From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Psychology

2021

Imperfect Inquiry: How Curiosity and Satisfaction Guide Learning

Liquin, Emily Grace

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp012f75rc163

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

IMPERFECT INQUIRY: HOW CURIOSITY AND SATISFACTION GUIDE LEARNING

Emily G. Liquin

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY

THE DEPARTMENT OF

PSYCHOLOGY

Adviser: Tania Lombrozo

September 2021

© Copyright by Emily G. Liquin, 2021. All rights reserved.

ABSTRACT

Humans are remarkably effective learners, and we are particularly unique in our desire to understand why and how things happen. Why do humans pursue explanations with such fervor? In this dissertation, I explore the phenomenological state that drives learners to pursue explanations (explanation-seeking curiosity) and the phenomenological state that signals inquiry has succeeded (explanatory satisfaction)—the states that together comprise "explanatory phenomenology." In Chapter 1, I present several competing accounts of explanatory phenomenology, which posit different connections between explanatory phenomenology and the goal of learning useful explanations. In Chapters 2-4, I test these accounts. In Chapter 2, I report empirical evidence that explanation-seeking curiosity is related to individuals' reported expectations that inquiry will lead to useful learning, above and beyond heuristic cues to learning. In Chapter 3, I show that explanation-seeking curiosity is related to expected learning/uncertainty about the query of interest, as operationalized by an optimal model. However, the optimal utility of information and optimal expected learning about a global target of inquiry are less consistently associated with curiosity. In Chapter 4, I report evidence that explanatory satisfaction is associated with perceived learning but is less consistently related to several measures of actual learning. In Chapter 5, I summarize these results and explore their implications for understanding when and why humans experience explanatory phenomenology. Taken together, this work supports the view that explanatory phenomenology is an imperfect motivating force behind useful learning. By experiencing explanatory phenomenology and pursuing inquiry, learners are guided towards asking questions that have strong but fallible potential to lead to informative and useful explanations.

ACKNOWLEDGEMENTS

I feel so lucky to have been able to spend the last five years pursuing the questions that make me most curious. This would not have been possible without the support and encouragement of my advisors, colleagues, friends, and family.

First, I am incredibly grateful for my advisor, Tania Lombrozo, who supervised the work presented in this dissertation. Tania, this dissertation would not exist without your guidance—but more importantly, I would not be the researcher and person I am today without your support. I cannot thank you enough for your mentorship. I am also deeply grateful for the support of Alison Gopnik. Alison, thank you for continuing to be a dedicated mentor from afar after I moved to Princeton. I hope that I've gained even a fraction of your wisdom from my time working with you. I also thank the other members of my dissertation reading committee, Tom Griffiths and Diana Tamir, who have provided invaluable feedback on this document, and my other oral exam committee member, Yael Niv.

Fred Callaway made important contributions to the work presented in Chapter 3. I have also had the pleasure of working with several collaborators, co-authors, and research assistants on work not presented here. Thank you all for shaping my ideas and for contributing your time and talents to our joint research.

The past and present members of the Concepts and Cognition Lab and the Gopnik Cognitive Development Lab have provided friendship and support and have assisted my research at all stages. I especially thank those with whom I've shared an office throughout the last five years: Ny Vasil, Patricia Mirabile, and Corey Cusimano. I also owe special thanks to the past and present lab managers of both labs: Ellie Kon, Maureen Gill, Casey Lewry, Sarah de la Vega, Teresa Garcia, and Rose Reagan. I also thank Casey Lew-Williams, Lauren Emberson, and the

members of the Princeton Baby Lab, who welcomed me to their lab meetings and developmental testing spaces as an unofficial Baby Lab member.

I am deeply grateful for Kelly Sheehan, David Uttal, and Mark Beeman, who introduced me to research as an undergraduate and paved the way for my success in graduate school.

I thank the Department of Psychology at UC Berkeley and the Department of Psychology at Princeton, especially the administrative staff who helped me navigate degree requirements, reimbursements, and more.

I was lucky enough to end up with three cohorts: the 2016 cohort at UC Berkeley (with whom I shared the experience of starting graduate school), the 2018 cohort at Princeton (with whom I began my time at Princeton), and the 2016 cohort at Princeton (who welcomed me into their cohort despite my late arrival in the department). Thank you all for being great company on this journey. I am also so grateful for the friends I met during graduate school, especially Shoshana Jarvis, Micah Brush, Fred Callaway, Zidong Zhao, and Juliana Trach. And of course, I am endlessly grateful for Nikki Koutas, Kristen Fairbanks, Alex Hoffman, Lauren Burke, Angela Metcalf, and Lana Kammerer, who have made up my support system since well before I started graduate school.

Lastly, I would not have finished this dissertation without the support of my partner, Amanda, and my family: my parents, Jane and Daryl, my brother, Chris, and my grandparents, aunts, uncles, and cousins. I know you're all sometimes confused about what I'm studying and why it matters. Thank you for cheering me on anyway. Mom, Dad, and Chris, I am eternally grateful for your unwavering support and your pride in all of my accomplishments, large and small. This dissertation is dedicated to you.

Funding Sources

The work in this dissertation was supported by the Berkeley Fellowship for Graduate Study and the National Science Foundation Graduate Research Fellowship Program (grant numbers DGE-1656466, DGE-1752814, and DGE-1106400). The work in Chapter 2 was also supported by a James S. McDonnell Foundation Scholar Award in Understanding Human Cognition and a Templeton Foundation Grant awarded to Tania Lombrozo. Chapters 3 and 4 were also supported by research funds awarded to TL by Princeton University.

TABLE OF CONTENTS

Page

Abstract iii

Acknowledgements iv

List of Tables ix

List of Figures x

1 Chapter 1: Introduction 1

1.1 Explanations are valuable; curiosity and satisfaction help us find them 3

1.2 What elicits explanatory phenomenology? 9

1.3 The present research 16

2 Chapter 2: A Functional Approach to Explanation-Seeking Curiosity 21

2.1 Introduction 21

2.2 Study 1 35

2.3 Study 2 43

2.4 Study 3 52

2.5 General Discussion 64

3 Chapter 3: Quantifying and Dissociating Elicitors of Curiosity 70

3.1 Introduction 70

3.2 Study 4 77

3.3 Study 5 91

3.4 Study 6 94

3.5 Study 7 111

3.6 General Discussion 118

4 Chapter 4: In Search of Satisfying Explanations 127

4.1 Introduction 127
4.2 Study 8 136
4.3 Study 9 145
4.4 Study 10 153
4.5 Study 11 159
4.6 General Discussion 172
5 Chapter 5: Conclusion 182
5.1 Summary of empirical work 183
5.2 Limitations 187
5.3 Open questions on imperfect goal-alignment 189
5.4 Explanatory phenomenology and useful learning 194
6 Appendices 196
Appendix A 196
Appendix B 197
Appendix C 198
Appendix D 201
References 205

LIST OF TABLES

Table Page

Table 1 Potential Triggers of Curiosity, Studies 1-3 34

Table 2 Model-estimated Features and Summary of Results, Studies 4-7 81

Table 3 Individual Regression Analyses Significance Tests, Studies 4-7 88

Table 4 Multiple Regression Analyses Significance Tests, Studies 4-7 88

Table 5 Predictions of Aligned and Brute Motivation Accounts, Studies 8-11 135

Table 6 Items, Studies 8-9 138

LIST OF FIGURES

Figure Page

Figure 1 Regression Results, Studies 1-3 40

Figure 2 Factor Analysis for Explanation-Seeking Questions, Studies 2-3 49

Figure 3 Factor Analysis for Fact-Seeking Questions, Study 3 62

Figure 4 Procedure for Studies 4-7. 79

Figure 5 Regression Analysis Results, Studies 4-7. 85

Figure 6 Associations Between Epistemic Features and Curiosity, Studies 4-5. 87

Figure 7 Model-Estimated Feature Correlations, Studies 4-7 89

Figure 8 Instrumental value, Choice, and Curiosity, Studies 6-7 106

Figure 9 Added Variance Explained by Instrumental Features, Studies 6-7 110

Figure 10 Explanatory Satisfaction and Perceived Learning, Studies 8-9 143

Figure 11 Explanatory Satisfaction and Subsequent Curiosity, Study 9 151

Figure 12 Explanatory Satisfaction and Actual Learning, Studies 10-11 158

Figure 13 Manipulating Perceived Learning and Satisfaction, Study 11 164

Figure 14 Mediation Analyses, Study 11 171

1 CHAPTER 1: INTRODUCTION

Why is the sky blue? Why do some people hate the taste of cilantro? Why is there human suffering? "Why" questions such as these reflect the process of explanatory inquiry: thinking, performing experiments, and asking questions in search of explanations. From a very young age, children seek explanations by asking questions (Chouinard, 2007) and performing informal experiments (Gopnik & Wellman, 2012; L. E. Schulz, 2012). As adults, we turn to Google when encountering mundane uncertainties (from my own search history, "Why does MATLAB take up so much space on my computer?"), and we turn to science and religion to explain more challenging unknowns (e.g., "Why do people get sick?"; Legare et al., 2012; Legare & Gelman, 2008; Shtulman & Lombrozo, 2016). At its best, this desire to understand why and how things happen leads humans to develop theories of gravitation, evolution, or cognition. But even in its more mundane forms, explanatory inquiry leads humans to build and revise intuitive theories, by which we explain and interpret our everyday observations (Carey, 1985; Gopnik & Meltzoff, 1997; Murphy & Medin, 1985; Wellman & Gelman, 1992).

But why do humans pursue explanations with such fervor? As argued by Gopnik (2000), in order for explanatory inquiry to get off the ground, there must be a motivational system in place that drives learners to perform experiments, ask questions, and otherwise seek explanations ("explanation-seeking curiosity"), and that allows learners to recognize when these efforts have succeeded ("explanatory satisfaction"). Provocatively, Gopnik likens explanation to orgasm: just as the phenomenology of orgasm motivates learners to perform the behaviors that support reproduction, the phenomenology of explanatory inquiry (curiosity and satisfaction) motivates learners to perform the behaviors that support the learning of explanations. Others have likened

curiosity to hunger (Lau et al., 2020): just as hunger motivates the search for food, curiosity motivates the search for information.

While these analogies appear apt, they raise questions about how exactly the phenomenology (orgasm, hunger, curiosity/satisfaction) is linked to the ultimate goal of the corresponding behavior (reproduction, adequate nutrition, learning). Orgasm motivates sex relatively indiscriminately: humans feel sexual desire even when sex is unlikely to lead to reproduction (e.g., when using birth control). Likewise, though hunger motivates us to seek out food, we often crave foods that are not nutritionally valuable, and we typically find donuts more satisfying than spinach. In other words, orgasm and craving are experienced in a way that is fairly independent of the goals they serve—they motivate the appropriate behavior (sex and eating), but they provide little guarantee that the corresponding goal (reproduction and nutrition) will be achieved.[1] In contrast, it is easy to imagine how orgasm and craving could be aligned with their ultimate goals: orgasm might be most likely when reproduction is also likely, and cravings might be specifically directed towards foods that are most valuable in our nutritional environment (e.g., healthy foods).

In this dissertation, I will apply these ideas to explanatory phenomenology, providing insight into when and why explanatory inquiry occurs. One possibility is that curiosity and satisfaction are dissimilar from orgasm and craving. In particular, curiosity and satisfaction might be aligned with useful learning, so that we selectively experience curiosity when good and useful explanations are likely to be found and selectively experience satisfaction when we find

[1] It is possible that orgasms and cravings serve goals besides ensuring reproduction and adequate nutrition, respectively. Likewise, it is possible that orgasms or cravings are aligned with their goals but only in certain environments—for example, if calorie-dense foods are rare, craving donuts rather than spinach could be beneficial. For the sake of motivating this program of research, I rely on what are likely to be oversimplified accounts of these motivational states.

them. If this were the case, then inquiry would reliably lead to good explanations. In contrast, another possibility is that curiosity and satisfaction are independent of the goal of explanatory inquiry: the learning of useful explanations. This independence could take many forms. For example, someone might experience curiosity about any question concerning dinosaurs, even if the question does not promise much useful information (e.g., "Why did dinosaurs have heads?"), or someone might selectively experience curiosity about fictional detective shows, which might provide some information but have little real-world utility (see Dubey & Griffiths, 2020). If curiosity and satisfaction were unrelated to the goal of explanatory inquiry in this way, just as orgasm and craving are largely unrelated to the goals of reproduction and nutrition, then inquiry would often fail to result in useful explanations—but we could still learn a great deal if inquiry were pursued often enough or if good explanations were plentiful enough.

These ideas will be elaborated in the remainder of this chapter. In Section 1.1, I will define explanation and motivate the study of explanatory phenomenology. In Section 1.2, I will outline four possible views of explanatory phenomenology and supporting empirical evidence for each view. In Section 1.3, I will propose a series of questions that must be answered if we are to understand which of these views is correct, and I will summarize my approach to answering these questions.

1.1 Explanations are valuable; curiosity and satisfaction help us find them

Explanation and curiosity, and to a lesser degree satisfaction, have all received attention as crucial components of how humans come to comprehend the world. In what follows, I will (loosely) define explanation, and I will review theoretical perspectives and empirical evidence on explanation's importance for learning and action. Then, I will argue that the study of curiosity and satisfaction can provide new insight into open questions on explanation.

1.1.1 Explanation

What is an explanation? One early account of explanation (Hempel & Oppenheim, 1948) proposed that a scientific explanation for a piece of empirical evidence subsumes the evidence under a general law. On the basis of such a general law, the evidence can be logically deduced and therefore could have been predicted prior to its occurrence. Hempel and Oppenheim brought attention to explanation in the philosophy of science, and subsequent accounts refined this initial treatment. I will not review these accounts in detail (see Salmon, 1989 for a comprehensive review). However, these theories provide two useful insights. First, they sketch a general definition of "explanation": an explanation relates some observation to general patterns or regularities, often by a causal relation. Second, they provide us with a sense of why explanations might be epistemically and practically valuable: they allow us to make predictions, intervene on causal systems (see Woodward, 2003), and reduce our many observations to a set of general principles (see Wilkenfeld, 2019).

