no effect on recall fidelity, we fit a Bayesian mixed-effects regression model predicting recall fidelity with reductive details as a fixed effect, using the R package rstanarm (Goodrich et al., 2020) and default weakly informative priors. We computed the Bayes factor in favor of the point-null (i.e., that the regression coefficient is exactly zero) using the Savage-Dickey method (Wagenmakers et al., 2010) with the R package bayestestR (Makowski et al., 2019). The evidence in favor of the null was strong, BF = 28.35. Thus, our manipulation of reductive details did not influence actual learning, suggesting that actual learning is not determined by satisfaction. However, these results are difficult to interpret, given the lack of overall correlation between actual learning and satisfaction.

5.2.5. Does perceived learning explain explanatory vices?

Finally, we tested one possible implication of the perceived learning prediction: that individuals fall prey to explanatory vices precisely when they believe they have learned from the misleading information. As noted already, a model predicting satisfaction as a function of explanation quality, inclusion of reductive details, and the interaction between these variables revealed that explanations were judged more satisfying when they were good versus bad and when they contained versus omitted reductive information. Likewise, the best fitting model for predicting perceived relevant learning included explanation quality, $b = 0.21$, 95% CI [0.14, 0.29], $\chi^2(1) = 35.36$, $p < .001$, and inclusion of reductive details, $b = 0.17$, 95% CI [0.10, 0.24], $\chi^2(1) = 21.76$, $p < .001$, as fixed effects, with no interaction, $\chi^2(1) = 0.18$, $p = .67$ (see Fig. 4).

Building on these results, we fit two mediation models using the R package lavaan (Rosseel, 2012), to test whether perceived relevant learning mediated the association between explanation quality and satisfaction, and, more critically, inclusion of reductive details and satisfaction.[9] We evaluated the statistical significance of all paths using bias-corrected bootstrap confidence intervals, with 95% confidence intervals excluding zero indicating statistical significance. All parameters are reported as standardized estimates. We preregistered a simple mediation model that does not account for the clustering of observations within participants and scientific phenomenon. However, it is more appropriate to fit a multi-level structural equation model (SEM). We report below the results of a multi-level SEM with all mediation paths at the level of observations (level 1), and with variances and covariances for endogenous variables (i.e., perceived learning and satisfaction) at the level of participants (level 2). As lavaan can account for only one clustering variable, we centered perceived learning and satisfaction within scientific phenomena (to control for this additional clustering variable) before fitting the model. We report the results of the preregistered mediation model in the supplementary material; all

[9] The conclusions from mediation analyses are valid only if a number of assumptions are met (MacKinnon, 2008). For example, there must be no omitted variables that cause both the mediator (perceived learning) and the outcome variable (satisfaction), and the outcome variable (satisfaction) must not be causally related to the mediator variable (perceived learning). In our case, both assumptions are potentially problematic, as little is known about the causal relation between perceived learning, satisfaction, and other variables (though we explore the causal relation between perceived learning and satisfaction in Study 5). Therefore, the results of all mediation analyses reported here should be taken as tentative and subject to further investigation (for further problems associated with mediation analysis, see Bullock et al., 2010; D. P. Green et al., 2010).

conclusions from the preregistered analysis were identical to those presented below.

For explanation quality, the indirect effect through perceived learning was significant, $\beta = 0.19$, 95% CI [0.13, 0.25]. The remaining direct effect was also significant, $\beta = 0.57$, 95% CI [0.48, 0.67]. In other words, perceived learning partially mediated the effect of explanation quality on satisfaction. For inclusion of reductive details, the indirect effect through perceived learning was significant, $\beta = 0.16$, 95% CI [0.09, 0.23], and the remaining direct effect was not significant, $\beta = -0.001$, 95% CI [-0.10, 0.10]. In other words, perceived learning fully mediated the effect of reductive details on satisfaction.

These results are suggestive, but it is also possible that general learning—rather than relevant learning—explains the association between explanation type and satisfaction. Indeed, perceived general learning was also best predicted by explanation quality, $b = 0.18$, 95% CI [0.11, 0.25], $\chi^2(1) = 27.07$, $p < .001$, and inclusion of reductive details, $b = 0.15$, 95% CI [0.08, 0.21], $\chi^2(1) = 17.39$, $p < .001$, as fixed effects, with no interaction, $\chi^2(1) = 0.41$, $p = .52$ (see Fig. 4). To rule out this possibility, we fit a multiple mediation model following the procedure suggested by Preacher and Hayes (2008), including separate indirect effects for perceived relevant learning and for perceived general learning, and controlling for the covariance between these potential mediators. We directly contrasted these mediators by defining a model parameter for the contrast between the two indirect effects. Again, we fit two models, one testing mediation of the explanation quality effect, and the other testing mediation of the reductive detail effect (see Fig. 5). We again report multi-level mediation models here, but preregistered simple mediation models are included in the supplementary material. For explanation quality, the indirect effect through relevant learning was significant, $\beta = 0.20$, 95% CI [0.13, 0.27], but the indirect effect through general learning was not, $\beta = -0.01$, 95% CI [-0.03, 0.01]. Moreover, the contrast between these two indirect effects was significant, $\beta = 0.21$, 95% CI [0.13, 0.29]. The remaining direct effect of explanation quality on satisfaction was also significant, $\beta = 0.57$, 95% CI [0.48, 0.67], indicating partial mediation. Likewise, for reductive details, the indirect effect through relevant learning was significant, $\beta = 0.17$, 95% CI [0.09, 0.24], but the indirect effect through general learning was not, $\beta = -0.01$, 95% CI [-0.03, 0.01]. Moreover, the contrast between these two indirect effects was significant, $\beta = 0.17$, 95% CI [0.09, 0.26]. The remaining direct effect of reductive details on satisfaction was not significant, $\beta = -0.001$, 95% CI [-0.10, 0.10], indicating full mediation. In summary, perceived *explanatorily relevant* learning partially mediated the effect of explanation quality and fully mediated the effect of reductive details on satisfaction, and general learning did not play an additional role above and beyond relevant learning. Together, these results suggest that reductive information—and perhaps explanatory vices more generally—specifically inflate explanatory satisfaction when their presence leads to the belief that relevant information has been gained.

5.3. Discussion

In Study 4, we failed to replicate the association between satisfaction and explanation recall across the full set of explanations, counter to the actual learning prediction. However, recall was associated with satisfaction specifically for "good" explanations, which contained explanatorily relevant details. Further, exploratory analyses revealed that explanation recall was similarly unrelated to perceived learning across the full set of explanations but was specifically related to perceived learning for good explanations. While the latter results emerged from exploratory analyses and thus should be treated as tentative, they suggest potential limits on when satisfaction tracks actual learning. Specifically, satisfaction might only track actual learning through perceived learning, and thus be bound by metacognitive calibration. In the current study, participants showed some metacognitive calibration for good explanations (e.g., those that provide a causal mechanism), but not for bad explanations (e.g., circular explanations). Future research is needed to further test this possibility.

In addition, we found evidence against the alternative hypothesis that satisfaction determines actual learning: though satisfaction was influenced by the inclusion of reductive details, actual learning was not. Consequently, it is likely that in cases where actual learning and satisfaction *are* related, this is not because satisfaction itself encourages more careful encoding, and thus facilitates better recall.

Finally, we found that perceived learning—specifically about relevant information—partially mediated the effect of explanation quality on satisfaction and fully mediated the effect of reductive information on satisfaction. In other words, the preference for explanations that contained "explanatorily irrelevant" reductive details was fully explained by participants' belief that those explanations in fact had taught them new, relevant information. This goes beyond prior work in suggesting a possible mechanism behind the preference for reductive explanations.

6. Study 5

In Study 5, we test a final key prediction of the aligned motivation account: that learning is causally responsible for satisfaction. Having established in Studies 1–4 that perceptions of learning appear to be a stronger predictor of satisfaction than actual learning, we focus on perceived learning as a determinant of satisfaction in Study 5. To address the causal connection between perceived learning and satisfaction, we develop a manipulation of perceived learning that is independent of explanation content: participants take a phony "assessment," which purportedly predicts how well a given participant will learn from a given explanation. Participants then read a series of explanations that are randomly labeled as "high learning" or "low learning," allegedly based on the predictions of the assessment (we refer to these as predicted-high explanations and predicted-low explanations, respectively). Prior research shows that expectations about future memory are malleable and susceptible to the influence of explicit beliefs about memorability (Mueller et al., 2014; Mueller & Dunlosky, 2017). Analogously, we expect this manipulation of participants' explicit beliefs about learning to impact their perceptions of learning. Consequently, if perceived learning is causally responsible for satisfaction, we would expect participants to report higher satisfaction with predicted-high explanations and lower satisfaction with predicted-low explanations.

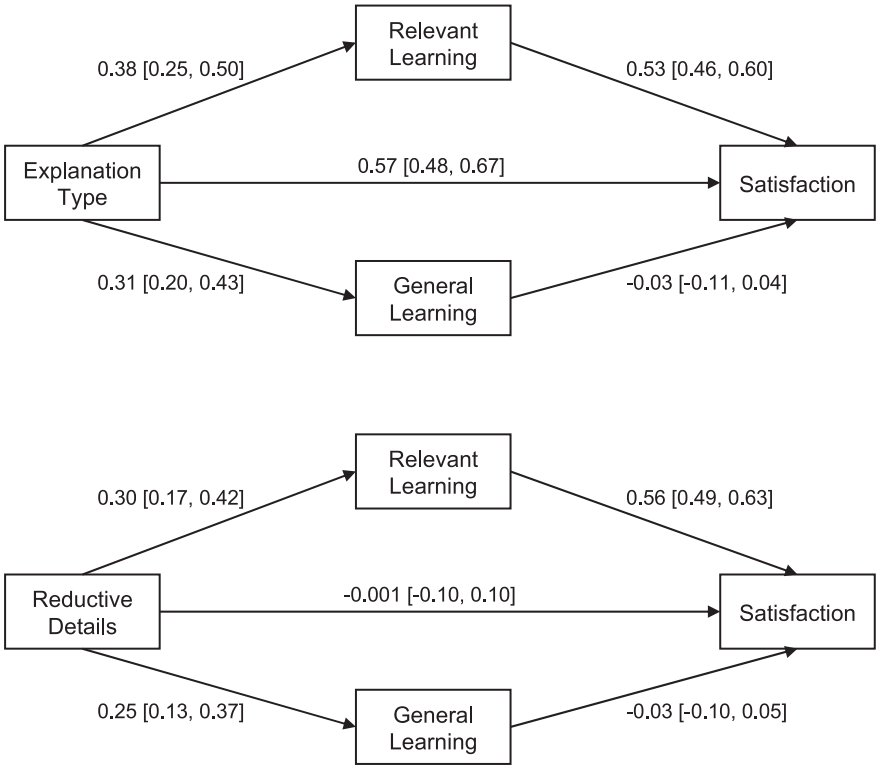

Fig. 5. Mediation Analyses, Study 4. Study 4 mediation models (fit using SEM) testing mediation by perceived relevant learning and perceived general learning. Results were consistent with the proposed mediation model: the effects of explanation type and reductive details on satisfaction were both mediated by perceived relevant learning rather than perceived general learning.

We also test the reverse causal direction by manipulating satisfaction (using a similar phony assessment) and measuring perceived learning. The aligned motivation account predicts that satisfaction is causally determined by perceived learning, but the reverse is also plausible: individuals might use feelings of satisfaction to gauge how much they believe they learned from an explanation. The aligned motivation account does not necessarily exclude the possibility of a bidirectional causal relationship between perceived learning and satisfaction. However, testing the reverse causal direction provides further insight into how learning and satisfaction are connected and together shape explanatory inquiry.

Study 5 was preregistered (https://aspredicted.org/4vv6u.pdf); any departures from our preregistered analysis plan are detailed in the Results section.

6.1. Method

6.1.1. Participants

Participants were 204 adults (72 male, 127 female, and 5 other, ages 18–78) recruited from Prolific. Participants were required to reside in the United States and to have completed at least 100 prior studies on Prolific with a minimum 95% approval rate. Eleven additional participants completed the study but were excluded from analysis because they failed to pass an attention check. The target sample size of 204 was determined by power analysis based on pilot data effect sizes, using the R package simr (P. Green & MacLeod, 2016).

6.1.2. Materials

In the main task, we used the same set of 20 question–answer pairs as used in Study 1, drawn from *1000 Questions & Answers Factfile* (Kerrod et al., 2006). We also used an additional 16 question–answer pairs drawn from the same source for the phony assessment. The full set of materials can be found at https://osf.io/hf3cj/.

6.1.3. Procedure

Participants were randomly assigned to the learning-assessment condition ($N = 101$) or the satisfaction-assessment condition ($N = 103$). Participants were initially instructed as follows:

> This study is part of a large-scale effort to better understand how people *[learn from explanations/evaluate explanations]*. Research shows that there are individual differences in the kinds of explanations people *[learn from best/find most satisfying]*. For example, some people *[learn best from/are most satisfied with]* explanations that use examples to illustrate a point, while other people *[learn best from/are most satisfied with]* explanations that discuss abstract principles. We have recently developed an assessment that can capture these individual differences and predict which explanations a given individual will *[learn from best/find most satisfying]*. In the first part of this study, you will take this assessment. Based on your answers, we will determine the kinds of explanations you are most likely to *[learn from/find satisfying]* and least likely to *[learn from/find satisfying]*. In the second part of this study, you will read some explanations and rate them on several dimensions.

The text in brackets varied according to whether a participant was in the learning-assessment condition or the satisfaction-assessment condition. Italics are included here for emphasis but were not presented to participants.

Following these instructions, participants completed the phony assessment. On each trial, participants were shown a randomly ordered pair of explanations from *1000 Questions & Answers Factfile* (Kerrod et al., 2006) on similar topics (e.g., two explanations about submarines), drawn from a larger set of eight paired explanations. For each pair of explanations, participants in the learning-assessment condition responded to the prompt, "Which explanation do you feel you learned more from?" and participants in the satisfaction-assessment condition responded to the prompt, "Which explanation do you feel is more satisfying?". Subsequently, participants were presented with the selected explanation from that trial and were prompted to select the sentence that contributed most to their learning/satisfaction. This task was repeated on six trials.

After completing the phony assessment, participants were instructed that they would read some new explanations that the assessment had classified as high learning/satisfaction and low learning/satisfaction. Upon reading each explanation, they would rate the manipulated construct (i.e., perceived learning in the learning-assessment condition; satisfaction in the satisfaction-assessment condition). In addition, we told participants, "You will also assess each explanation on several other dimensions. Our assessment was not designed to predict these other dimensions, but we are interested in your responses."