While these accounts largely concern explanation in science, the insights from this literature are consistent with work in psychology on more commonplace forms of explanation. A large body of research in social psychology characterizes explanation as fundamental to how humans make sense of our own and other's behavior (e.g., Heider, 1958; Kelley, 1973; Malle, 2004; Weiner, 1985a; P. T. Wong & Weiner, 1981). In addition, influential accounts of concepts (Carey, 1985; Murphy & Medin, 1985) and intuitive theories (Gopnik & Meltzoff, 1997; Wellman & Gelman, 1992) characterize explanations as central components of humans' cognitive capacities (see also Keil, 2003, 2006). Indeed, empirical evidence suggests that children and adults use explanations (along with explanatory causal knowledge and intuitive theories) to interpret new objects and events (e.g., Ahn et al., 2002, 2003; Griffiths &

Tenenbaum, 2009; Karmiloff-Smith & Inhelder, 1974; Khemlani & Johnson-Laird, 2012; Murphy & Allopenna, 1994; Pennington & Hastie, 1992; Waldmann & Hagmayer, 2001), make predictions and inferences about novel objects and events (e.g., Bonawitz et al., 2012, 2019; Gweon et al., 2010; Lombrozo, 2009; Lombrozo et al., 2018; Lombrozo & Gwynne, 2014; Patalano et al., 2006; Rehder & Hastie, 2001, 2004), and guide action and exploration, (e.g., Bonawitz et al., 2012; Callanan et al., 2020; Danovitch & Mills, 2018; Gweon & Schulz, 2011; Legare, 2012; Legare et al., 2015; Meltzoff, 1995; Ruggeri et al., 2019; L. E. Schulz & Gopnik, 2004). For example, Legare (2012) prompted children to generate an explanation for an expectation-violating event (e.g., a block failing to activate a machine), then subsequently gave children the opportunity to explore the participating objects. The kind of explanation a child provided (e.g., "It's out of batteries" vs. "It's the wrong kind of block") was related to how that child later explored (e.g., looking inside the block vs. sorting blocks into categories). Together, this research illustrates the value of explanations: even from a very young age, we use explanations to inform learning and action.

In addition to these benefits of having explanations, there are epistemic benefits to seeking and generating explanations (Wilkenfeld & Lombrozo, 2015). Influential research in education (Chi et al., 1989, 1994) has documented a benefit to "self-explaining" during learning. In one study, students who were prompted to generate explanations while reading an educational text performed better on a subsequent knowledge test than those who were not prompted to explain, and those who provided many self-explanations demonstrated deeper understanding (Chi et al., 1994; see Rittle-Johnson et al., 2017 for a recent review and meta-analysis). Lombrozo (2016) has proposed that generating explanations draws a learner's attention to the features that make an explanation good—for example, simplicity, breadth, and causal structure.

Consistent with this account, when children and adults are prompted to explain while learning, they are more likely to generate and favor hypotheses that are simple (Walker, Bonawitz, et al., 2017) and broad (Walker, Lombrozo, et al., 2017; Walker & Lombrozo, 2017; Williams & Lombrozo, 2010, 2013), and they are more likely to attend to causally-relevant features (Legare & Lombrozo, 2014; Walker et al., 2014).

These findings are also consistent with the philosophical literature on "inference to the best explanation" (e.g., Harman, 1965; Lipton, 2004; Thagard, 1978), which has argued that an explanation's "loveliness"—its possession of desirable explanatory virtues like simplicity and breadth—provides a guide to its "likeliness"—its probability of being correct (Lipton, 2004). Recent empirical work has shown individuals do incorporate explanatory considerations into their probabilistic belief updating and inference (Douven & Mirabile, 2018; Douven & Schupbach, 2015a, 2015b). Further, it has been argued that inference to the best explanation could be rationally justified, in particular if one's aim is to quickly converge on a hypothesis that is likely to be true, as opposed to minimize inaccuracy in the long-term (Douven, 2013, 2020; Douven & Wenmackers, 2017).

Taken together, the work reviewed in this section demonstrates the special value of having and seeking explanations. First, explanations allow us to understand our observations by organizing them under broad (often causal) patterns (Friedman, 1974; Grimm, 2010; Keil, 2006; Strevens, 2013), so that receiving a new explanation tends to increase both the amount of knowledge we have and the coherence and simplicity of our body of knowledge. And second, explanations enable us to act. By having an explanation, we can design interventions on causal systems (Woodward, 2003), predict what is going to happen (Craik, 1943; Hempel &

Oppenheim, 1948), and decide what information to seek (Nyrup, 2015). Thus, explanation has (at least) two goals: epistemic and instrumental (see Lombrozo, 2011).

But when are these epistemic and instrumental goals likely to be achieved? In order to predict when these positive benefits of explanation are likely to occur, we need to determine when individuals are motivated to seek out explanations in the first place. An answer to this question holds potential for informing interventions that spark explanation search when it would be desirable—for example, in educational settings. In the following section, I introduce the study of curiosity and satisfaction as a particularly promising route to answering this question.

1.1.2 Curiosity and satisfaction

While relatively little research has addressed when and why individuals are motivated to seek explanations, there is a large body of research on curiosity, which has addressed when and why individuals are motivated to seek information that is not necessarily explanatory. The empirical study of curiosity was pioneered by psychologist Daniel Berlyne (e.g., Berlyne, 1950, 1960, 1978). Curiosity was initially defined as a *non-instrumental* drive for information: we feel curious about information because we want that information for its own sake, and not because the information will help us achieve any immediate practical goal (e.g., Berlyne, 1954; Loewenstein, 1994). However, later formulations of curiosity have relaxed the assumption that curiosity is completely divorced from instrumental motives (e.g., Dubey & Griffiths, 2020; Gottlieb et al., 2020). I broadly define curiosity as a phenomenological state characterized by the desire for information—leaving the question of whether it is instrumental or non-instrumental to empirical study (see Chapter 3). Following a plausible folk definition (e.g., "my curiosity has been satisfied"), I define satisfaction in relation to curiosity: satisfaction is the sense that the desire for information (curiosity) has been reduced to an acceptable degree.

Curiosity and satisfaction have benefits for learning and memory, so that an explanation attained while experiencing these phenomenological states might be better learned. For example, curiosity about a given target (e.g., the answer to a trivia question) is associated with better memory for that target (Fandakova & Gruber, 2021; Galli et al., 2018; Gruber et al., 2014; Kang et al., 2009; Ligneul et al., 2018; McGillivray et al., 2015; Mullaney et al., 2014; Stare et al., 2018) and in some cases, for incidentally presented stimuli unrelated to the target information (Fandakova & Gruber, 2021; Galli et al., 2018; Gruber et al., 2014; Stare et al., 2018). Satisfaction appears to have similar effects: the difference between rated satisfaction with a trivia answer and initial curiosity about the trivia question ("information prediction error") is related to subsequent memory for the answer (Marvin & Shohamy, 2016; see also Fandakova & Gruber, 2021; Fastrich et al., 2018; Jepma et al., 2012; McGillivray et al., 2015). Taken together, these findings suggest that information learned when curious or satisfied is better remembered. To the extent that these results extend beyond trivia facts to explanations, we might expect that the explanations found during inquiry will be especially powerful in shaping learning, inference, and action, as those explanations are likely to be especially memorable.

More importantly, the study of curiosity and satisfaction promises to shed light on when and why we pursue explanations because these phenomenological states motivate action. For example, individuals are selectively willing to pay a cost to reveal the answers to questions that they were more curious about (Kang et al., 2009; Marvin & Shohamy, 2016). Indeed, costly information search has often been used as a behavioral measure of curiosity (Abir et al., 2020; Baranes et al., 2015; T. C. Blanchard et al., 2015; Brydevall et al., 2018; Dubey & Griffiths, 2020; FitzGibbon et al., 2019; Hsee & Ruan, 2016; Jirout & Klahr, 2012; Kobayashi et al., 2019; Subbotsky, 2010; van Lieshout et al., 2018; Wang & Hayden, 2019). While satisfaction has

received less empirical study, there is evidence that children are less likely to re-ask a question after receiving a genuine explanation (which is perhaps satisfying) as opposed to a non-answer (Frazier et al., 2009, 2016; Kurkul & Corriveau, 2018; Mills et al., 2019; see also Ünlütabak et al., 2019), consistent with the idea that satisfaction halts inquiry. Thus, in an explanation-seeking context, we are likely to seek explanations when we feel curious, and we are likely to halt our search when we feel satisfied. Notably, the research reviewed here concerns the observable behavioral consequences of feeling curiosity or satisfaction—for example, curiosity might motivate an individual to ask a question. It is likely that curiosity and satisfaction also motivate unobservable cognitive consequences, such as trying to come up with an answer on one's own (which has been linked to learning in explanation-seeking contexts; e.g., Chi et al., 1989, 1994).

1.2 What elicits explanatory phenomenology?

In the previous section, I argued that explanation-seeking curiosity motivates us to search for explanations. This search for explanations (if all goes well) will provide us with a satisfying explanation, allowing us to terminate inquiry and reap the epistemic and instrumental benefits of having found an explanation. However, much can go wrong in this process: for example, curiosity might motivate us to search for explanations where none are available, and satisfaction might misfire, leading us to halt inquiry before an informative and useful explanation has been found. In other words, the success of this process depends heavily on when curiosity and satisfaction are experienced. To return to an earlier example, an individual who is selectively curious about questions concerning dinosaurs might miss non-dinosaur-related opportunities for learning. As an even more extreme example, imagine someone who only experiences curiosity about questions that contain the word "and" and who experiences satisfaction only about answers that contain the word "that." To this person, "Why are the cups made of glass and in the

cabinet?" would elicit a strong feeling of curiosity. They would feel motivated to seek information and would pose the question to their partner, who might say, "That's an odd question." Hearing the word "that," the person would feel satisfied, despite not having received an explanation, let alone a good one. This example is obviously contrived, but it illustrates a key point: experiencing curiosity and satisfaction provide no guarantee that a good explanation will be found. Curiosity and satisfaction must be experienced in precisely the right moments.

What are "the right moments"? Obviously, the goal of explanation is not to answer "and" questions with "that" answers, making the above scenario implausible. In contrast, based on the research reviewed previously, explanatory phenomenology would be maximally useful to the extent that it motivates inquiry towards achieving the goal of "useful learning": the learning of general patterns that inform inference, learning, and action. In what follows, I outline three possible accounts of when and why humans are motivated to seek explanations, each of which proposes a different connection between explanatory phenomenology and the epistemic and instrumental goals of explanation. Then, I review prior research relevant to these accounts, which ultimately does not clearly adjudicate between them.

1.2.1 Theoretical accounts of explanatory phenomenology

An initial possibility, the **indiscriminate account**, predicts that curiosity and satisfaction are experienced indiscriminately: explanation-seeking curiosity might motivate us to seek any explanation, and we might be equally satisfied by any explanation we receive. If this were the case, then curiosity and satisfaction would only be associated with useful learning insofar as they motivate inquiry in the appropriate context. For example, if curiosity motivates indiscriminate inquiry in an "epistemically friendly" environment, where many explanations are in fact informative and useful (e.g., a classroom), inquiry would tend to lead to informative and useful

explanations—not because curiosity and satisfaction guide inquiry towards these explanations, but because good explanations are plentiful and easy to happen upon by chance. However, if the environment is less friendly (e.g., a politically biased blog), curiosity and satisfaction would not tend to guide inquiry towards useful learning.

Second, it is possible that curiosity and satisfaction are selective but misaligned with the goal of useful learning: the **goal-misaligned account**. As an analogy, food cravings are selective—we tend to crave sweet and salty snacks more than we crave green vegetables—but the selectivity of craving is misaligned with the ultimate goal of adequate nutrition. Likewise, curiosity and satisfaction might be experienced for some questions and explanations more than others, but the cues that trigger them are misaligned with the goal of useful learning. This could arise if, in humans' evolutionary or ontogenetic past, curiosity and satisfaction developed to support some non-epistemic or non-instrumental function (e.g., forming better arguments; see Mercier, 2016; Mercier & Sperber, 2011). This is in fact one account of modern overeating: in our evolutionary past, the consumption of large amounts of calorically dense food guarded against starvation during periods when food was scarce (Higginson et al., 2016). However, our cravings for calorically dense foods are no longer suited to our present-day environment. As is the case for the indiscriminate account, any link between curiosity, satisfaction, and useful learning on the goal-misaligned account would require a relatively friendly epistemic environment.

Third, it is possible that curiosity and satisfaction are selective in a way that is aligned with the goal of learning useful explanations: the **goal-aligned account**. If this were the case, we would experience curiosity about questions that are likely to lead to the learning of useful explanations, and we would experience satisfaction with explanations are epistemically or

practically valuable. On this account, an epistemically friendly environment is not required—the learner is able to seek and identify explanations that have epistemic value even if they are few and far between.

Finally, many **imperfectly goal-aligned accounts**, which fall on a continuum bounded by the goal-aligned and goal-misaligned accounts, are possible. For example, explanatory phenomenology might be aligned with individuals' *perceptions* of useful learning, so that we are curious about and satisfied with explanations that we believe to be epistemically or practically valuable (whether or not they actually are). Unfortunately, judgments of learning are not always a good guide to actual learning (Glenberg & Epstein, 1985; Lin & Zabrucky, 1998; Maki, 1998), so explanatory phenomenology would likely fall short of ideal "goal-alignment" on this account. Similarly, it is possible that curiosity and satisfaction are selective with respect to a set of heuristic guides to useful learning, which might be easier to compute than useful learning, more robust to error, or have some other non-epistemic benefit. Such heuristics would fall short in predictable ways. As an analogy, we might use the number of people eating at a restaurant as an easy heuristic cue to how good the food is, but this cue is imperfect—a restaurant might be crowded if it has a happy hour or live music, even if it has bad food. If curiosity and satisfaction are guided by heuristic cues, explanatory phenomenology may sometimes appear to be aligned with actual or perceived useful learning—but in fact, curiosity and satisfaction are shaped by imperfect cues that only approximate alignment with these goals.

1.2.2 Prior research on (non-)explanatory (non-)phenomenology

Unfortunately, very little research has directly addressed the elicitors of curiosity and/or satisfaction in an explanation-seeking context. However, a growing body of research has investigated curiosity about trivia facts, information search about facts or explanations (perhaps

motivated by curiosity), and evaluations of explanations (perhaps tied to satisfaction). I provide a brief review of this research on (non-)explanatory (non-)phenomenology to inform my approach to testing the accounts of explanatory phenomenology, though I provide more comprehensive discussion of this prior work in Chapters 2-4.