Finally, each participant was presented with 10 explanations, randomly selected from the full set of 20. Each explanation was accompanied by the phony assessment results (e.g., "Assessment results: **high** learning"), with five predicted-high explanations and

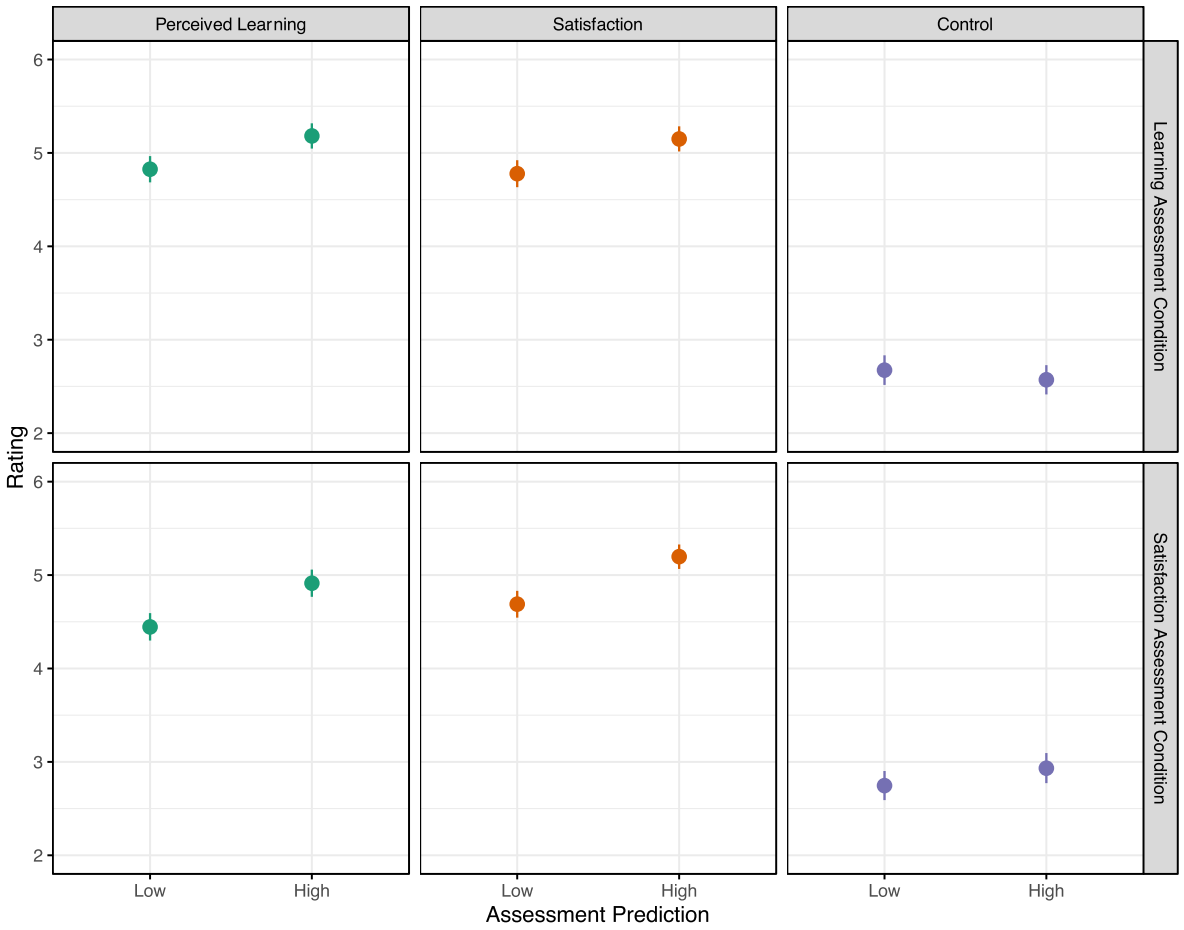

Fig. 6. Effects of manipulation on perceived learning, satisfaction, and control. Results of Study 5 manipulations of learning (top panels) and satisfaction (bottom panels). Points indicate the average rating on scale ranging from 1 to 7; error bars indicate 95% confidence intervals.

five predicted-low explanations, presented in separate blocks. Participants were randomly assigned to read predicted-high explanations first ($N = 100$) or predicted-low explanations first ($N = 104$). After each block of five explanations, participants indicated how well they thought the assessment had predicted their learning/satisfaction so far on a seven-point scale (i.e., "manipulation endorsement").

For each explanation, participants first responded to a manipulation check question (learning-assessment condition: "To what extent did you learn from this explanation?"; satisfaction-assessment condition: "How satisfying do you find this explanation?") on a seven point scale, with endpoints marked "Didn't learn at all"/"Not at all satisfying" (1) and "Learned a lot"/"Very satisfying" (7). On the following page (with the assessment results and explanation still visible, but the manipulation check question no longer visible), participants responded to two items in a random order: the unmanipulated construct (satisfaction in the learning-assessment condition; learning in the satisfaction-assessment condition), and a control question ("How much does this explanation remind you of a movie or TV show that you've seen?," endpoints marked "Not at all" and "A lot"). The control question was included to obscure the experiment's purpose to participants. In addition, it allowed us to test the deflationary hypothesis that any effect of the assessment manipulation is due to anchoring on the higher or lower ends of the scale.

Finally, participants completed an attention check question, selecting four of the ten questions explained during the main task from a list including four additional distractors, and they provided demographic information (age and gender). Participants were then debriefed about the purpose of the study and the nature of the phony assessment.

6.2. Results

6.2.1. Analytic approach

We followed the same analytic approach as in Studies 1–4. For all models, we fit random intercepts for participant and for question–answer pair. Categorical measures were dummy coded for regression analyses; the reference group for each measure is indicated below. As preregistered, all analyses were conducted separately for the learning-assessment condition and the satisfaction-assessment condition.

6.2.2. Manipulation checks

First, we tested whether the assessment manipulation indeed influenced perceived learning (in the learning-assessment condition) and satisfaction (in the satisfaction-assessment condition). Within each condition, we fit a mixed-effects regression model predicting the manipulation check measure (perceived learning or satisfaction), with assessment-predicted quality (reference group: predicted-low) as a fixed effect. As predicted, perceived learning was significantly higher for predicted-high explanations than predicted-low explanations in the learning-assessment condition, $b = 0.22$, 95% CI [0.13, 0.32], $\chi^2(1) = 20.33$, $p < .001$, and satisfaction was significantly higher for predicted-high explanations than predicted-low explanations in the satisfaction-assessment condition, $b = 0.33$, 95% CI [0.22, 0.43], $\chi^2(1) = 37.38$, $p < .001$ (see Fig. 6).

6.2.3. Does perceived learning cause satisfaction?

Next, we tested whether our manipulation of perceived learning influenced satisfaction (in the learning-assessment condition). First, we fit a mixed-effects regression model predicting satisfaction, with assessment-predicted learning (reference group: predicted-low) as a fixed effect. In support of the hypothesized causal relation, satisfaction was significantly higher when participants were told that an explanation would be high in learning as opposed to low, $b = 0.24$, 95% CI [0.14, 0.33], $\chi^2(1) = 22.71$, $p < .001$ (see Fig. 6). Next, we attempted to rule out the possibility that this effect was merely a result of anchoring on the higher or lower ends of the response scale. To do so, we fit a mixed-effects regression model with measure (satisfaction/control, reference group: control), assessment-predicted learning, and their interaction as fixed effects. The interaction effect was significant, $\chi^2(1) = 13.22$, $p < .001$, indicating that assessment-predicted learning had a differential effect on participants' responses to the control measure relative to the satisfaction measure. Indeed, there was no evidence that responses to the control measure differed as a function of assessment-predicted learning, $b = -0.04$, 95% CI [-0.12, 0.05], $\chi^2(1) = 0.71$, $p = .40$ (see Fig. 6).

6.2.4. Does satisfaction cause perceived learning?

Next, we repeated the above analyses in the satisfaction-assessment condition, in order to test whether satisfaction causes perceived learning. In support of this causal relation, perceived learning was significantly higher when participants were told that an explanation would be high in satisfaction as opposed to low, $b = 0.29$, 95% CI [0.20, 0.39], $\chi^2(1) = 34.30$, $p < .001$. In addition, this effect was selective: the interaction between measure (perceived learning vs. control) and assessment-predicted satisfaction was significant, $\chi^2(1) = 4.47$, $p = .03$. Responses to the control measure were significantly higher when assessment-predicted satisfaction was higher, $b = 0.11$, 95% CI [0.02, 0.19], $\chi^2(1) = 6.32$, $p = .01$, but this effect was weaker than the effect of assessment-predicted satisfaction on perceived learning (see above).

Our preregistration specified that we would test whether perceived learning mediated the causal relation between assessment-predicted learning and satisfaction (in the learning-assessment condition), as well as whether satisfaction mediated the causal relation between assessment-predicted satisfaction and learning (in the satisfaction-assessment condition). However, one assumption of mediation analysis is that the mediator causes the outcome variable, with no reverse causal relation (MacKinnon, 2008). Because we found evidence for a bidirectional causal relation between perceived learning and satisfaction, these mediation analyses are not valid. However, we report these preregistered analyses in the supplementary material.

6.2.5. Does manipulation endorsement moderate effects?

Finally, we tested whether participants' endorsement of the assessment's predictions moderated the effects reported above. In the learning-assessment condition, we fit a mixed-effects regression model predicting satisfaction, with assessment-predicted learning, manipulation endorsement (averaged across the two manipulation endorsement ratings), and their interaction as fixed effects. There was evidence for a significant interaction, with each one standard deviation increase in manipulation endorsement predicting a 0.21 increase in the difference in z-scored satisfaction between predicted-low explanations and predicted-high explanations, $b = 0.21$, 95% CI [0.11, 0.30], $\chi^2(1) = 17.92$, $p < .001$. We fit an analogous model in the satisfaction-assessment condition, revealing similar results: there was evidence for a significant interaction, with each one standard deviation increase in manipulation endorsement predicting a 0.29 increase in the difference in z-scored perceived learning between predicted-low explanations and predicted-high explanations, $b = 0.29$, 95% CI [0.20, 0.39], $\chi^2(1) = 35.17$, $p < .001$.

6.3. Discussion

In Study 5, we tested whether perceived learning is causally related to satisfaction, as the aligned motivation account predicts. In support of this prediction, we found that a manipulation of perceived learning affected individuals' reported satisfaction. This effect was unlikely to be due to low-level features of the task (e.g., anchoring on the higher or lower ends of the scale according to the instruction that learning should be high or low), as this effect did not extend to an unrelated control measure.

In addition, we found support for the reverse causal relation: a manipulation of satisfaction with an explanation affected individuals' reported perceptions of learning. This is not predicted by the aligned motivation account, but this is also not ruled out by the aligned motivation account: satisfaction might be partly determined by perceptions of learning, yet at the same time inform those same perceptions. Alternatively, satisfaction might be determined by perceptions of learning when perceptions of learning are more salient than feelings of satisfaction, and the reverse might be true when feelings of satisfaction are more salient than perceptions of learning. We discuss these and other possibilities in the General Discussion.

7. General discussion

Why do learners experience satisfaction upon receiving some explanations but not others? In the present research we hypothesized that explanatory satisfaction is aligned with (useful) learning, such that we experience satisfaction selectively for explanations that support learning of query-relevant information. This aligned motivation account (in contrast to the brute motivation account) makes three key predictions: that explanatory satisfaction should track how much has been learned from an explanation (actual learning prediction), that explanatory satisfaction should track subjective judgments of learning (perceived learning prediction), and that explanatory satisfaction should selectively motivate subsequent inquiry (selective reinforcement prediction). Across four studies, we found mixed support for these predictions.

Support for the actual learning prediction was tenuous. We found in Study 3 that explanatory satisfaction was related to the fidelity with which an explanation was recalled, and we found in Study 4 that satisfaction and recall fidelity were related among good explanations. However, we did not find evidence for an association between explanatory satisfaction and a multiple-choice learning assessment in Studies 2–3, and we did not replicate the association between recall fidelity and satisfaction across both good and bad explanations in Study 4. This failure to find a stable association between explanatory satisfaction and actual learning plausibly reflects the limits of metacognition, and it is unsurprising in this light: in Study 2 and among bad explanations in Study 4, we also failed to find associations between perceived learning and actual learning. In contrast, among good explanations in Study 4, there was an association between perceived learning and actual learning.

Support for the perceived learning prediction was relatively strong. We found in Studies 1 and 2 that explanatory satisfaction was predicted by perceptions of useful learning, above and beyond judgments of several explanatory virtues. In particular, controlling for other measures, significant variance in explanatory satisfaction was explained by judgments of how much was learned from the explanation and how useful the explanation would be in the future, as well as how much information the explanation contained (in Study 1 only). Consistent with prior research, several explanatory virtues explained unique variance in satisfaction, as well: the extent to which the explanation picked out a general pattern or regularity, whether the explanation seemed to be expert-produced (in Study 1 only), and how simple the explanation was (in Study 2 only). These findings reveal that explanations that are "good" in the sense that they are judged satisfying are also "good" in the sense that they are judged to support the learning of useful information. In addition, we found in Study 5 that perceptions of learning are partly causally responsible for variation in satisfaction, though we also found evidence for the reverse causal direction.

Support for the selective reinforcement prediction was mixed. In Studies 1–2, we found that explanatory satisfaction predicted subsequent curiosity about questions other than the target question. Moreover, this association was selective: satisfying answers were associated with greater curiosity about related follow-up questions, with a weaker association for unrelated follow-up questions. However, in Study 2, we did not find evidence that satisfaction was related to decreased curiosity about the initial question.

In the supplementary material, we tested one additional prediction of the aligned motivation account: that participants should be well-calibrated in their expectations about learning, so that curiosity reliably guides inquiry towards explanations that support (perceived) useful learning. Across several studies, we found evidence in support of this prediction: expected learning from the answer to a particular question (e.g., "Why do some stars explode?") was significantly associated with perceived learning when the answer was received, both within and between subjects.

Finally, we explored one possible implication of our findings: that individuals might fall prey to "explanatory vices" that contain no

new explanatory content precisely when these "vicious" explanations appear to fulfill epistemic goals. In Study 4, we found evidence in support of this prediction: perceived relevant learning (above and beyond perceived general learning) fully mediated the effect of irrelevant reductive information on satisfaction.

7.1. Implications

In combination with prior research (Liquin & Lombrozo, 2020a), the results from these studies provide evidence that the phenomenological states associated with explanatory inquiry—explanation-seeking curiosity and explanatory satisfaction—play a role in guiding learning. However, explanatory satisfaction is only partially aligned with epistemic success. On the one hand, our results suggest that satisfaction tracks actual learning as well as humans are consciously able: through perceptions of learning. (Relatedly, curiosity is more closely related to perceived prior knowledge than to actual prior knowledge; Wade & Kidd, 2019; see also Martí et al., 2018.) This correspondence with (perceived) learning is consistent with the aligned motivation account. On the other hand, the divergence between perceived and actual learning implies that satisfaction will often guide inquiry away from actual possibilities for learning—a key prediction of the brute motivation account. Our data thus suggest that the correct account of explanatory phenomenology falls somewhere in between the aligned motivation account and the brute motivation account: curiosity and satisfaction indeed track actual useful learning, but only when perceptions of learning are accurate.

There are at least three ways to understand this result. First, it could be that explanatory satisfaction is merely a consequence of perceived learning, but because satisfaction is rewarding, it ultimately plays a role in reinforcing and motivating subsequent inquiry that maximizes perceived learning. A second and stronger variant of this view could hold that explanatory satisfaction has the function of tracking *actual* learning, but (like perceived learning), it just does so with mixed success. A third possibility, however, is that explanatory satisfaction instead has a different function: not of tracking the actual learning that has occurred from a given explanation, but of motivating and reinforcing inquiry in such a way as to promote actual learning in general. This third view will depart from the first two if there are conditions under which learning is best supported by a strategy for information search that does not always maximize immediate expected learning, but instead sometimes explores or learns simply for the sake of learning—whether or not the learning that ensues is useful or relevant to a given explanatory query.