Some evidence appears to support a goal-aligned account of curiosity or exploration. When asking questions to achieve a specified learning goal, adults (and sometimes children) select questions that are reasonably well-described by optimal measures of expected learning (Coenen et al., 2015; Kachergis et al., 2017; Meder et al., 2019; Nelson, 2005; Nelson et al., 2010, 2014; Nussenbaum et al., 2020; Rothe et al., 2018; Ruggeri et al., 2016, 2017; for a review, see Coenen et al., 2019; Jones et al., 2020). Relatedly, adults decide which of several learning activities to complete by considering the rate at which learning has progressed in each activity (Ten et al., 2020; see also Oudeyer, 2018; Oudeyer et al., 2007). These findings could be consistent with the goal-aligned account: exploration (and perhaps curiosity) is guided by the probability of receiving information. However, expected learning has been confounded with other features in the majority of these tasks—for example, in some studies, the stimuli that provided the best opportunity for learning were also the ones about which individuals were the most uncertain (see Markant et al., 2015; Markant & Gureckis, 2014a). There is independent evidence that uncertainty can trigger curiosity and exploration (Baranes et al., 2014; Busch & Legare, 2019; Cook et al., 2011; Jepma et al., 2012; Kobayashi et al., 2019; Ligneul et al., 2018; L. E. Schulz & Bonawitz, 2007; van Lieshout et al., 2018), and uncertainty is likely a reasonable cue to expected learning. Nonetheless, the association between uncertainty and expected learning is not perfect: for example, static noise on a television screen evokes constant uncertainty, but being curious about it would surely be a mistake, as there is no possibility for learning (Gottlieb

et al., 2013; Schmidhuber, 2010). Because expected learning and uncertainty have not been empirically differentiated, particularly in the study of curiosity as opposed to information search, it is unclear whether these findings provide support for the goal-aligned account, or for a somewhat-aligned "heuristic" account.

Children and adults are also more curious or more inclined to seek information when they are surprised (Berlyne & Frommer, 1966; Bonawitz et al., 2012; Perez & Feigenson, 2020; Sim & Xu, 2017; Stahl & Feigenson, 2015, 2017; Vogl et al., 2019, 2020), encounter novel stimuli (Baranes et al., 2014; Berlyne & Frommer, 1966; Carvalho et al., 2018; Smock & Holt, 1962; Wittmann et al., 2008), and when social information suggests to-be-received information is valuable or interesting (Dubey et al., 2021; Y. Wu & Gweon, 2020). Again, these might point towards the goal-aligned use of expected learning as a cue to curiosity or information search, where surprise, novelty, and social information provide evidence that useful learning is likely to occur. On the other hand, these might constitute heuristic cues that are over-extended even when useful learning is unlikely: static noise on a television is also infinitely surprising and novel but should not trigger curiosity, and the reliability of social information depends on the knowledge state of one's social partner.

Finally, there is evidence that curiosity is triggered by several cues that are not obviously connected to learning at all. For example, individuals experience more curiosity about questions that have positive rather than negative valence (Marvin & Shohamy, 2016; see also Charpentier et al., 2018; Kobayashi et al., 2019) and when they generate a guess prior to receiving an answer (Brod & Breitwieser, 2019). These findings provide some support for the goal-misaligned account: curiosity might be sensitive to cues with no direct connection to useful learning.

On the basis of this evidence, the indiscriminate account appears to be implausible: there is clear selectivity in curiosity. However, the research that exists does not clearly discriminate between the goal-aligned account, the goal-misaligned account, and some account in between (like the heuristic account). Moreover, the bulk of this research has concerned the goal of learning, but very little research has considered the possible instrumental goals associated with information search—and research that has addressed this question has found conflicting results (Abir et al., 2020; Bramley et al., 2015; Dubey & Griffiths, 2020; Markant & Gureckis, 2012; Meder & Nelson, 2012). As a result, it is currently unclear what goal curiosity might be aligned with (learning or useful learning), if any.

Similar findings exist regarding satisfaction. For example, explanations are found more satisfying to the extent that they invoke a minimal number of unexplained causes (Lombrozo, 2007; Pacer & Lombrozo, 2017), and "simplicity" defined as such is lauded as one of the epistemic virtues of a good explanation (e.g., Friedman, 1974; Kitcher, 1989; Lombrozo, 2016). However, people appear use simplicity as a heuristic guide to useful learning (see Johnson et al., 2019), favoring simple explanations even when probabilistic considerations point to a complex alternative (Lombrozo, 2007; Pacer & Lombrozo, 2017). In addition, there is evidence that children are better able to remember explanations that are rated more satisfying by adults (Frazier et al., 2016; see also Fastrich et al., 2018; Marvin & Shohamy, 2016; McGillivray et al., 2015), perhaps suggesting that satisfaction is aligned with the goal of learning. However, satisfaction is also sensitive to a number of cues that appear unlinked to learning, supporting the goal-misaligned account. For example, people judge explanations as more satisfying or better when they include logically irrelevant neuroscience information (Weisberg et al., 2008), when they are longer (Weisberg et al., 2015), and when they appeal to a named category (Giffin et al.,

2017). In sum, as is the case for curiosity, the indiscriminate account appears implausible. However, existing evidence does not clearly adjudicate between the goal-aligned account, the goal-misaligned account, or something in between. In addition, instrumental goals and epistemic goals have not been clearly disentangled in the study of satisfaction.

1.3 The present research

To summarize, there is clear evidence that curiosity and satisfaction are selective, allowing us to reject the indiscriminate account. However, the existing evidence is fairly ambiguous between the goal-aligned account and a hybrid account that relies on perceptions of learning or heuristic cues. Further, even the goal-misaligned account has received some support. Thus, the correct account of explanatory phenomenology remains unresolved. To test these accounts, I report three sets of studies investigating the cues that shape explanation-seeking curiosity and explanatory satisfaction.

1.3.1 Approach

Testing these accounts of explanatory phenomenology relies on a reasonable operationalization of curiosity and satisfaction. Some researchers have argued that curiosity is a naïve or folk concept that is difficult or impossible to objectively define (Murayama et al., 2019; see also Kidd & Hayden, 2015), so researchers should focus on studying the processes underlying curiosity rather than curiosity itself (Murayama et al., 2019). Corresponding to these viewpoints, curiosity has often been operationalized and defined using behavioral measures. For example, curiosity has been defined by participants' choices between several pieces of information (e.g., Kobayashi et al., 2019) or choices to pay a cost to receive a piece of information (e.g., Dubey & Griffiths, 2020). As a consequence, all acts of information search in these paradigms are taken to signal the operation of curiosity. Similar arguments have not been

made for satisfaction, but the fact that various measures have been used to quantify the "goodness" of an explanation—including satisfaction (e.g., Giffin et al., 2017; Khemlani et al., 2011; Lim & Oppenheimer, 2020; Lombrozo, 2007; Lombrozo & Carey, 2006; Pacer & Lombrozo, 2017; Weisberg et al., 2008), quality (e.g., Bechlivanidis et al., 2017; T. Blanchard, Lombrozo, et al., 2018; Hopkins et al., 2016; Vasilyeva et al., 2017; Zemla et al., 2017), and probability (e.g., T. Blanchard, Lombrozo, et al., 2018; Johnson et al., 2019; Khemlani et al., 2011; Khemlani & Johnson-Laird, 2011)—implies that these judgments are seen as relatively interchangeable, despite the fact that only one (satisfaction) directly measures a phenomenological state.

In this dissertation, I take a different approach to the study of explanation-seeking curiosity and explanatory satisfaction. Following the definitions provided earlier, I take curiosity and satisfaction to be defined by their phenomenology: curiosity is not synonymous with information search and satisfaction is not synonymous with a goodness judgment; they are mental states with a particular phenomenology. Like emotions, these phenomenological states tend to have measurable behavioral consequences (e.g., happiness is often accompanied by smiling, curiosity is often accompanied by information search). However, the behavioral consequences can also arise when the mental state is not present (e.g., we might smile in the absence of happiness, we might seek information in the absence of curiosity), and the mental state can arise without the behavioral consequence following (e.g., we might feel happy but not smile, we might feel curious but not explore). Thus, measuring information search as a signal of curiosity (or quality as a signal of satisfaction) is misleading. Moreover, I assume that people have some introspective access to the phenomenology of inquiry, like they do for emotions. On the basis of this assumption, I measure curiosity and satisfaction by self-report across all studies,

and I take these reports to be a reliable indicator of the relevant phenomenology. However, I also validate these measures by testing the association between self-reported curiosity and information search (Chapter 2) and self-reported satisfaction and subsequent inquiry (Chapter 4).

Research on explanation-seeking curiosity and explanatory satisfaction also navigates a trade-off between ecological validity and experimental control (like most empirical work in psychology). Some research on explanatory satisfaction has used controlled experimental designs that are somewhat removed from everyday experience—for example, contrasting explanations for symptoms of a fictitious alien disease (Johnson et al., 2019; Lombrozo, 2007; Pacer & Lombrozo, 2017). Other studies have maintained reasonable experimental control while using more realistic explanations (e.g., Bechlivanidis et al., 2017; Giffin et al., 2017; Hopkins et al., 2016; Khemlani & Johnson-Laird, 2011; Lombrozo & Carey, 2006; Weisberg et al., 2008), while a small subset of studies has examined evaluations of real-world explanations (Frazier et al., 2009; Kurkul & Corriveau, 2018; Zemla et al., 2017). Research on curiosity has also spanned this range, with some studies investigating information search about highly controlled games of chance (e.g., Kobayashi et al., 2019), other studies investigating curiosity about real-world trivia questions (e.g., Kang et al., 2009), and still other studies investigating curiosity in real-world educational settings (e.g., Bernacki & Walkington, 2018). In the present research, I navigate this trade-off by studying curiosity and satisfaction with a diverse set of real-world questions and explanations in Chapters 2 and 4, as well as studying curiosity in a well-defined experimental task in Chapter 3.

1.3.2 Summaries of chapters

In the following chapters, I address several questions, which are designed to adjudicate between the goal-aligned and the goal-misaligned accounts of explanatory phenomenology, and

to define the contours of a plausible in-between account (e.g., on the basis of perceived useful learning or heuristic cues). In addition to differentiating these accounts, I probe what goal curiosity and satisfaction might be aligned with: learning information regardless of its instrumental value (epistemic goal) or learning for the purpose of gaining useful information (instrumental goal).

In particular, I address these questions:

1. When expected useful learning and alternative heuristic cues are disentangled, is explanation-seeking curiosity aligned with the goal of useful learning?
2. To the extent that curiosity is aligned with the goal of useful learning, is it aligned with the actual likelihood of learning, or subjective expectations about learning?
3. Is curiosity aligned with epistemic goals (learning in general) or instrumental goals (learning useful information)?

I address these questions in Chapter 2 and Chapter 3. In Chapter 2, I test whether curiosity ratings in response to explanation-seeking and fact-seeking questions are associated with participant judgments of surprise, uncertainty, novelty, expected learning, expected utility, and other relevant measures. I find that perceived expected learning and perceived expected utility both explain variance in curiosity above and beyond heuristic cues. In Chapter 3, I test whether curiosity about an explanation-seeking question over the course of a learning task tracks trial-by-trial changes in surprise, uncertainty, expected learning, and expected instrumental utility as operationalized by a Bayesian model. I find that uncertainty or expected learning about the question of interest (which are confounded) explain the most variance in curiosity, while the role of surprise, uncertainty about the ultimate learning goal, and expected learning about the ultimate learning goal are more variable. In addition, the instrumental value of information has a

fragile influence on curiosity. Together, these findings suggest that heuristic cues and perceptions of useful learning likely play the strongest role in guiding curiosity, though actual expected useful learning sometimes contributes, as well.

In addition, I ask:

1. Is explanatory satisfaction aligned with the goal of useful learning?
2. To the extent that satisfaction is aligned with the goal of useful learning, is it aligned with actual useful learning, or perceived useful learning?
3. Is satisfaction aligned with epistemic goals (learning in general) or instrumental goals (learning useful information)?

I address these questions in Chapter 4. I test whether satisfaction ratings in response to explanations are associated with participant judgments of learning, utility, and other explanatory virtues (simplicity, breadth). In addition, I test whether satisfaction ratings are associated with actual learning, using multiple objective learning assessments. I find that variance in satisfaction is partly explained by participants' judgments of learning and utility, and these learning judgments explain the effect of one goal-misaligned cue from prior research. However, satisfaction is only associated with actual learning in particular settings and using particular learning measures. Together, these findings suggest that satisfaction is imperfectly aligned with the goal of useful learning, generally co-occurring with perceived learning and perceived instrumental utility rather than actual learning.

Finally, in Chapter 5, I summarize the contributions of this dissertation, and I discuss promising directions for future research.

2 CHAPTER 2: A FUNCTIONAL APPROACH TO EXPLANATION-SEEKING CURIOSITY*

2.1 Introduction

Why do humans ask why? Much of human learning, from childhood through adulthood, is achieved by asking questions. Children begin to ask information-seeking questions before the age of two, and they ask increasingly more *explanation*-seeking questions (often "why" questions) between the ages of two and three (Chouinard, 2007; Hickling & Wellman, 2001). Explanation remains an important mechanism for learning throughout the human lifespan: children ages two to six use self-generated explanations to guide their exploration (Legare, 2012), and adults use explanatory principles to guide their inferences (Lombrozo, 2016).

Despite the prevalence of "why" questions and the importance of explanation, humans do not always ask questions when they lack an explanation. In fact, there are many things in the world we *do not* question: why a particular tree is 57 feet tall, why dogs have eyes, or why a certain conference room has an odd number of chairs (to offer only a small number of possible cases). These examples illustrate that our *explanation-seeking curiosity* (ESC) is highly selective—humans are "why-curious" about some things, but not about others. Of course, there

* A version of this chapter was published in Liquin, E. G., & Lombrozo, T. (2020). A functional approach to explanation-seeking curiosity. *Cognitive Psychology, 119*, 101276. https://doi.org/10.1016/j.cogpsych.2020.101276. An earlier version of this work was published in Liquin, E. G. & Lombrozo, T. (2018). Determinants and consequences of the need for explanation. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 696-701). Cognitive Science Society. Components of this work were also presented in the following: Liquin, E. G. & Lombrozo, T. (2019, March). Determinants of explanation-seeking curiosity across development. Talk presented at the biennial meeting of the Society for Research in Child Development, Baltimore, MD. Liquin, E. G. & Lombrozo, T. (2018, July). Determinants and consequences of the need for explanation. Talk presented at the annual meeting of the Cognitive Science Society, Madison, WI. Liquin, E. G. & Lombrozo, T. (2018, July). What demands an explanation? Expertise and utility drive explanation search. Talk presented at the annual meeting of the Society for Philosophy and Psychology, Ann Arbor, MI.

may be some circumstances under which these observations would prompt explanation-seeking: for example, if a 57-foot tree were surrounded by trees that were 56 feet tall, or for a person who expects trees of that species not to exceed 50 feet in height. In other words, ESC is highly context-dependent and person-dependent, as well. However, little empirical work has addressed what distinguishes those things that elicit why-curiosity from those that do not, or what explains why ESC differs across contexts and individuals.