While this seems suboptimal at first glance, there may be benefits to "imperfectly aligned motivation." For example, a motivational system that perfectly tracks moment-by-moment learning might miss the opportunity to explore topics that require a nontrivial investment in time or resources before leading to large epistemic gains (e.g., scientific research) or to reap the epistemic benefits of seeking and generating explanations without necessarily finding an accurate explanation (Wilkenfeld & Lombrozo, 2015). While speculative, this proposal is consistent with research on reinforcement learning: the best way to maximize reward in the long run is to occasionally explore, even when this exploration might lead to non-reward-maximizing behavior in the short run (Sutton & Barto, 1998). Analogously, the best way to maximize learning in the long run may be to occasionally seek and be satisfied with information that does not maximize learning in the short run—even information that appears to have little epistemic promise at first glance. By occasionally pursuing what appears to be a suboptimal line of inquiry, the learner could unexpectedly stumble upon something epistemically remarkable. This proposal raises several questions for future research: for example, in what environments might an imperfectly aligned agent ultimately learn more than a perfectly aligned agent, and what mixture of aligned and brute motivation is best?

Even if imperfectly aligned motivation has some benefits, there are clearly drawbacks, as well. First, some learning environments (e.g., echo chambers) might take advantage of an imperfectly aligned agent, by presenting false information that nonetheless makes one feel as if they learned something new and useful (see Nguyen, 2021). In addition, while there may be benefits to occasionally seeking information where learning is unlikely to occur, there will also be instances where a learner does not seek information when learning is likely. For example, people are susceptible to an "illusion of explanatory depth," believing that their explanatory understanding is much more complete than it actually is (Rozenblit & Keil, 2002). Thus, although individuals have incomplete explanatory understanding of phenomena ranging from common artifacts (Alter et al., 2010; Lawson, 2006; Rozenblit & Keil, 2002) to political policies (Alter et al., 2010; Fernbach et al., 2013; Vitriol & Marsh, 2018), their failure to recognize these gaps in their understanding might prevent them from pursuing inquiry and learning about these phenomena. In general, a more detailed characterization of the links between phenomenology, learning, and perceptions of learning will provide further insight into how humans learn from self-directed inquiry.

Our findings also raise new questions for understanding why and how "explanatory vices" sway explanatory satisfaction away from what appears to be epistemically rational. In Study 4, we found that participants' inflated satisfaction with reductive explanations (Hopkins et al., 2016) was fully explained by their judgments that these explanations provided explanatorily relevant information—even though the explanations were carefully constructed so that this was not the case (by experts' lights). Why, then, do people judge that reductive information provides genuinely explanatory content? One possibility is that reductive information serves as a "placeholder" for unknown but deep causal structure. As an analogous case, there is evidence that people prefer explanations that explain unusual behavior by appeal to labeled categories (e.g., "Randy has Depathapy, a tendency to imitate the actions of others and obey commands directed at them") as opposed to matching explanations that simply omit the category labels ("Randy has a tendency to imitate the actions of others and obey commands directed at them"). In these studies, the preference is largely explained by the inferences participants draw from the presence of the label: that it refers to a known causal source (Giffin et al., 2017). In the case of reductive information, learning that a psychological phenomenon is related to brain activity in the intraparietal sulcus could (in certain circumstances) provide little real explanatory benefit, but this explanation points to underlying causal structure and a relevant target of inquiry—the intraparietal sulcus—for further understanding why the psychological phenomenon occurs. That is, beyond the

possibility that novices over-generalize a criterion for explanatory quality that is useful in other cases (as suggested by Hopkins et al., 2016), it could be that the way in which reductive information structures inquiry has epistemic value (for novices, if not also for experts).

Our findings also have implications for understanding what motivates inquiry. First, we suspect that the generally strong associations between satisfaction and curiosity for both related and unrelated questions observed in Study 2 were driven by two inferences that participants made in parallel: one about the value of topical inquiry, as we predicted (e.g., "How much is there to learn about about dinosaur digestion?"; see Liquin & Lombrozo, 2020a; Oudeyer et al., 2007; Schmidhuber, 2010), and the other about the value of available informants (e.g., "How good are explanations from this source?"). Indeed, prior research has shown that informative responses from a particular source lead to selective trust in that source, as well as further information search directed towards that source (e.g., Corriveau & Kurkul, 2014; Koenig et al., 2004; Koenig & Harris, 2005; Landrum et al., 2015). Therefore, it is possible that curiosity about unrelated questions is driven by an expectation that the source (rather than the topic) will provide useful and informative explanations.

Both inferences (about the value of a particular topic and a particular source) could be linked to the results on calibration reported in the supplementary material: that individuals form systematic expectations about the extent to which forthcoming explanations will support learning and exhibit explanatory virtues, which are in fact related to perceptions of learning and explanatory virtues when the explanation is received. In particular, one puzzling question raised by these results is how individuals form these expectations. One possibility is that individuals form these expectations by inferring the value of topical inquiry and available sources on the basis of a previously received explanation. For example, if a learner was satisfied with a previous explanation on the topic of dinosaur digestion from a particular source—perhaps because they judged that they learned a lot and that the explanation was generalizable—they might infer that a new explanation on the topic of dinosaur digestion or from the same source would have similar properties. Thus, satisfaction's link to subsequent curiosity might be mediated by these inferences. Several additional mechanisms are possible: for example, individuals might guess the possible content of the to-be-received explanation, then evaluate that guessed content (see optimal models of question asking; e.g., Coenen et al., 2019). Alternatively, individuals might maintain higher-order knowledge about the typical causal structures invoked in a given domain (Strickland et al., 2017), which could guide domain-specific expectations about unanswered questions. Finally, individuals might use properties of the event or phenomenon being explained (e.g., its complexity) to make predictions about properties of the to-be-received explanation (Lim & Oppenheimer, 2020). Further research is needed to determine which of these mechanisms (or which combination of mechanisms) explains how individuals form reasonably accurate expectations about learning and explanatory virtues.

Our failure to find an association between satisfaction and decreased curiosity about the original question was surprising. One possibility is that the explanations presented to participants were not maximally satisfying, so participants were curious to receive other (more satisfying) explanations in response to the same question. Explanations are rarely "complete" (see Korman & Khemlani, 2020): for example, a causal explanation (e.g., "the water spilled because the dog bumped into the table") can always be enriched by adding more temporally distant causes (e.g., "the water spilled because the dog bumped into the table because he was clumsily rushing towards the treat on the other side of the room"). As a result, even a satisfying explanation might not be expected to halt inquiry; instead, individuals might become more curious about additional explanatory details. In addition, it is possible that curiosity and satisfaction are in part "observer neutral," in that they track the general epistemic value of an explanation independent of the learner's own knowledge (e.g., for the average individual). This idea is reflected in the measure we label "information content" in Studies 1–2. If curiosity and satisfaction track information content, we might expect curiosity to persist even after a satisfying explanation is received.

Finally, our findings have implications for how we might define motivational states such as curiosity and satisfaction. Curiosity has traditionally been defined as "desiring information for information's sake," with no instrumental motive (Loewenstein, 1994; see also Gottlieb et al., 2013). However, in the present research, we propose that curiosity and satisfaction may be aligned with *useful* learning, and we find that satisfaction is related to the perceived utility of information. Likewise, we report in previous research (Liquin & Lombrozo, 2020a) that curiosity is related to the expected utility of information (see also Abir et al., 2020; Dubey & Griffiths, 2020). Thus, evidence increasingly suggests that curiosity and satisfaction are in some part instrumental: they guide learners towards gaining *useful* information, blending both epistemic and instrumental motives. It is an open question how exactly these motives are weighed and combined, as well as whether curiosity and satisfaction are actually experienced as instrumental to learners. That is, individuals might feel as if they are pursuing information for information's sake, while actually being partly guided by instrumental motives (see Liquin & Lombrozo, 2020b). By combining self-report measures of curiosity and satisfaction—as we have done in this and other research (Liquin & Lombrozo, 2020a)—with other methodological approaches that emphasize behavioral measures (e.g., Abir et al., 2020; Baranes et al., 2015; Dubey & Griffiths, 2020; Hsee & Ruan, 2016; Kobayashi et al., 2019), we can gain further insight into this question.

7.2. Limitations

Several limitations of these studies must be acknowledged. First, these studies used predominantly correlational methods, which leaves open questions about the causal links between satisfaction, learning, and inquiry. We found in Study 4 that satisfaction is unlikely to cause actual learning, and we found in Study 5 that perceived learning has a causal influence on satisfaction. However, several additional causal links remain unexplored: for example, it remains unclear whether satisfaction causally influences subsequent inquiry, or whether satisfaction and subsequent curiosity are both caused by some third variable (e.g., perceived learning). In addition, Study 5 found evidence for both the predicted causal relation between perceived learning and satisfaction (that perceived learning causes satisfaction) and the reverse causal relation (that satisfaction causes perceived learning). One possible explanation for these

results is that people do not differentiate between perceived learning and satisfaction—in other words, it might be more accurate to conclude that satisfaction *is* perceived learning (at least in the minds of our participants), rather than that satisfaction is *caused by* perceived learning. This seems somewhat unlikely: though satisfaction ratings and learning ratings were correlated across studies, this correlation was far from perfect. However, satisfaction and perceived learning do appear to be tightly linked, and thus our attempt to disentangle them might not have sufficiently isolated the causal influence of each construct.

On the other hand, the aligned motivation account does not necessarily rule out a bidirectional association between satisfaction and perceived learning. In particular, it is possible that perceived learning partly determines satisfaction, and then experiencing satisfaction further reinforces one's perceptions of having learned. It is also possible that only a single causal direction is at play in any given instance, depending on which features of the explanation (or one's phenomenological response to it) are salient. For example, we might expect to find evidence for a different causal relation in a classroom, where perceived learning is likely salient, than in curiosity-motivated Wikipedia browsing, where satisfaction is likely salient. If this were the case, satisfaction would still be aligned with learning in many cases.

Indeed, prior research provides reason to suspect both causal relations may exist. First, research on *situational interest*—an affective and cognitive response to some external stimulus, characterized by heightened attention and feelings of enjoyment (Grossnickle, 2016; Hidi & Renninger, 2006; Schraw & Lehman, 2001)—has shown that interest is triggered by appraisals that a stimulus is new and complex yet can be comprehended (Silvia, 2005; see also Murayama et al., 2019)—in other words, perceptions that learning is possible. Satisfaction, which is characterized by similar affective and cognitive components, might analogously be triggered by learning-related appraisals. On the other hand, research on judgments of learning has demonstrated that inducing an emotional state in learners inflates their judgments of learning relative to a neutral emotional state (Baumeister et al., 2015). Thus, the affective experience associated with satisfaction might influence perceptions of learning. Together, in combination with our results from Study 5, these findings suggest that satisfaction and perceived learning are closely tied and likely bidirectionally related. However, further research is needed to determine the nature of these associations.

There are also several limitations of our measures of actual learning. First, the aligned motivation account predicts that satisfaction should be aligned with the in-the-moment learning that occurs while an explanation is being processed (i.e., when the explanation is first presented). Instead, our measures of multiple-choice performance and recall measured downstream memory several minutes after the explanation was presented. While memory for the explanation provides some evidence that in-the-moment learning occurred, it is an indirect measure. Second, it is possible for downstream memory to be high even for an explanation that is completely false. However, in this situation, we would likely claim that in-the-moment useful learning is very low—attaining false information does not typically constitute useful learning. Third, even if participants recall true information, they might achieve reasonable success on the multiple choice or recall tasks by recalling surface-level details rather than meaningful explanatory content. While developing a measure of in-the moment actual learning is challenging (and thus our measures of downstream memory are reasonable proxies), it remains possible that these limitations prevented us from finding a robust association between actual learning and satisfaction.

Finally, our materials were drawn from a fairly limited set of questions and explanations—in particular, those targeted towards sparking and satisfying curiosity on a variety of topics (Study 1), targeted towards teaching academic material (Studies 2–3), or targeted towards explaining scientific phenomena (Study 4). With the exception of the "bad" explanations in Study 4, the explanations we provided to participants were hand-crafted to teach specific information in a satisfying way. Future research should explore a broader range of questions and explanations. In addition, it would be valuable to understand whether these results extend to explanation-seeking questions in other domains (i.e., beyond the natural and social scientific questions used in the present research) and whether these results extend to how-questions in addition to why-questions. Moreover, there is evidence that participants' satisfaction about the answers to non-explanation-seeking questions shapes later memory for these answers (Marvin & Shohamy, 2016). This suggests that some components of our account of explanatory phenomenology may not be specific to explanation at all, an intriguing possibility for future research to explore.

7.3. Conclusion: imperfectly aligned motivation

The present research provides evidence supporting some aspects of the aligned motivation account and some aspects of the brute motivation account. Despite weak ties to multiple objective assessments of learning, explanatory satisfaction is robustly tied to multiple dimensions of perceived learning. Furthermore, the link between satisfaction and perceived learning helps explain previously puzzling "explanatory vices" (Hopkins et al., 2016; Weisberg et al., 2008). Thus, despite falling short of perfect (or even near-perfect) correspondence with actual learning, satisfaction is as aligned with learning as we could reasonably expect given human metacognitive limitations (Glenberg & Epstein, 1985; Lin & Zabrucky, 1998; Maki, 1998). While this leads us to an account of satisfaction that falls somewhere between "aligned motivation" and "brute motivation" (perhaps "imperfectly aligned motivation"), our findings raise important new questions about whether and when misaligned satisfaction might in fact be beneficial for a learner—for example, when certain avenues of inquiry do not initially appear promising but have the potential for later epistemic gains. Answers to these open questions promise to shed light on how phenomenological states (e.g., satisfaction, curiosity) combine with behavior (e.g., inquiry) to make us the deeply imperfect but often effective learners we are.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Casey Lewry for coding a subset of the recalled explanations from Study 3. We would also like to thank Diana Tamir, Tom Griffiths, and members of the Concepts and Cognition Lab for their useful feedback on this work. Some of the results reported here were presented at the 2019 meeting of the Cognitive Science Society, and we are grateful to these audiences for their discussion and feedback. In addition, subsets of this work were submitted in partial fulfillment of EL's dissertation requirement at Princeton University. This work was supported by research funds awarded to TL by Princeton University, as well as an NSF Graduate Research Fellowship to EL [grant number DGE-1656466]. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cogpsych.2021.101453.

References

Abir, Y., Marvin, C., Geen, C. van, Leshkowitz, M., Hassin, R., & Shohamy, D. (2020). *Rational curiosity and information-seeking in the COVID-19 pandemic*. PsyArXiv. https://doi.org/10.31234/osf.io/hcta4.

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology, 99*(3), 436–451. https://doi.org/10.1037/a0020218

Aronowitz, S., Lewry, C., & Lombrozo, T. (in prep). *Experiential explanations in iterated learning.*

Baranes, A. F., Oudeyer, P.-Y., & Gottlieb, J. (2015). Eye movements reveal epistemic curiosity in human observers. *Vision Research, 117*, 81–90. https://doi.org/10.1016/j.visres.2015.10.009

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumeister, R. F., Alquist, J. L., & Vohs, K. D. (2015). Illusions of learning: Irrelevant emotions inflate judgments of learning. *Journal of Behavioral Decision Making, 28*(2), 149–158.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28*(5), 610–632. https://doi.org/10.1016/0749-596X(89)90016-8

Blanchard, T., Lombrozo, T., & Nichols, S. (2018). Bayesian Occam's razor is a razor of the people. *Cognitive Science, 42*(4), 1345–1359. https://doi.org/10.1111/cogs.12573

Blanchard, T., Vasilyeva, N., & Lombrozo, T. (2018). Stability, breadth and guidance. *Philosophical Studies, 175*(9), 2263–2283. https://doi.org/10.1007/s11098-017-0958-6

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology, 48*(4), 1156–1164. https://doi.org/10.1037/a0026471

Bromme, R., & Thomm, E. (2016). Knowing who knows: Laypersons' capabilities to judge experts' pertinence for science topics. *Cognitive Science, 40*(1), 241–252. https://doi.org/10.1111/cogs.12252

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology, 98*(4), 550–558.