In this chapter, we develop and test a novel set of predictions about what triggers explanation-seeking curiosity. We call our approach a "functional" approach because our predictions derive from considering the functional consequences of explanation-seeking within the process of inquiry. Within this process, observations or facts will sometimes spark a learner's explanation-seeking curiosity, and thereby motivate explanation search, whether it takes the form of reflection, exploration, or asking others. Importantly, explanation-seeking has a distinctive profile of functional consequences: as we argue below, explanation-seeking facilitates learning about the explanandum (i.e., what is being explained), learning for export (i.e., to generalize beyond the explanandum), and learning from the domain-based expertise of others. If ESC is in part sensitive to how likely it is that these functional consequences will be achieved, then *expectations* about these consequences—whether implicit or explicit—could guide when ESC is experienced and to what degree. In other words, learners should experience (more) ESC for observations or facts that, if explained, would likely support learning about the explanandum, for export, and/or from the domain-based expertise of others.

We begin, in the section that follows, by noting the most relevant previous research on information search and curiosity. We then turn to previous literature on the functional consequences of explanation search, motivating the three functions indicated above. For each

function, we derive corresponding triggers of ESC, many of which have not been posited or investigated in prior research. We next evaluate these new triggers, along with potential triggers from prior research, by testing how well they predict ESC across diverse sets of questions. Specifically, we report three experiments employing a novel method that combines the diversity and ecological validity of real-world stimuli with the rigors of the lab: we sample explanation-seeking questions posed by users on the internet, and we have new groups of participants indicate their curiosity about the answers, evaluate the questions for the presence of candidate triggers, and/or reveal a subset of answers. In this way we can identify which triggers predict ESC. Finally, we explore (in Study 3) how well these potential triggers predict *fact-seeking curiosity* (FSC) in comparison to ESC, to help define the contours of our approach.

2.1.1 Prior approaches to curiosity and information-seeking

Mirroring our functional approach, a subset of prior work on question-asking and curiosity has focused on learners' expectations about the consequences of seeking information. We characterize such considerations as "forward-looking," as they involve an evaluation of the expected outcomes of pursuing information, as opposed to "backward-looking" considerations (such as novelty or surprise) which involve an evaluation of how a stimulus fits in with a learner's prior beliefs. For example, research on question-asking has considered how questions are shaped by their potential to solve problems (Chouinard, 2007; Legare et al., 2013) and reduce uncertainty (Rothe et al., 2018; Ruggeri et al., 2016, 2017; Ruggeri & Lombrozo, 2015).[2] Turning to curiosity, the influential *information gap* theory (Loewenstein, 1994) is also

[2] A body of research in social psychology has also considered the social and emotional consequences of engaging in explanation (Anderson et al., 1996; Malle, 2004). For example, an explanation could redirect blame, persuade, impress, or make one feel better about oneself. These are surely important psychological roles, but they are beyond the scope of the current paper, which focuses on the role of explanation in the process of inquiry, and hence on more epistemic functions.

articulated as forward-looking, focusing on the discrepancy between one's current knowledge state and one's desired knowledge state. According to this proposal, curiosity is sparked by a modest information gap—that is, one large enough to indicate there's something to be learned, but not so large that satisfying one's curiosity seems unlikely. More recently, Dubey and Griffiths (2020) have proposed a rational theory of curiosity according to which curiosity functions to increase the value of an agent's current beliefs regarding the phenomena the agent expects to encounter in the future.

These forward-looking proposals have not been extended to explanation-seeking in particular, and thus have not considered functional consequences that may be unique to explanation search as compared to information search more generally. More importantly, the small handful of papers that have focused on explanation-seeking (as opposed to "fact-seeking") have investigated "backward-looking" triggers of information search, not an inference to anticipated (future) consequences. Both classic (e.g., Berlyne, 1950, 1966) and contemporary research (for a review, see Gottlieb et al., 2013) suggest that curiosity can be triggered by *novelty* and *surprise*, and these factors can indeed prompt explanation search, as well (Frazier et al., 2009). Even the information gap theory, which has also been applied to explanation-seeking behavior (Mills et al., 2019), has typically been tested by focusing on how confident participants are that they know the answer to a question (e.g., Gruber et al., 2014; Kang et al., 2009) or by how completely a question has been answered (Mills et al., 2019), both evaluations involving the relationship between the question and current beliefs, and not necessarily expectations about the consequences of seeking an answer.

Two theoretical proposals concerning explanation-seeking, both from philosophy, similarly focus on backward-looking triggers, though they emphasize the forward-looking goals

these triggers serve. These accounts address "need for explanation"—the sense that a given event or phenomenon demands an explanation. Wong and Yudell (2015) propose that it is normatively correct to seek an explanation when a given event or phenomenon does not fit one's "map" of the world, a judgment which we refer to as *"map mismatch."* Relatedly, Grimm (2008) suggests that one experiences this sense of need for explanation when one can easily imagine an alternative way the world could have been—that is, there is a salient contrast between *fact* (what actually occurred) and *foil* (what could have occurred instead). Both judgments involve an evaluation of the relationship between an event or phenomenon and prior beliefs (and are thus backward-looking), but the authors highlight their connection to subsequent epistemic consequences. For example, for Grimm, upon receiving an explanation that differentiates between the fact and the foil, "the proto-understanding in terms of which we experience the world would, optimistically, be approximating ever more closely the various ways the world actually might be" (Grimm, 2008, p. 495).

In sum, prior research on question-asking has considered the expected consequences of pursuing information, but almost never in the context of explanation-seeking; research on fact-seeking and explanation-seeking curiosity, by contrast, has almost always relied on backward-looking cues (namely surprise, novelty, information gap, map mismatch, and fact-and-foil; see Table 1). In the following section, we bring a forward-looking, functional approach to the forefront, focusing in particular on the functional consequences of pursuing explanations.

2.1.2 Functional consequences of explanation search

In this section, we consider the functional consequences of seeking an explanation, and we derive candidate triggers of ESC on the basis of these consequences. Our guiding idea is that (a) ESC triggers explanation search, and (b) explanation search has a distinctive profile of

downstream consequences. Under these assumptions, if ESC is well-calibrated with respect to the downstream consequences of explanation search, we would expect ESC to be triggered precisely when a learner judges it likely that these consequences will come about. In this way, identifying the functional consequences of engaging in explanation search can be used to generate testable hypotheses about potential triggers of ESC.

In the following sections, we review evidence to support three functional consequences of explanation search: learning about the explanandum (what is being explained), attaining information that is "exportable" beyond the explanandum, and learning from the domain-based expertise of others. We derive multiple candidate triggers from each function, as summarized in Table 1.

Learning about the explanandum

The first functional consequence of explanation search is obvious but important: explanation search, especially when successful, has the consequence of providing information about the explanandum (that is, the fact or event being explained), and in particular about how or why it is the case. For example, the explanation answering the question "Why is rust red?" will provide information about the properties of rust, and the inquirer is therefore likely to learn something new about the target of the explanation (rust, redness, and associated causal mechanisms). Learning about the explanandum can also occur when an individual "self-explains," even in the absence of feedback (Chi et al., 1994). These findings motivate predictions about the triggers of ESC: ESC should be triggered whenever a person anticipates learning something new about the explanandum by engaging in explanation search (an idea that we refer to as *anticipated learning*, Table 1, Row 1), or when the explanation is expected to support

learning, independent of the person's existing knowledge (an idea that we refer to as *learning potential,* Table 1, Row 2).

This prediction aligns with previous research on information search and question-asking. Lindley (1956) formalized a measure of "expected information gain" that identifies the evidence one should collect to maximally reduce uncertainty in the space of candidate hypotheses. Since the introduction of this forward-looking measure to psychology (Oaksford & Chater, 1994), several studies have shown that adults' exploration is sensitive to expected information gain and thus the future value of information (e.g., Bramley et al., 2015; Coenen et al., 2015; Steyvers et al., 2003). Furthermore, adults can evaluate (but not always generate) maximally informative questions in a Battleship-like game, with the goal of finding the location of several concealed ships on a game board (Rothe et al., 2018). Even preschoolers can select the more effective of two questions for solving a particular problem as measured by expected information gain (Ruggeri et al., 2017), and older children can often generate such questions themselves (Ruggeri & Lombrozo, 2015). These studies suggest that anticipated learning drives the evaluation of questions in tasks with well-defined goals; no prior work (to our knowledge) has investigated whether it affects the experience of curiosity about why or how something is the case.

Learning for Export

Seeking explanations not only supports learning about the explanandum: it can also support learning that is "exportable" to other situations (Lombrozo & Carey, 2006). For example, an explanation for why some people are allergic to cats is likely to reveal general facts about the immune system and individual differences that are applicable well beyond the explained example. Indeed, self-explanation prompts are often beneficial (relative to other tasks) in helping students *transfer* the principle or procedure from explained problems to novel problems (Aleven

& Koedinger, 2002; Atkinson et al., 2003) or apply what they've learned to make novel inferences (Chi et al., 1994). Evidence suggests that these consequences are not incidental features of explanation-seeking, but rather arise because generalizability is among the criteria for what makes an explanation good. For instance, explanatory claims are judged more appropriate when they identify relationships that are "stable" in the sense that they hold despite variation in background circumstances, which should generally track exportability to new contexts (Vasilyeva et al., 2018; see also T. Blanchard, Vasilyeva, et al., 2018).

These ideas motivate the prediction that ESC should be triggered by evidence that a to-be-received explanation will be exportable and applicable, as pursuing explanations under these conditions should increase the odds of obtaining good (i.e., exportable) explanations. More precisely, we might expect that people will be more curious about the answer to an explanation-seeking question when they expect it to pick out a genuine pattern or *regularity* that extends beyond the explanandum (Table 1, Row 3), and when they expect it to have high *future utility* (Table 1, Row 4). Dubey and Griffiths (2020) similarly predict effects of expected future utility on curiosity, but this prediction has not been evaluated in an explanation-seeking context.

Considerations regarding export also motivate more specific predictions concerning the roles of simplicity and breadth as triggers of ESC. Research suggests that all else being equal, people prefer explanations that are simple and broad (see Lombrozo, 2016 for a review). For example, when prompted to explain (rather than perform a control task), children and adults are more likely to discover patterns that are broad in scope (Walker, Lombrozo, et al., 2017; Williams et al., 2013; Williams & Lombrozo, 2010) and simple in the sense that they appeal to few features (Kon & Lombrozo, in prep; Walker, Bonawitz, et al., 2017). Importantly, such patterns may be more exportable than narrow and complex alternatives: an explanation that is

broader in scope will necessarily explain more observations (Lombrozo & Carey, 2006), and simpler explanations may be more likely to generalize and support future communication and intervention (Pacer & Lombrozo, 2017). Correspondingly, we might expect ESC to be triggered when explanation-seeking is anticipated to help the learner identify a pattern that is expected to be *simple* (Table 1, Row 5) and *broad* (Table 1, Row 6), as these are the kinds of patterns that explanation-seeking seems to privilege.[3]

Learning from others' domain-based expertise

The preceding considerations (learning about the explanandum and learning for export) are likely to characterize explanation search quite broadly, whether the learner attempts to generate an explanation herself, engages in exploration to uncover an explanation, or asks a knowledgeable informant. However, this third method of explanation search—asking a knowledgeable informant—might introduce additional factors that could trigger ESC. A great deal of learning involves the testimony of others (Harris, 2012; Harris & Koenig, 2006), and testimony can often communicate content that is difficult to uncover through direct observation or exploration, such as causal structure or unobservable mechanisms. In fact, there's evidence that even young children will explore observable entities directly, but instead *ask questions* about unobservable entities (Fitneva et al., 2013). Moreover, there is good evidence that people are adept at exploiting the "division of cognitive labor" (e.g., Keil, 2003; Keil et al., 2008; Lutz &

[3] It is unclear exactly how a learner would generate expectations about the simplicity or breadth of an explanation they have not yet received. However, there is already evidence that both children and adults have systematic expectations about the simplicity/complexity of causal mechanisms (Kominsky et al., 2018), as well as theoretical arguments to the effect that if pursuing an explanation is expected to have high epistemic utility (because it would be a good explanation if true), this could support its pursuit (Nyrup, 2015).

Keil, 2002), deferring to appropriate experts in a given domain depending on what they want to know.

While deference to experts is likely to characterize question-asking in general, we might expect domain-based expertise to be especially relevant when it comes to explanation-seeking questions, given lay beliefs that domains are united by distinct patterns of causes and effects (Strickland et al., 2017). In addition, there is evidence that people's willingness to defer to others to answer questions within a given domain is related to their attributions of explanatory understanding: attributions that another person understands why something in that domain occurred more closely track participants' willingness to defer to that person than do attributions that the person knows why (Wilkenfeld et al., 2016). Furthermore, adults' judgments of the quality of an explanation are predicted by their judgments of whether the explanation was written by an expert and their judgments of the explanation's complexity (Zemla et al., 2017). Finally, children and adults are more likely to ask for help using or fixing a device that is judged to have a complex mechanism than a device that is judged to have a simple mechanism (Kominsky et al., 2018).

On our forward-looking approach, expectations about the necessity or value of consulting others could itself drive explanation-seeking curiosity. Specifically, if one way to pursue ESC is by asking knowledgeable others, then we might expect ESC to be triggered when *expertise* in some domain is needed to answer a given question (Table 1, Row 7), and/or when the answer is anticipated to be *complex* (Table 1, Row 8). Under these conditions, the learner might not expect to discover the answer on their own, but explanation-seeking directed towards domain-based experts can exploit the division of cognitive labor and support effective learning. While prior work shows that children and adults have the capacity to form expectations about expertise and

complexity (e.g., Keil et al., 2008; Kominsky et al., 2018; Lutz & Keil, 2002), research has not investigated whether these expectations guide ESC.

2.1.3 The present research

Having articulated several potential triggers of ESC based on the downstream consequences of explanation seeking, and having reviewed several backward-looking triggers from prior work on question-asking, curiosity, and the need for explanation, we next turn to an empirical evaluation.

As described above, several of the potential triggers of ESC that we test have received previous empirical support, primarily for their role in driving information search more broadly (surprise, novelty, information gap, anticipated learning, and future utility). With the exception of surprise (Frazier et al., 2009) and information gap (Mills et al., 2019), however, prior work has not evaluated these factors as triggers of explanation-seeking curiosity or explanation search, in particular. Thus, one contribution of our work is to evaluate whether these candidate triggers indeed prompt ESC and explanation search. More importantly, however, our functional approach to ESC motivates a set of potential triggers that have not (to our knowledge) been tested as drivers of general curiosity and information search or of ESC and explanation search: expected simplicity/complexity, expected breadth, learning potential, and expected regularity. Additionally, while the literature on "need for explanation" has proposed normative standards for when an event *should* demand explanation, these standards (map mismatch, fact-and-foil) have not been empirically evaluated as triggers of ESC. A second contribution of our research is thus to evaluate whether these novel triggers prompt ESC and explanation search. Moreover, by testing all hypothesized triggers simultaneously, we can identify the relative contribution of each, and we can uncover patterns of shared variance that might indicate a meaningful, latent structure.

These are important advantages over a more piece-meal approach focusing on the role of a single trigger.

Importantly, we can differentiate two hypotheses about the factors that drive ESC and explanation search. First, it is possible that a learner need not consider forward-looking triggers (such as anticipated learning and future utility) to successfully seek explanations in a way that achieves the downstream consequences of explanation. This is because backward-looking triggers, such as surprise and novelty, may be good (enough) proximal cues to the downstream consequences of explanation, obviating the need for a learner to also form expectations about anticipated learning and export. This hypothesis motivates the prediction that forward-looking triggers should play little to no role in predicting explanation-seeking curiosity once backward-looking triggers are considered. However, it is also possible that these forward-looking triggers have some added value for a learner above and beyond the backward-looking triggers, in which case both forward- and backward-looking triggers should predict a learner's experience of ESC. In the present research, we test these two hypotheses.

Methodologically, we also address a common limitation of previous research on explanation-seeking behavior. As spontaneous explanation-seeking is, by its very nature, difficult to capture in the lab, research has typically been motivated by a small number of realistic cases or relied on highly artificial contexts in the lab. The present research circumvents this challenge by drawing upon a novel source of data: large-scale on-line databases of user-generated questions that allow us to better capture explanation search in the course of everyday life. Specifically, we sample why-questions from on-line forums (e.g., "Why aren't clouds a uniform sheet of vapor in the sky?"), and we use these as stimuli in studies that prompt participants to report their curiosity about the explanandum (e.g., about why clouds aren't a

uniform sheet of vapor in the sky) and to evaluate the question and explanandum along a variety of dimensions designed to track hypothesized triggers of ESC. This approach is correlational, which limits our ability to make claims about whether the hypothesized triggers *cause* a learner to experience ESC. However, the strengths of this approach—capturing a range of real-world explanation-seeking and testing many potential triggers of ESC in a single set of studies—offset this drawback.

In three studies, we address the following questions: (1) How well do our hypothesized triggers of ESC (Table 1) predict both explanation-seeking curiosity and explanation search itself? (2) Do forward-looking triggers contribute to the prediction of ESC above and beyond backward-looking triggers? (3) Is there some meaningful *structure* to the predictors of ESC (i.e., in terms of latent factors) that sheds light on function(s) of ESC? (4) How do the potential triggers of ESC compare to the potential triggers of fact-seeking curiosity?

In Study 1, we address the first and second questions by asking participants to rate ESC for a series of "why" questions, along with measures for the potential triggers of ESC identified in the prior sections. Additionally, we ask participants to select a subset of questions for which they would like to receive answers, thus allowing us to assess actual explanation-seeking behavior. In Study 2, we address the second and third questions by performing exploratory factor analysis on the potential trigger measures, using a stimulus set representing a broader sample of possible "why" questions. Finally, in Study 3, we answer the fourth question by comparing the triggers of ESC to the triggers of FSC, using fact-seeking questions from previous research on curiosity (Dubey & Griffiths, 2020; Kang et al., 2009) in Study 3a, and using a novel set of both explanation-seeking and fact-seeking questions in Study 3b.

Table 1 Potential Triggers of Curiosity, Studies 1-3

Potential triggers (with their primary functional consequences as applicable), and items used to measure each potential trigger in Studies 1-2 and Study 3.

Potential Trigger	Item, Studies 1-2	Item, Study 3
Function: Learning about the explanandum		
Anticipated Learning	To what extent do you think the answer to this question would teach you something new?	To what extent do you think the answer to this question would teach you something new?
Learning Potential	Do you think there is something to be learned from the answer to this question (even if you yourself already know the answer)?	Do you think there is something to be learned from the answer to this question (even if you yourself already know the answer)?
Function: Learning for export		
Regularity	Do you think answering this question would help reveal a genuine pattern, structure, or regularity?	Do you think answering this question would help reveal a genuine pattern, structure, or regularity?
Future Utility	To what extent would knowing the answer to this question be useful to you in the future?	To what extent would knowing the answer to this question be useful to you in the future?
Simplicity (Reverse-Scored Complexity)	Do you think the answer to this question is likely to be simple or complex?	Do you think the answer to this question is likely to be simple or 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)?	Do you think the answer to this question is narrow (only applies to what this question is asking about) or broad (also applies to other similar cases)?
Function: Learning from others' domain-based expertise		
Expertise	Do you think that answering this question correctly requires special expertise in some domain?	Do you think that answering this question correctly requires special expertise in some domain?
Complexity (Reverse-Scored Simplicity)	Do you think the answer to this question is likely to be simple or complex?	Do you think the answer to this question is likely to be simple or complex?
Backward-looking triggers from prior research		
Surprise	To what extent do you think it is surprising that *[premise]*?	Is this question asking about something surprising?
Novelty	How novel for you is the claim that *[premise]*?	Is this question asking about something that you find novel or new?
Information Gap*	How confident are you that you know the answer to this question?	How confident are you that you know the answer to this question?
Map Mismatch*	How well does the claim that *[premise]* fit in with your current beliefs about the world?	Does what this question is asking about fit in with your current beliefs about the world?
Fact-and-Foil	How easily can you imagine a world in which it is not the case that *[premise]*?	
Additional triggers		
Negative Valence*	To what extent do you think the main claim of the question (that *[premise])* is negative, positive, or neutral?	Is this question asking about something negative or positive?

* *reverse-scored*

Broadly, these studies advance our understanding of ESC, both in terms of its relationship to other forms of curiosity and its potential triggers. Most notably, they highlight the importance of forward-looking considerations in guiding ESC above and beyond backward-looking considerations, and they shed light on the functional role of ESC and explanation seeking within the process of inquiry.

2.2 Study 1

In Study 1, we present participants with several questions and ask them to rate how curious they are about the answer to each question. Additionally, they rate each question on 13 potential trigger items, based on the potential triggers of ESC identified in the Introduction (see Table 1). We additionally include "negative valence" (Table 1, Row 13), a trigger motivated by work in attribution.[4] Subsequently, participants are asked to select the questions for which they would *most* like to receive an answer, thus providing the opportunity to seek actual explanations. This design allows us to assess whether ratings of ESC predict actual explanation search, validating our measure of ESC. In addition, we can test whether the 13 potential triggers identified in the Introduction do, in fact, predict ESC and explanation search. Finally, we investigate whether forward-looking triggers increase our ability to predict ESC beyond backward-looking triggers.

2.2.1 Method

Participants

Participants in all studies were recruited from Amazon Mechanical Turk and paid at the rate of $7.50 per hour, based on the expected average time to complete the task (20 minutes for

[4] Prior work on attribution shows that people tend to generate attributions for outcomes with negative emotional valence (Weiner, 1985b; P. T. Wong & Weiner, 1981)—for example, a person is more likely to explain a failing grade on a test than a passing grade. While this potential trigger is not obviously forward- or backward-looking in nature, we include it due to its potential relevance to explanation search.

Study 1 and 16 minutes for Studies 2-3). All participants provided informed consent under a protocol approved by the Institutional Review Board at University of California, Berkeley (Studies 1 and 2) or a protocol approved by the Institutional Review Board at Princeton University (Study 3). Participation in Studies 1 and 2 was restricted to users in the United States with a minimum approval rating of 95% based on at least 50 previous Mechanical Turk tasks.

The sample size for Study 1 was determined using the "simr" package for R (Green & MacLeod, 2016), simulating data based on a similar study from Liquin and Lombrozo (2018). This analysis suggested 80% power to detect an effect of ESC on explanation search with sample sizes between 150 and 200 participants. Our final sample for Study 1 included 195 adults (97 males and 98 females) ranging from 19 to 78 years of age (M_{age} = 35, SD = 11). Twenty-five additional participants were excluded for failing an attention check (explained below).

Materials

Forty "why" questions were sampled from the Reddit *Explain Like I'm Five* (ELI5) webpage (www.reddit.com/r/explainlikeimfive). On this page, Reddit users submit questions and other users provide easy-to-understand explanations. The questions were randomly sampled (from all questions posted between September and December 2017) and edited lightly for grammar and readability. Sample questions include: "Why are most pencils painted yellow? Why not a different color?"; "Why do ice cubes crackle when liquid is poured on them?"; and "Why does the bowl always get hotter than the food in the microwave?" For each question, we selected an answer provided by another user on the basis of its succinctness and accuracy; these were also edited lightly for grammar and readability. (See https://osf.io/gyuqa/ for a complete set of questions and answers.)

Procedure

Participants were randomly assigned to eight questions from the ELI5 stimulus set. For each question, participants were asked to rate their curiosity in response to that item: "Consider the following: [that ice cubes crackle when liquid is poured on them]. How curious are you about why this is the case?" This rating was completed on a seven-point scale, with the endpoints marked "Not at all curious about why" and "Very curious about why." Additionally, participants rated each question on 13 other measures corresponding to the 13 potential triggers of ESC (with simplicity and complexity collapsed to a single dimension), summarized in Table 1. These 14 judgments (including ESC) were presented in a random order.

After completing these ratings, participants completed a short distractor task, which involved performing simple arithmetic problems. This task doubled as an attention check, as the task required participants to remember what they had seen previously (the answer to each arithmetic problem was an argument in the next arithmetic problem). Participants who made more than two errors were excluded.[5]

After the distractor task, participants were presented with the eight questions they had rated, and were instructed, "You get to see the answers to three of these questions. Please choose the three questions that you want answered." Participants read the explanations to the three questions they selected, after which they were asked to provide their age and gender (male, female, other, or prefer not to specify).

[5] While the primary function of this attention check was to ensure that participants were reading instructions and paying attention, it could be argued that working memory and arithmetic abilities are not diagnostic of attentive completion of this study. As a result, we also conducted all analyses without any exclusions. The pattern of results reported remained unchanged.

2.2.2 Results

Data preparation

Across all studies, listwise deletion was used to eliminate missing data for any of the potential trigger measures or ESC. Unless otherwise specified, all variables were z-scored to facilitate the interpretation of regression model parameters across studies. Descriptive statistics for curiosity and all potential trigger measures (across all studies) are reported in Appendix A.

ESC and explanation search

First, we tested whether explanation search (i.e., revealing the answer to a question in the final phase of the study) could be predicted by ESC ratings. A generalized linear mixed model was fit to the answer reveal data (using the R package "lme4," Bates et al., 2015, with a logit link function). The model contained ESC as a fixed effect, with random intercepts for participant and item.[6] This model was compared to a null model, which contained no fixed effects, using a likelihood ratio test. This analysis revealed a significant effect of ESC, OR = 1.69, 95% CI [1.51, 1.90], $\chi^2(1) = 84.94$, $p < .001$. This indicates that our measure of ESC predicts explanation search, and thus supports the validity of our measure of ESC.

Triggers of ESC

Next, we tested whether the 13 potential triggers predicted ESC. To test how well each trigger predicted ESC accounting for all other measures, we fit a regression model to the rating

[6] Following the recommendation of Barr et al. (2013) and an anonymous reviewer, we attempted to fit all models with by-participant and by-item random slopes, in addition to random intercepts. In nearly all cases, the model with random slopes and random intercepts resulted in singular fit, indicating that the random effects structure was too complex to be supported by the data. Following the recommendation of Singmann & Kellen (2019), we subsequently simplified the models, first eliminating by-item random slopes, then all random slopes. In most cases, only the fully simplified (random-intercept only) models converged, so we report these models. In cases where more complex models converged, the parameter estimates for fixed effects were comparable to the random-intercept only models. All alternative models are included in our analysis scripts at https://osf.io/gyuqa/.

data. The potential triggers were fit as fixed effects, and random intercepts were fit for participant and item. The contribution of each potential trigger to the model was calculated by comparing the full model to a reduced model excluding that measure, using likelihood ratio tests. Seven of the 13 potential triggers made a significant contribution to the model predicting ESC, holding all other potential triggers fixed at their mean values: anticipated learning, learning potential, future utility, expertise, surprise, information gap,[7] and map mismatch (see Figure 1; numeric values for all regression coefficients and 95% confidence intervals depicted in Figure 1 are provided in Appendix B).

We also investigated whether the 13 potential triggers were related to explanation search. Fitting a generalized linear mixed-effects model to the answer reveal data showed that anticipated learning, OR = 1.44, 95% CI [1.19, 1.71], $\chi^2(1) = 16.66$, $p < .001$, surprise, OR = 1.17, 95% CI [1.01, 1.37], $\chi^2(1) = 4.80$, $p = .03$, and information gap, OR = 1.26, 95% CI [1.10, 1.45], $\chi^2(1) = 11.58$, $p < .001$, made a significant contribution to the full model above and beyond the other potential trigger measures. The remaining measures did not reach statistical significance.

Because many of the 13 potential trigger measures were modestly correlated with each other, the failure of some measures to reach significance could be due to their shared variance, rather than their lack of relation to ESC. To address this, we also tested to what extent each

[7] We also tested whether the quadratic effect of information gap on curiosity was significant, given the proposed "U-shaped curve" relating the magnitude of the information gap to curiosity (Loewenstein, 1994). To do so, we fit a multi-level model with random intercepts for participant and item, and fixed linear and quadratic terms for information gap. This quadratic model was compared to a model excluding the quadratic term using a likelihood ratio test. The results of this test revealed a significant quadratic effect, $\chi2(1) = 6.76$, $p = .009$. However, inspection of the model coefficients reveals that the magnitude of this effect was quite small, $\beta = -0.07$, 95% CI [-0.12, -0.02]. Therefore, we take the linear effect alone to be a reasonable approximation of the effect of an information gap on curiosity in all subsequent analyses.

measure predicted ESC and explanation search in isolation from all other measures (analyses in supplementary material at https://osf.io/gyuqa/). As expected, 10 of the 13 measures were significant predictors of ESC in isolation (all but fact-and-foil, map mismatch, and negative valence), and nine of the 13 potential triggers were significant predictors of explanation search (all but breadth, fact-and-foil, map mismatch, and negative valence).