Chater, N., & Loewenstein, G. (2016). The under-appreciated drive for sense-making. *Journal of Economic Behavior & Organization, 126*, 137–154. https://doi.org/10.1016/j.jebo.2015.10.016

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2019). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review, 26*(5), 1548–1587. https://doi.org/10.3758/s13423-018-1470-5

Corriveau, K. H., & Kurkul, K. E. (2014). "Why does rain fall?": Children prefer to learn from an informant who uses noncircular explanations. *Child Development, 85*(5), 1827–1835. https://doi.org/10.1111/cdev.12240

Danovitch, J. H., & Keil, F. C. (2004). Should you ask a fisherman or a biologist?: Developmental shifts in ways of clustering knowledge. *Child Development, 75*(3), 918–931. https://doi.org/10.1111/j.1467-8624.2004.00714.x

Dubey, R., & Griffiths, T. L. (2020). Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review, 127*(3), 455–476. https://doi.org/10.1037/rev0000175

Dubey, R., Griffiths, T. L., & Lombrozo, T. (2019). In *If it's important, then I am curious: A value intervention to induce curiosity* (pp. 282–288). Cognitive Science Society.

Fall, A., Weber, B., Pakpour, M., Lenoir, N., Shahidzadeh, N., Fiscina, J., & Bonn, D. (2014). Sliding friction on wet and dry sand. *Physical Review Letters, 112*(17), Article 175502. https://doi.org/10.1103/PhysRevLett.112.175502

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political Extremism Is Supported by an Illusion of Understanding. *Psychological Science, 24*(6), 939–946. https://doi.org/10.1177/0956797612464058

Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult-child conversation. *Child Development, 80*(6), 1592–1611. https://doi.org/10.1111/j.1467-8624.2009.01356.x

Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2016). Young children prefer and remember satisfying explanations. *Journal of Cognition and Development, 17*(5), 718–736. https://doi.org/10.1080/15248372.2015.1098649

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy, 71*(1), 5–19. https://doi.org/10.2307/2024924

Giffin, C., Wilkenfeld, D., & Lombrozo, T. (2017). The explanatory effect of a label: Explanations with named categories are more satisfying. *Cognition, 168*, 357–369. https://doi.org/10.1016/j.cognition.2017.07.011

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 702–718. https://doi.org/10.1037/0278-7393.11.1-4.702

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). *rstanarm: Bayesian applied regression modeling via Stan.* R package version 2.21.1. https://mc-stan.org/rstanarm.

Gopnik, A. (2000). Explanation as orgasm and the drive for causal knowledge: The function, evolution, and phenomenology of the theory formation system. In F. C. Keil, & R. A. Wilson (Eds.), *Explanation and Cognition* (pp. 299–323). The MIT Press.

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. F. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences, 17*(11), 585–593. https://doi.org/10.1016/j.tics.2013.09.001

Green, D. P., Ha, S. E., & Bullock, J. G. (2010). Enough already about "black box" experiments: Studying mediation is more difficult than most scholars suppose. *The Annals of the American Academy of Political and Social Science, 628*(1), 200–208.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Grossnickle, E. M. (2016). Disentangling curiosity: Dimensionality, definitions, and distinctions from interest in educational contexts. *Educational Psychology Review, 28*(1), 23–60.
Gwynne, N. Z., & Lombrozo, T. (2010). In *The cultural transmission of explanations: Evidence that teleological explanations are preferentially remembered* (pp. 1301–1306). Cognitive Science Society.
Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(1), 22–34. https://doi.org/10.1037//0278-7393.29.1.22
Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*(2), 111–127.
Hopkins, E. J., Weisberg, D. S., & Taylor, J. C. V. (2016). The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition, 155*, 67–76. https://doi.org/10.1016/j.cognition.2016.06.011
Hsee, C. K., & Ruan, B. (2016). The Pandora effect: The power and peril of curiosity. *Psychological Science, 27*(5), 659–666. https://doi.org/10.1177/0956797616631733
Johnson, S. G. B., Rajeev-Kumar, G., & Keil, F. C. (2016). Sense-making under ignorance. *Cognitive Psychology, 89*, 39–70. https://doi.org/10.1016/j.cogpsych.2016.06.004
Johnson, S. G. B., Valenti, J. J., & Keil, F. C. (2019). Simplicity and complexity preferences in causal explanation: An opponent heuristic account. *Cognitive Psychology, 113*, Article 101222. https://doi.org/10.1016/j.cogpsych.2019.05.004
Johnston, A. M., Sheskin, M., Johnson, S. G. B., & Keil, F. C. (2018). Preferences for explanation generality develop early in biology but not physics. *Child Development, 89*(4), 1110–1119. https://doi.org/10.1111/cdev.12804
Keil, F. C., Stein, C., Webb, L., Billings, V. D., & Rozenblit, L. (2008). Discerning the division of cognitive labor: An emerging understanding of how knowledge Is clustered in other minds. *Cognitive Science, 32*(2), 259–300. https://doi.org/10.1080/03640210701863339
Kerrod, R., Madgwick, W., Reed, S., Collins, F., & Brooks, P. (2006). *1000 questions & answers factfile*. Kingfisher.
Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). Harry Potter and the sorcerer's scope: Latent scope biases in explanatory reasoning. *Memory & Cognition, 39*(3), 527–535. https://doi.org/10.3758/s13421-010-0028-1
Kim, N. S., & Keil, F. C. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory & Cognition, 31*(1), 155–165. https://doi.org/10.3758/BF03196090
Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher, & W. C. Salmon (Eds.), *Scientific Explanation* (pp. 410–505). University of Minnesota Press.
Kobayashi, K., Ravaioli, S., Baranès, A., Woodford, M., & Gottlieb, J. (2019). Diverse motives for human curiosity. *Nature Human Behaviour, 3*, 587–595. https://doi.org/10.1038/s41562-019-0589-3
Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science, 15*(10), 694–698. https://doi.org/10.1111/j.0956-7976.2004.00742.x
Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development, 76*(6), 1261–1277. https://doi.org/10.1111/j.1467-8624.2005.00849.x
Kominsky, J. F., Zamm, A. P., & Keil, F. C. (2018). Knowing when help is needed: A developing sense of causal complexity. *Cognitive Science, 42*(2), 491–523. https://doi.org/10.1111/cogs.12509
Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349. https://doi.org/10.1037/0096-3445.126.4.349
Korman, J., & Khemlani, S. (2020). Explanatory completeness. *Acta Psychologica, 209*, Article 103139. https://doi.org/10.1016/j.actpsy.2020.103139
Kurkul, K. E., & Corriveau, K. H. (2018). Question, explanation, follow-up: A mechanism for learning from others? *Child Development, 89*(1), 280–294. https://doi.org/10.1111/cdev.12726
Landrum, A. R., Eaves, B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences, 19*(3), 109–111. https://doi.org/10.1016/j.tics.2014.12.007
Lawson, R. (2006). The science of cycology: Failures to understand how everyday objects work. *Memory & Cognition, 34*(8), 1667–1675. https://doi.org/10.3758/BF03195929
Lim, J. B., & Oppenheimer, D. M. (2020). Explanatory preferences for complexity matching. *PLoS ONE, 15*(4), Article e0230929. https://doi.org/10.1371/journal.pone.0230929
Lin, L.-M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23*(4), 345–391. https://doi.org/10.1006/ceps.1998.0972
Liquin, E. G., Callaway, F., & Lombrozo, T. (2020). In *Quantifying curiosity: A formal approach to dissociating causes of curiosity* (pp. 309–315). Cognitive Science Society.
Liquin, E. G., & Lombrozo, T. (2020a). A functional approach to explanation-seeking curiosity. *Cognitive Psychology, 119*, Article 101276. https://doi.org/10.1016/j.cogpsych.2020.101276
Liquin, E. G., & Lombrozo, T. (2020b). Explanation-seeking curiosity in childhood. *Current Opinion in Behavioral Sciences, 35*, 14–20. https://doi.org/10.1016/j.cobeha.2020.05.012
Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin, 116*(1), 75–98. https://doi.org/10.1037/0033-2909.116.1.75
Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55*(3), 232–257. https://doi.org/10.1016/j.cogpsych.2006.09.006
Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences, 20*(10), 748–759. https://doi.org/10.1016/j.tics.2016.08.001
Lutz, D. J., & Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child Development, 73*(4), 1073–1084. https://doi.org/10.1111/1467-8624.00458
MacKinnon, D. (2008). *Introduction to Statistical Mediation Analysis*. Taylor & Francis Group.
Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in Educational Theory and Practice* (pp. 117–144). Taylor & Francis.
Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). BayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software, 4*(40), 1541. https://doi.org/10.21105/joss.01541.
Martí, L., Mollica, F., Piantadosi, S., & Kidd, C. (2018). Certainty is primarily determined by past performance during concept learning. *Open Mind, 2*(2), 47–60. https://doi.org/10.1162/opmi_a_00017
Marvin, C. B., & Shohamy, D. (2016). Curiosity and reward: Valence predicts choice and information prediction errors enhance learning. *Journal of Experimental Psychology: General, 145*(3), 266–272. https://doi.org/10.1037/xge0000140
Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences, 20*(9), 689–700. https://doi.org/10.1016/j.tics.2016.07.001
Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968
Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). *Advances in Pre-Training Distributed Word Representations*. http://arxiv.org/abs/1712.09405.
Mills, C. M., Sands, K. R., Rowles, S. P., & Campbell, I. L. (2019). "I want to know more!": Children are sensitive to explanation quality when exploring new information. *Cognitive Science, 43*(1), Article e12706. https://doi.org/10.1111/cogs.12706
Mueller, M. L., & Dunlosky, J. (2017). How beliefs can impact judgments of learning: Evaluating analytic processing theory with beliefs about fluency. *Journal of Memory and Language, 93*, 245–258. https://doi.org/10.1016/j.jml.2016.10.008
Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language, 70*, 1–12. https://doi.org/10.1016/j.jml.2013.09.007