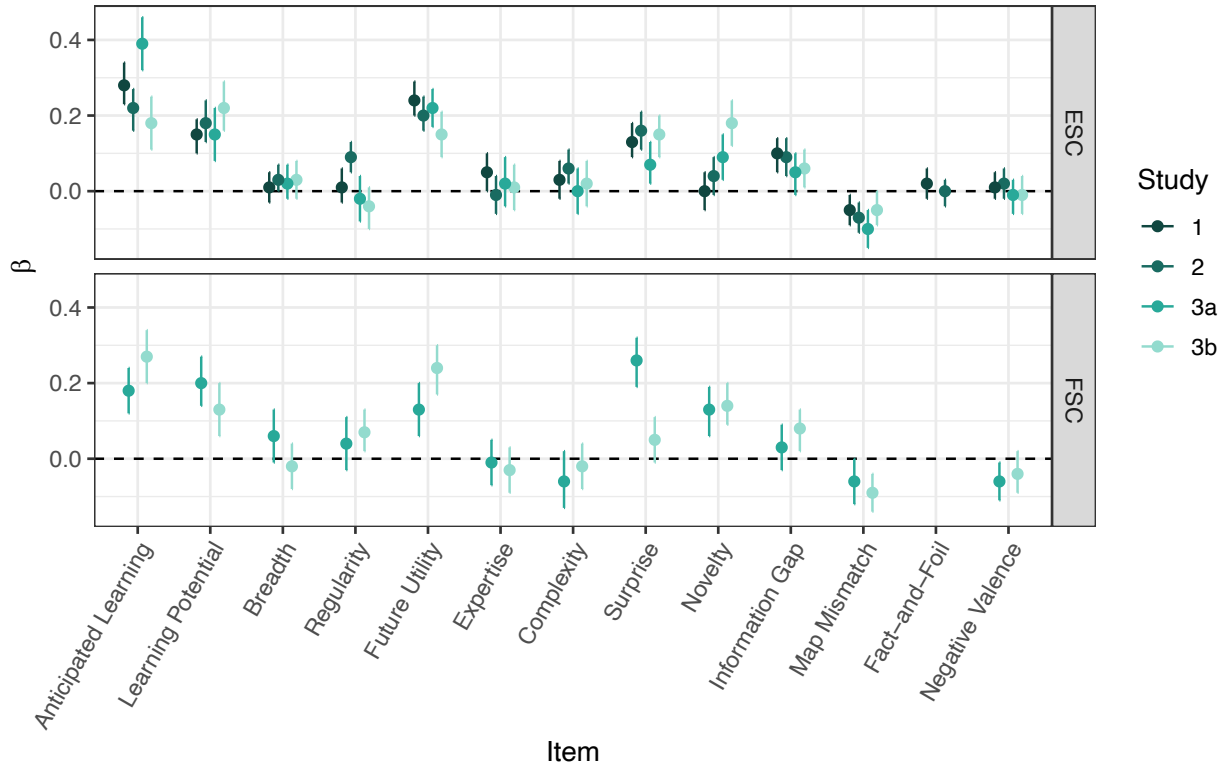


Figure 1 Regression Results, Studies 1-3
Standardized regression coefficients (with 95% CIs) predicting ESC/FSC for each potential trigger.

Forward-looking and backward-looking triggers

Finally, we tested whether forward-looking triggers of ESC significantly contributed to the prediction of ESC and explanation search, over and above the predictive power afforded by backward-looking triggers. In other words, do forward-looking expectations about the functional consequences of engaging in explanation search capture variance in ESC or explanation search

that was not already captured by backward-looking considerations (such as surprise and information gap)? We fit two multi-level models to the ESC rating data. The first model contained all backward-looking triggers as fixed effects: surprise, novelty, information gap, map mismatch, and fact-and-foil. The second model also contained these, plus fixed effects for simplicity/complexity, breadth, anticipated learning, learning potential, regularity, future utility, and expertise (negative valence was excluded from this analysis). Both models included random intercepts for participant and for item. The two models were compared using a likelihood ratio test. This analysis suggested strongly that the forward-looking triggers of explanation search significantly contributed to the prediction of ESC above and beyond backward-looking triggers, $\chi^2(7) = 578.19$, $p < .001$. This analysis was repeated for the answer reveal data (removing random intercepts for participant, as all participants selected exactly three answers to reveal), and similarly, the forward-looking triggers significantly increased the model fit, $\chi^2(7) = 41.89$, $p < .001$. According to Nakagawa and Schielzeth's (2013) measure of R^2 for generalized linear mixed effects models, adding the forward-looking potential triggers increased the marginal R^2 (measuring only the percentage of variance explained by the fixed effects of the models) from 21% to 51% for the prediction of ESC and from 5% to 9% for the prediction of answer reveal. Thus, while backward-looking triggers may serve as proximal cues to the downstream consequences of explanation seeking, our forward-looking triggers play a role in predicting ESC and explanation-seeking that is not well captured by these backward-looking triggers.

2.2.3 Discussion

These results suggest that our measure of ESC predicts actual explanation search, and that many of the potential triggers of ESC identified in the Introduction predict both ESC and explanation search. Most strikingly, the strongest three predictors of ESC—anticipated learning,

learning potential, and future utility—align with forward-looking goals of learning about the explanandum and learning for export. Other predictors of ESC—information gap, surprise, and map mismatch—are backward-looking, and are more familiar from previous research on curiosity and need for explanation. In Studies 2-3, we consider the relationships among triggers and their associations with ESC more systematically, but for now we note that the best predictors of ESC are "forward-looking" as opposed to "backward-looking" in that they concern the potential benefits of explaining the explanandum, rather than the relationship between the explanandum and current beliefs. Furthermore, forward-looking triggers predict ESC and explanation search above and beyond backward-looking triggers, suggesting that there may be value to considering expectations about the future when seeking explanations that goes beyond attention to novelty, surprise, and other backward-looking considerations.

While some potential triggers failed to significantly predict ESC or explanation search, even in isolation (see supplementary material at https://osf.io/gyuqa/), they may not be irrelevant to ESC in all contexts. In particular, the weak association between negative valence and ESC is plausibly a consequence of the nature of our stimuli: while research in attribution theory tends to focus on personally-relevant events, the questions used in the present research were not particularly personally relevant, nor did they seek to explain success or failure at achieving a given goal. We also failed to find a positive association between simplicity and ESC; instead, ESC was higher for questions with answers that were expected to be more *complex*. Even if some forms of explanatory simplicity are valued (T. Blanchard, Lombrozo, et al., 2018; Bonawitz & Lombrozo, 2012; Lombrozo, 2007; Pacer & Lombrozo, 2017; Read & Marcus-Newhall, 1993), this could be because complex explanations can make the explanandum appear more "necessary" (Zemla et al., 2017) or probable (Johnson et al., 2019), or for the reasons that

led us to predict an independent effect of complexity: a complex explanation might require greater reliance on experts, and reveal more about the explanandum.

Somewhat surprisingly, the fact-and-foil and map mismatch accounts (Grimm, 2008; W. Wong & Yudell, 2015), the two most fully developed accounts of "need for explanation" to date, faired quite poorly in predicting ESC. Fact-and-foil was not a significant predictor of ESC, and map mismatch was a *negative* predictor of ESC in the simultaneous regression model, indicating that a higher mismatch between one's map of the world and a current observation predicted lower levels of curiosity. Again, however, we do not take our results as evidence that these considerations are irrelevant to ESC. Here, several caveats are in order: First, these theories of explanation search were not necessarily intended to be descriptive psychological models, and thus people may in fact *not* use these considerations as cues to ESC, even if (perhaps) they should. Second, these accounts are plausibly directed at differentiating explanation-seeking questions that arise from those that do not even arise, while in this study, we focused on variation in ESC among questions that had in fact been asked. Put differently, our stimuli were restricted to "why" questions that already met some threshold for "need for explanation," and it could be that the fact-and-foil and map mismatch accounts identify factors that differentiate what falls under versus over this threshold, rather than accounting for variation in strength of ESC well above this threshold. In Study 2, we address this concern.

2.3 Study 2

In Study 2, we replicate Study 1 with a broader set of "why" questions drawn from two sources. The *Explain Like I'm Five* question set likely reflects a very limited range of all possible questions, as these questions were posted on a website in search of answers. To truly discover the sources of ESC, we would ideally probe such "asked" questions alongside "unasked"

questions—that is, questions that haven't been posted on a webpage specifically designed to answer questions, and that reflect questions that people *would not* normally be compelled to ask. This is precisely what we do in Study 2 by constructing a set of "unasked" questions in addition to a new set of *Explain Like I'm Five* questions. This allows us to test the 13 potential triggers more fairly, using a broader set of potential "why" questions. The addition of this question set also increases the probability that any detected triggers of ESC distinguish ESC from its absence (rather than distinguishing high levels of ESC from moderate levels of ESC), and it addresses the potential worry that questions posed to others (on an on-line forum) could inflate the importance of expertise as a predictor of ESC.

Additionally, as suggested by the modest correlations between several of the 13 potential triggers of explanation search, it is likely that these judgments reflect some underlying structure. If we assume the potential trigger ratings are responded to on the basis of several latent variables, we can use factor analysis to extract these latent variables. These latent variables, in turn, might provide new insights about both the triggers of ESC and how they map onto the functional consequences of explanation search. In Study 2, we explore this potential structure, finding three latent variables that predict ESC.

2.3.1 Method

Participants

Participants were 187 adults (80 males, 106 females, and 1 other) ranging from 19 to 70 years of age (M_{age} = 36, SD = 12). Fourteen additional participants were excluded for failing an attention check (explained below). The sample size was planned to match the sample size of Study 1.

Materials

Forty new "why" questions (not used in Study 1) were sampled from the Reddit *Explain Like I'm Five* webpage (www.reddit.com/r/explainlikeimfive). Questions were edited lightly for grammar and readability. (See https://osf.io/gyuqa/ for the complete set of the stimuli.)

Since ELI5 is comprised only of questions that users actually asked (and thus these questions likely already surpass some threshold of curiosity beyond which a question is posed), we also sought "unasked questions" from ConceptNet, an open source semantic network (http://conceptnet.io; Speer et al., 2017). This semantic network contains words and phrases, which are linked to other words and phrases by the relations they hold to each other. For example, the words "lizard" and "garden" are linked by the relation "location of." Subsequently, a "why" question can be constructed from these two words and their relation: "Why are lizards found in gardens?" To constrain our construction of questions, we focused solely on the domain of non-human animals. We selected ten animals (lizard, dog, beaver, camel, horse, bear, dolphin, duck, deer, and hamster) and four relation types (is a type of, is capable of, location of, and types of), then constructed 40 why questions based on the entries under these animals and relations in ConceptNet. Sample questions include: "Why are dogs a type of canine?"; "Why are dromedaries a type of camel?"; and "Why are bears capable of fishing for salmon?"

Procedure

Participants were assigned to eight questions from either the ELI5 ($N = 96$) or ConceptNet ($N = 91$) question set.[8] For the ConceptNet question set, participants were randomly assigned two questions from each relation type.

[8] Participants for each stimulus set were in fact recruited separately, but the studies are reported here as one study, as methods and analyses were identical other than the stimulus set used.

First, as in Study 1, participants rated each item on ESC and the 13 other potential trigger measures. After completing these ratings for each of the eight questions, participants completed an attention check, in which they were asked to identify four of the questions they had read previously from four distractor questions. The distractor questions were four additional questions from ConceptNet or from ELI5 that were not included in the full question set. The attention check was scored for accuracy out of eight (one point for each question correctly selected and one point for each distractor question not selected), and participants who scored lower than six were excluded.

2.3.2 Results

Triggers of ESC

First, we tested the association between each potential trigger and ESC, accounting for the other potential triggers by holding them fixed at their mean values. We thus repeated the analysis used in Study 1, finding that 8 of the 13 potential triggers made a significant contribution to the model predicting ESC: anticipated learning, learning potential, regularity, future utility, complexity, surprise, information gap,[9] and map mismatch (see Figure 1). These results largely replicate Study 1; as in Study 1, many forward-looking potential triggers (anticipated learning, learning potential, regularity, future utility, and complexity) were significant predictors of ESC even holding backward-looking triggers fixed. We also replicate the results of the individual regression analyses in the supplementary material at https://osf.io/gyuqa/.

[9] Again, we tested the quadratic effect of information gap on curiosity, using the same method as in Study 1. This revealed a significant quadratic effect, $\chi 2(1) = 13.65$, $p < .001$. However, the magnitude of this effect was again quite small, $\beta = -0.10$, 95% CI [-0.16, -0.05]. We thus take the linear effect alone to be a reasonable approximation of the effect of an information gap on curiosity in all subsequent analyses.

Stimulus set

We were also interested in whether the stimulus set (ELI5 or ConceptNet) moderated the relationship between any potential trigger and ESC. We expanded upon the model reported above by adding regression coefficients for stimulus set and for the interactions between stimulus set and each potential trigger measure. The significance of each interaction was tested by comparing (using likelihood ratio tests) the full model to a reduced model excluding a single interaction term. The interactions between stimulus set and complexity, β =-0.13, 95% CI [-0.22, -0.03], $\chi^2(1) = 7.06$, $p = .008$, and stimulus set and information gap, β =0.10, 95% CI [0.004, 0.19], $\chi^2(1) = 4.18$, $p = .04$, made a significant contribution to the full model. Complexity was a stronger predictor of ESC in the ConceptNet question set than the ELI5 question set, while information gap was a stronger predictor of ESC in the ELI5 question set than the ConceptNet question set. In general, this analysis suggests only minor differences between "asked" (ELI5) and "unasked" (ConceptNet) questions in the extent to which the potential triggers predict ESC.

Factor structure of predictors

Having replicated the basic pattern of results from Study 1, we next performed a factor analysis on the 13 potential triggers of ESC. Raw ratings on the 13 items were group mean centered on the basis of individual questions (e.g., the mean "learning potential" rating for the question "Why are dogs a type of canine?" was subtracted from all "learning potential" ratings in response to that question). This technique reduces the ICC, which in turn reduces the influence of clustering on parameter estimates (see Huang & Cornell, 2016). Initial eigenvalues for the first three factors were over one, and explained 29%, 14%, and 10% of the variance in ratings, respectively. The next three factors explained 7%, 7%, and 6% of the variance and had

eigenvalues less than one. In accordance with the results of parallel analysis, we retained a three-factor solution.

An initial three-factor solution, using varimax rotation, was inspected. Two potential triggers—fact-and-foil and negative valence—were eliminated because they failed to load onto any factor with a factor loading above 0.4 or any several factors with loadings above 0.3. The final three-factor solution with varimax rotation, excluding these two items, explained 47% of the total variance in ratings. The three factors corresponded roughly to learning-related considerations (anticipated learning, learning potential, information gap, expertise, and complexity), export-related considerations (learning potential, regularity, future utility, and breadth), and backward-looking considerations (surprise and novelty). Though map mismatch loaded onto the backward-looking factor in the initial factor analysis, its factor loading did not exceed 0.4 in the final solution (once fact-and-foil and negative valence were removed). The factor loading matrix for this analysis is presented in Figure 2.