Murayama, K., FitzGibbon, L., & Sakaki, M. (2019). Process account of curiosity and interest: A reward-learning perspective. *Educational Psychology Review, 31*(4), 875–895. https://doi.org/10.1007/s10648-019-09499-9
Nguyen, C. T. (2021). The seductions of clarity. *Royal Institute of Philosophy Supplement, 89*, 227–255. https://doi.org/10.1017/S1358246121000035
Oudeyer, P.-Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation, 11*(2), 265–286. https://doi.org/10.1109/TEVC.2006.890271
Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General, 146*(12), 1761–1780. https://doi.org/10.1037/xge0000318
Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879
Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology, 65*(3), 429–447. https://doi.org/10.1037/0022-3514.65.3.429
Reber, R., & Greifeneder, R. (2017). Processing fluency in education: How metacognitive feelings shape learning, belief formation, and affect. *Educational Psychologist, 52*(2), 84–103. https://doi.org/10.1080/00461520.2016.1258173
Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02.
Rozenblit, L., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1
Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development, 2*(3), 230–247. https://doi.org/10.1109/TAMD.2010.2056368
Schraw, G., & Lehman, S. (2001). Situational Interest: A Review of the Literature and Directions for Future Research. *Educational Psychology Review, 13*(1), 23–52. https://doi.org/10.1023/A:1009004801455
Silvia, P. J. (2005). What is interesting? Exploring the appraisal structure of interest. *Emotion (Washington, D.C.), 5*(1), 89–102. https://doi.org/10.1037/1528-3542.5.1.89
Strevens, M. (2004). The causal and unification approaches to explanation unified—Causally. *Noûs, 38*(1), 154–176. https://doi.org/10.1111/j.1468-0068.2004.00466.x
Strickland, B., Silver, I., & Keil, F. C. (2017). The texture of causal construals: Domain-specific biases shape causal inferences from discourse. *Memory & Cognition, 45*(3), 442–455. https://doi.org/10.3758/s13421-016-0668-x
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). MIT Press.
Thagard, P. R. (1978). The best explanation: Criteria for theory choice. *The Journal of Philosophy, 75*(2), 76–92. https://doi.org/10.2307/2025686
Trout, J. D. (2008). Seduction without cause: Uncovering explanatory neurophilia. *Trends in Cognitive Sciences, 12*(8), 281–282. https://doi.org/10.1016/j.tics.2008.05.004
Ünlütabak, B., Nicolopoulou, A., & Aksu-Koç, A. (2019). Questions asked by Turkish preschoolers from middle-SES and low-SES families. *Cognitive Development, 52*, Article 100802. https://doi.org/10.1016/j.cogdev.2019.100802
Vitriol, J. A., & Marsh, J. K. (2018). The illusion of explanatory depth and endorsement of conspiracy beliefs. *European Journal of Social Psychology, 48*(7), 955–969. https://doi.org/10.1002/ejsp.2504
Vogl, E., Pekrun, R., & Loderer, K. (2021). Epistemic Emotions and Metacognitive Feelings. In D. Moraitou, & P. Metallidou (Eds.), *Trends and Prospects in Metacognition Research across the Life Span: A Tribute to Anastasia Efklides* (pp. 41–58). Springer.
Vredenburgh, C., & Kushnir, T. (2016). Young children's help-seeking as active information gathering. *Cognitive Science, 40*(3), 697–722. https://doi.org/10.1111/cogs.12245
Wade, S., & Kidd, C. (2019). The role of prior knowledge and curiosity in learning. *Psychonomic Bulletin & Review, 26*, 1377–1387. https://doi.org/10.3758/s13423-019-01598-6
Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology, 60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001
Walters, D. J., Fernbach, P. M., Fox, C. R., & Sloman, S. A. (2017). Known unknowns: A critical determinant of confidence and calibration. *Management Science, 63*(12), 4298–4307.
Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*(3), 470–477. https://doi.org/10.1162/jocn.2008.20040
Weisberg, D. S., Taylor, J., & Hopkins, E. (2015). Deconstructing the seductive allure of neuroscience explanations. *Judgment and Decision Making, 10*(5), 429–441.
Wilkenfeld, D. A., & Lombrozo, T. (2015). Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science & Education, 24*(9–10), 1059–1077. https://doi.org/10.1007/s11191-015-9784-4
Wilkenfeld, D. A., Plunkett, D., & Lombrozo, T. (2016). Depth and deference: When and why we attribute understanding. *Philosophical Studies, 173*(2), 373–393. https://doi.org/10.1007/s11098-015-0497-y
Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General, 142*(4), 1006–1014. https://doi.org/10.1037/a0030996
Zemla, J. C., Sloman, S. A., Bechlivanidis, C., & Lagnado, D. (2020). *Not so simple! Mechanisms increase preference for complex explanations*. PsyArXiv. https://doi.org/10.31234/osf.io/jbn5f.
Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic Bulletin & Review, 24*(5), 1488–1500. https://doi.org/10.3758/s13423-017-1258-z