Factor scores and ESC

Next, we extracted factor scores from the final three-factor solution (using Thomson's method; Thomson, 1951), for use in predicting ESC. The z-scored factor scores were simultaneously entered as fixed effects in a model predicting ESC, with random intercepts for participant and item.[10] The contribution of each factor to the prediction of ESC was calculated by comparing this model using likelihood ratio tests to three reduced models, each excluding one of the three

[10] Lastovicka and Thamodaran (1991), among others, have suggested that factor scores may not be appropriate for use in regression analysis because these scores are estimates rather than measurements, resulting in biased regression coefficient estimates. We report this method in the main text because it is familiar and easy to interpret, but we also present the results from an alternative method, using Structural Equation Modeling (SEM), in the supplementary material at https://osf.io/gyuqa/. These additional analyses are provided for all analyses for which factor scores are used across all studies and produce largely consistent results with the factor score regression analysis.

factors. This analysis revealed that the factor scores for all three factors made a significant contribution to the prediction of ESC (see Figure 2), learning: β =0.38, 95% CI [0.35, 0.42], $\chi^2(1) = 386.31$, $p < .001$; export: β =0.37, 95% CI [0.33, 0.41], $\chi^2(1) = 276.43$, $p < .001$; backward-looking: β =0.20, 95% CI [0.16, 0.23], $\chi^2(1) = 124.95$, $p < .001$.

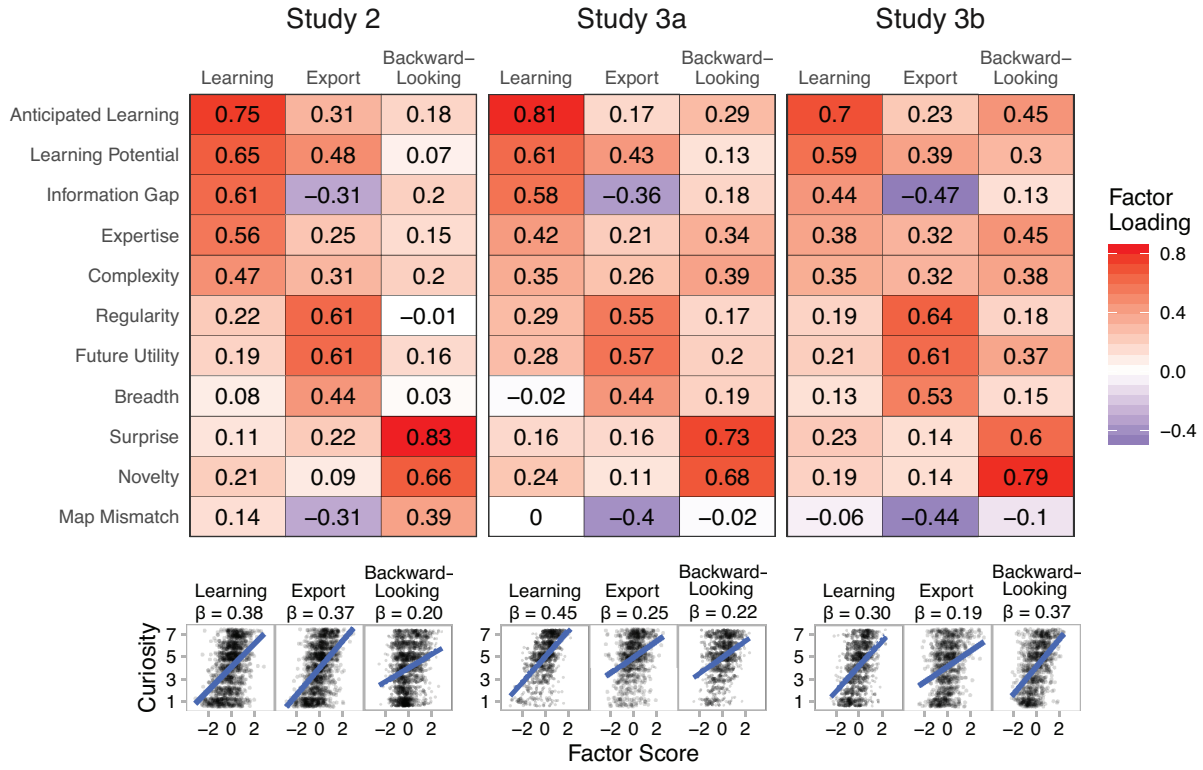


Figure 2 Factor Analysis for Explanation-Seeking Questions, Studies 2-3
Results of exploratory factor analysis on ratings in response to explanation-seeking questions. For Studies 3a and 3b factors are rearranged to facilitate comparison with Study 2 (Study 3a original order: learning, backward-looking, export; Study 3b original order: export, backward-looking, learning). Below, jittered scatterplots (with a point for each item rated by each participant) showing the relationship between extracted factor scores for each factor/study and ESC ratings, and the standardized regression coefficient relating these variables controlling for the nested structure of the data.

For each factor, we also fit a separate model to predict ESC ratings, with the factor score as a fixed effect and random intercepts for participant and item. Nakagawa and Schielzeth's (2013) measure of R^2 for generalized linear mixed effects models was used to compare the

variance in ESC explained by each model. The marginal R^2 of the learning model was 0.20, the marginal R^2 of the export model was 0.18, and the marginal R^2 of the backward-looking model was 0.06. These findings suggest that the learning model explained the most variance in ESC, followed by the export model then the backward-looking model.

2.3.3 Discussion

Study 2 offers several insights about explanation-seeking curiosity and the mapping between the factors that trigger it and the functional consequences of explanation search. First, we replicated the results of Study 1 on a broader set of "why" questions that included both "asked" (ELI5) and "unasked" (ConceptNet) questions. Additionally, ratings of the 13 potential triggers of ESC corresponded to a three-factor structure, suggesting that the triggers of explanation search correspond to three latent variables, tracking forward-looking potential for learning, forward-looking potential for export, and backward-looking considerations. Notably, these first two factors correspond fairly well with the functional consequences of explanation search reviewed in the Introduction (with the triggers supporting "learning about the explanandum" and "learning from others' domain-based expertise" collapsed into a single "learning" factor). Information gap, which we initially classified as a backward-looking trigger, in fact patterned with forward-looking triggers concerning expectations about learning. While this item ("How confident are you that you know the answer to this question?") does concern prior knowledge rather than expectations about the future, and is thus backward-looking in nature, it may be the case that this backward-looking measure is in fact a highly reliable cue to anticipated learning. For example, some theories of curiosity suggest that an information gap leads to a feeling of deprivation, which results in curiosity and motivates subsequent learning (e.g., Jirout & Klahr, 2012). Interestingly, information gap may in fact be a *better* cue to

expectations about learning than are surprise and novelty, as these measures load onto a separate factor and thus are less closely associated with anticipated learning.

Our backward-looking factor, which included measures from classic theories of curiosity (e.g., Berlyne, 1950, 1966; Berlyne & Frommer, 1966), captured the smallest portion of variance in people's judgments of ESC. Relatedly, the map mismatch account (W. Wong & Yudell, 2015) and the fact-and-foil account (Grimm, 2008), the principal theoretical accounts of need for explanation to date, were again fairly poor predictors of actual ESC judgments (replicating Study 1). Instead, ESC was best captured by triggers corresponding to learning and export: the extent to which an explanation might teach one something new, and the extent to which that information might be generalizable to other contexts in the future. This is surprising because it has been assumed in previous research on curiosity (whether implicitly or explicitly) that backward-looking considerations (e.g., surprise) are important drivers of curiosity and exploratory behavior *because* they are cues to important functional consequences, such as learning. As in Study 1, we show that forward-looking considerations (in the form of *expectations* about downstream functional consequences) may in fact be important correlates of curiosity above and beyond these backward-looking considerations.

Nonetheless, these results leave several questions unanswered. First, how well does the three-factor structure we found in this study extend to different sets of explanation-seeking questions? We test this in Studies 3a and 3b, conducting a close replication with 40 new questions from *Explain Like I'm Five* (Study 3a) and a far replication with 40 questions drawn from a new source, notably one for which the explananda are themselves novel for participants (Study 3b). Second, how do the potential triggers of explanation-seeking curiosity compare to the potential triggers of fact-seeking curiosity?

2.4 Study 3

In Study 3, we attempt to replicate the three-factor structure describing the potential triggers of ESC in Study 2. Additionally, we investigate whether the tested triggers of ESC are differentially relevant for the prediction of fact-seeking curiosity (FSC). We do so using two sets of questions as experimental stimuli. In Study 3a, we use a new set of ELI5 questions, and contrast these questions with fact-seeking trivia questions used in prior work (Dubey & Griffiths, 2020; Kang et al., 2009). In Study 3b, we use explanation-seeking and fact-seeking questions posted as comments in response to captioned images in an on-line forum. This latter set of stimuli has an additional notable feature: while the ELI5 questions and trivia questions largely ask about familiar, everyday objects or events (e.g., "Why do ice cubes crackle when liquid is poured on them?"), and the ConceptNet items probed familiar facts from common knowledge (e.g., "Why are dogs a type of canine?"), the questions used in Study 3b are in response to novel image stimuli. Thus, these questions provide a fairer test of the role of "backward-looking" considerations (e.g., map mismatch, surprise, and novelty) in predicting ESC, as novel stimuli have a much greater potential to invoke surprise, novelty, or a map mismatch than familiar stimuli.

Why might we expect differences in the predictors of ESC and FSC? Lowenstein (1994) drew upon research on explanation search to build his information gap theory of curiosity, and Schwitzgebel (1999) suggests that explanation-seeking curiosity could be either a different *type* of curiosity, or simply the same drive state directed towards a different type of information. However, some lines of research point towards differences in explanation-seeking and fact-seeking processes. First, the "explanation for export" account emphasizes the role of explanation in providing information that can be generalized to future contexts (Lombrozo & Carey, 2006).

While this account does not address the contrast between explanations and facts, it might suggest that export should be *more* characteristic of explanations than of other types of information (including facts), and thus we might expect export-relevant potential triggers—such as future utility, breadth, and regularity—to be stronger predictors of ESC than of FSC.

Additionally, there is evidence for a special link between explanation and deference to experts, which may not exist to the same extent for simple facts. Wilkenfeld, Plunkett, and Lombrozo (2016) presented participants with a series of hypothetical agents who possessed different depths of knowledge about a given causal connection, and asked the participants to evaluate the extent to which each agent knew why or understood why a relevant effect occurred. Relatively shallow levels of knowledge were sufficient for attributions of knowledge-why, but greater depths of knowledge were required for attributions of understanding-why. Notably, judgments of understanding-why (vs. knowledge-why) were also more strongly associated with a measure of deference to expertise: whether the agent would be a good person to consult for a related question in the same domain. This suggests a special link between understanding why something is the case (i.e., having an explanation) and expertise, which does not extend to mere knowledge. Therefore, in the context of this study, we might expect expertise-relevant potential triggers to be stronger predictors of ESC than of FSC.

2.4.1 Method

Participants

Participants in Study 3a were 234 adults (113 males, 119 females, and 2 other) ranging from 20 to 72 years of age (M_{age} = 36, SD = 10). Participants in Study 3b were 251 adults (109 males, 141 females, and 1 other) ranging from 19 to 73 years of age (M_{age} = 36, SD = 10). Across the two studies, fifty additional participants were excluded for failing attention checks (explained

below), and six additional participants were excluded because their responses came from duplicate IP addresses.

The sample sizes of these studies were planned to match Study 2 after exclusions. An initial version of Study 3b contained a significant proportion of low-quality data, likely due to automated responding or "bots" on Amazon Mechanical Turk (Dennis et al., 2018). To combat this, we added several attention checks and restricted participation to those in the United States with a minimum approval rating of 99% on at least 1000 previous Mechanical Turk tasks. We also included a Captcha Verification in these studies. We preregistered the studies and analyses,[11] but some of the analyses reported below go beyond our preregistration. All preregistered analyses are presented in the supplementary material at https://osf.io/gyuqa/, and all departures and additions are explained and justified.

Materials

For Study 3a, forty new "why" questions (not used in Studies 1 or 2) were sampled from the Reddit *Explain Like I'm Five* webpage (www.reddit.com/r/explainlikeimfive). Questions were edited lightly for grammar and readability. The 40 trivia questions used by Kang et al. (2009) and Dubey and Griffiths (2020) were used as fact-seeking questions (FSQs).

For Study 3b, we used Google's BigQuery to extract comments on the r/pics subreddit (www.reddit.com/r/pics/) posted between January 2017 and June 2018 that began with the words "who," "what," "when," where," "why," or "how." All comments were posted directly in response to the original posts (which were images with captions), rather than in response to another user's comment. From this set of comments, we extracted posts that had at least one why-question comment and at least one who/what/when/where/how-question comment, then

[11] The preregistration for Study 3a can be found at https://aspredicted.org/bx6bh.pdf and the preregistration for Study 3b can be found at https://aspredicted.org/8je27.pdf.

hand selected 40 posts with an explanation-seeking why-question and a fact-seeking who/what/when/where/how-question,[12] avoiding rhetorical questions, jokes, and inappropriate content. This produced a set of 40 explanation-seeking questions and 40 fact-seeking questions, posted in response to the same 40 images and captions. As an example, one image was a picture of two dogs on a snowy hill at sunset, with the caption "My mom took this at our cabin yesterday." The explanation-seeking question in response to this post asked, "Why is only one of your dogs on a leash and not the other?" and the fact-seeking question asked, "Whereabouts was this?" (See https://osf.io/gyuqa/ for a complete set of stimuli.)

Procedure

Participants were randomly assigned to eight questions from either the ESQ set (N_{3a} = 114; N_{3b} = 128) or the FSQ set (N_{3a} = 120; N_{3b} = 123). For each question, participants responded to the prompt "How curious are you about the answer to this question?" on a seven-point scale, with the endpoints marked "Not at all curious" and "Very curious." Participants also rated each question on 12 potential trigger measures (fact-and-foil was excluded due to our inability to produce a version of this item that applied to FSQs), each on a seven-point scale. Several of these items were reworded from the items in Studies 1 and 2 so that they would apply to both explanation-seeking and fact-seeking questions (see Table 1). These 12 items and the curiosity measure were presented in a random order.

After completing these ratings for each of the eight questions, participants completed three attention checks. First, they were asked to type a given word contained in a JPEG image into a text box. Next, they were asked to identify four of the questions (Study 3a) or images

[12] "How" questions are often explanation-seeking in nature (e.g., "How does a bicycle work?"). For the purpose of this study, we exclusively selected *fact-seeking* "how" questions (e.g., "How tall is the Eiffel Tower?").

(Study 3b) they had seen previously from four distractor questions/images. The distractor questions were four additional questions from ELI5 (for ESQs) or four additional fact-seeking trivia questions (for FSQs), and the distractor images were four additional images from other Reddit posts. This attention check was scored as in Study 2. Finally, participants solved an addition problem, for which the arguments were two random numbers between 1 and 20. Participants who did not successfully complete all three attention checks were excluded.

2.4.2 Results

Triggers of ESC and FSC

First, we tested whether the 12 potential triggers were differentially predictive of ESC and FSC. To do so, we fit a mixed-effects regression model predicting curiosity, with fixed effects for each potential trigger, question type (explanation-seeking vs. fact-seeking), and the interaction between each trigger and question type; and with random intercepts for participant and item. This full model was compared to a set of reduced models that each excluded one interaction term, using likelihood ratio tests. In Study 3a, anticipated learning was a stronger predictor of ESC than of FSC, β = 0.21, 95% CI [0.11, 0.31], $\chi^2(1) = 17.13$, $p < .001$, while surprise was a stronger predictor of FSC than of ESC, β = -0.18, 95% CI [-0.27, -0.10], $\chi^2(1) = 17.82$, $p < .001$. In Study 3b, surprise, β = 0.09, 95% CI [0.01, 0.17], $\chi^2(1) = 5.33$, $p = .02$, and learning potential, β = 0.10, 95% CI [0.002, 0.19], $\chi^2(1) = 3.97$, $p = .046$, were stronger predictors of ESC than of FSC. Regularity, β = -0.12, 95% CI [-0.20, -0.04], $\chi^2(1) = 8.68$, $p = .003$, and future utility, β = -0.09, 95% CI [-0.17, -0.002], $\chi^2(1) = 4.04$, $p = .04$, were stronger predictors of FSC than of ESC. Notably, these results are inconsistent across studies and do not align with our predictions; although ESC and FSC may diverge, the measures that predict one over the other may vary across settings.

Next, we fit a model predicting curiosity across conditions (collapsing across ESC and FSC), using the same analysis as in Studies 1 and 2. In Study 3a, eight of the 12 potential trigger measures made a unique contribution to the prediction of curiosity, holding all other measures fixed at their mean values: anticipated learning, β = 0.26, 95% CI [0.21, 0.30], learning potential, β = 0.19, 95% CI [0.15, 0.24], future utility, β = 0.19, 95% CI [0.14, 0.23], surprise, β = 0.17, 95% CI [0.13, 0.22], novelty, β = 0.10, 95% CI [0.06, 0.15], information gap, β = 0.05, 95% CI [0.01, 0.09],[13] map mismatch, β = -0.08, 95% CI [-0.12, -0.04], and negative valence, β = -0.04, 95% CI [-0.08, -0.01]. In Study 3b, all the same measures made unique significant contributions to the prediction of curiosity, with the exception of negative valence: anticipated learning, β = 0.23, 95% CI [0.18, 0.28], learning potential, β = 0.18, 95% CI [0.13, 0.23], future utility, β = 0.19, 95% CI [0.15, 0.23] surprise, β = 0.10, 95% CI [0.06, 0.14], novelty, β = 0.16, 95% CI [0.12, 0.20], information gap, β = 0.08, 95% CI [0.04, 0.11], and map mismatch, β = -0.07, 95% CI [-0.10, -0.04]. We also fit separate models for ESC and FSC to facilitate comparison with the previous studies, which are reported in Figure 1 and the supplementary material at https://osf.io/gyuqa/. These results are largely consistent with the results of Studies 1 and 2. In particular, we again see that the strongest predictors of curiosity are forward-looking (future utility, anticipated learning, and learning potential), while the backward-looking measures emphasized in prior research on curiosity (information gap, surprise, novelty) play a smaller role.

[13] As in Studies 1 and 2, we tested the quadratic effect of information gap on curiosity. This revealed a significant quadratic effect in both Study 3a, $\chi^2(1) = 35.15$, $p < .001$, and Study 3b, $\chi^2(1) = 5.74$, $p = .02$. The magnitude of this effect was again fairly small in both studies, Study 3a: β =-0.16, 95% CI [-0.22, -0.11]; Study 3b: β =-0.06, 95% CI [-0.12, -0.01]. We include only the linear effect of information gap in all subsequent analyses.

Factor structure and ESC

We next sought to verify the factor structure found in Study 2 in response to explanation-seeking questions. We did so using two approaches. First, we used confirmatory factor analysis to test whether the factor solution from Study 2 provided a good fit to the data from Study 3a and Study 3b. This analysis is presented in the supplementary material at https://osf.io/gyuqa/. The confirmatory factor analyses did not meet standard thresholds of good fit, so we then turned to exploratory factor analysis to investigate the factor structure of the potential trigger measures in these two new datasets (as preregistered, and as recommended by Flora & Flake, 2017).

As in Study 2, raw ratings on the measures were group-mean centered within questions, to eliminate the influence of clustered data on parameter estimates. For Study 3a, initial eigenvalues for the first three factors were over one, and explained 33%, 13%, and 9% of the variance in ratings, respectively. Consistent with parallel analysis, we retained a three-factor solution. For Study 3b, initial eigenvalues for the first two factors were over one and explained 37% and 13% of the variance in ratings, respectively. The third factor had eigenvalue 0.97 and explained 8% of the variance. Parallel analysis recommended a two-factor solution. Because three-factor solutions were used for all prior analyses, we inspected both two- and three-factor solutions; the three-factor solution produced better fit to the data than the two-factor solution; $TLI_{3\text{-factor}}$: 0.94, $RMSEA_{3\text{-factor}}$: 0.06, 95% CI [0.05, 0.07]; $TLI_{2\text{-factor}}$: 0.90, $RMSEA_{2\text{-factor}}$: 0.08, 95% CI [0.07, 0.09]. Additionally, the three-factor solution was more interpretable and would facilitate comparison to Studies 2 and 3a. Thus, we report the final two-factor solution in the supplementary material at https://osf.io/gyuqa/, but we retain the three-factor solution in the text that follows.

An initial three-factor solution, using varimax rotation, was inspected for each study. In both datasets, negative valence was eliminated because it failed to load onto any factor with a factor loading above 0.4 or on several factors with loadings above 0.3. The final three-factor solutions with varimax rotation, excluding this item, explained 44% of the total variance in ratings for Study 3a and 48% of the total variance in ratings for Study 3b. Notably, the structure of the latent variables was quite similar to that in Study 2 (see Figure 2).

To quantitatively compare the results of these factor analyses to the Study 2 factor analysis, we estimated the correlation between the factor loadings for each potential trigger measure onto each factor in Study 2 and the corresponding factor loadings in Study 3a and Study 3b. The factor solutions were highly correlated between Study 2 and Study 3a, $r = 0.90$, $t(31) = 11.79$, $p < .001$, between Study 2 and Study 3b, $r = 0.82$, $t(31) = 8.03$, $p < .001$, and between Study 3a and Study 3b, $r = 0.94$, $t(31) = 16.03$, $p < .001$. These results suggest that there is significant overlap between the factor solutions, despite using different sets of explanation-seeking questions across the three studies and changing the items used to measure some of the potential triggers between Studies 2 and 3.

Factor scores were extracted from these factor models using Thomson's method (Thomson, 1951) and were used to predict ESC as in Study 2. For Study 3a, the factor scores for all three factors made a significant contribution to the prediction of ESC, learning: $\beta = 0.45$, 95% CI [0.41, 0.50], $\chi^2(1) = 341.18$, $p < .001$; export: $\beta = 0.25$, 95% CI [0.20, 0.29], $\chi^2(1) = 111.09$, $p < .001$; backward-looking: $\beta = 0.22$, 95% CI [0.17, 0.26], $\chi^2(1) = 91.67$, $p < .001$. Additionally, when a separate model was fit to predict ESC for each individual factor, the marginal R^2 of the learning model was 0.31, the marginal R^2 of the export model was 0.08, and the marginal R^2 of the backward-looking model was 0.11.

For Study 3b, the factor scores for all three factors made a significant contribution to the prediction of ESC, learning: $\beta = 0.30$, 95% CI [0.26, 0.34], $\chi^2(1) = 201.52$, $p < .001$; export: $\beta = 0.19$, 95% CI [0.15, 0.23], $\chi^2(1) = 77.89$, $p < .001$; backward-looking: $\beta = 0.37$, 95% CI [0.33, 0.41], $\chi^2(1) = 286.61$, $p < .001$. Additionally, when a separate model was fit to predict ESC for each individual factor, the marginal R^2 of the learning model was 0.19, the marginal R^2 of the export model was 0.06, and the marginal R^2 of the backward-looking model was 0.23. As predicted, backward-looking measures played a larger role in predicting ESC when questions were posed in response to a novel image stimulus. Nonetheless, expectations about learning and export still predicted ESC controlling for backward-looking considerations.

Factor structure and FSC

Next, we investigated whether the potential trigger measures in response to FSQs followed a qualitatively different factor structure from that for ESQs. To do so, we conducted a second set of exploratory factor analyses, using only the subset of the data in response to FSQs. As before, raw ratings were first group-mean centered on the basis of question to control for the nested structure of the data. For Study 3a, eigenvalues for the first three factors were over one, and explained 32%, 13%, and 9% of the variance in ratings, respectively. As parallel analysis suggested a two-factor solution, we compared a two- and three-factor solution. The three-factor solution produced better fit to the data than the two-factor solution; $TLI_{3\text{-factor}}$: 0.89, $RMSEA_{3\text{-factor}}$: 0.07, 95% CI [0.06, 0.08]; $TLI_{2\text{-factor}}$: 0.83, $RMSEA_{2\text{-factor}}$: 0.09, 95% CI [0.08, 0.10]. Therefore, we retained a three-factor solution.

An initial three-factor solution, using varimax rotation, was inspected. Negative valence and map mismatch failed to load onto any factor with a factor loading above 0.4 or any several factors with loadings above 0.3, and thus were eliminated. Excluding these items, the final three-

factor solution with varimax rotation explained 47% of the total variance in ratings. The factors resulting from this analysis look notably different from the factors that emerged in response to explanation-seeking questions (see Figure 3). The lack of consistency between the results of these two factor analyses precludes a more quantitative comparison.

For Study 3b, we initially fit a confirmatory factor analysis model to the FSC data (reported in the supplementary material at https://osf.io/gyuqa/), using the factor structure discovered in the previous analysis for Study 3a. However, this model did not achieve good fit by any standard metrics, so we conducted a second exploratory factor analysis. Initial eigenvalues for the first three factors were over one, and explained 38%, 11%, and 10% of the variance in ratings, respectively. We retained a three-factor solution, as recommended by parallel analysis. As in Study 3a, negative valence and map mismatch failed to load onto any factor with a factor loading above 0.4 or any several factors with loadings above 0.3, and thus were eliminated. Excluding these items, the final three-factor solution with varimax rotation explained 52% of the total variance in ratings. As can be seen in Figure 3, the factors resulting from this analysis look qualitatively different from the factors that emerge in response to explanation-seeking questions across all studies. However, they are significantly though modestly correlated with the factor loadings found in Study 3a for FSQs, $r = 0.48$, $t(28) = 2.87$, $p = .008$.

As preregistered, we used the resulting factor scores to predict FSC. Due to the limited interpretability of these factors (and consequently, the limited interpretability of these regression analyses) we report these analyses in the supplementary material at https://osf.io/gyuqa/.

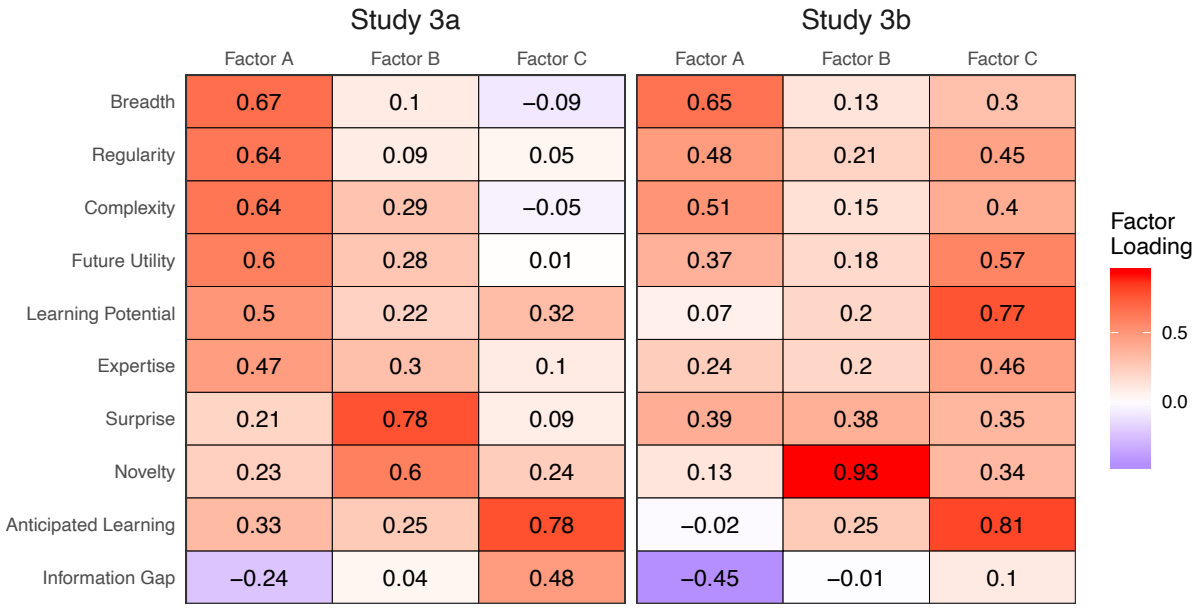


Figure 3 Factor Analysis for Fact-Seeking Questions, Study 3

Results of exploratory factor analysis on ratings in response to fact-seeking questions. Factors for Study 3b are rearranged to facilitate comparison with Study 3a (original order: factor C, factor A, factor B).

2.4.3 Discussion

In relation to Studies 1 and 2, we largely replicate our previous findings in Study 3. The relative strength of the 12 potential trigger measures in predicting curiosity was similar to the relative strength of these measures in Studies 1 and 2 (see Figure 1). Further, though confirmatory factor analysis resulted in a poorly fitting model, exploratory factor analysis revealed a three-factor structure that was qualitatively and quantitatively highly similar to the three-factor structure found in Study 2 (see Figure 2), across two different sets of explanation-seeking questions. The failure of the confirmatory factor analysis model despite the striking similarity between exploratory factor analysis solutions is likely at least in part driven by the constraint that each measure could only load onto one factor. As can be seen in Figure 2, many measures had small but potentially meaningful factor loadings onto more than one factor. By limiting each measure to load onto a single factor, we limited the amount of variance this model

could explain. Nonetheless, the exploratory factor analysis lends further support to the idea that ESC is driven by both backward-looking triggers and by forward-looking triggers that correspond to two functions—learning and export. We again found that triggers related to learning were highly predictive of ESC. Backward-looking triggers were much weaker in Study 3a, but stronger in Study 3b. This is potentially because the explanation-seeking questions in Study 3b were in response to novel image stimuli rather than familiar facts or phenomena—while the latter may be expected to elicit a fairly limited range of surprise or novelty, the former may be more sensitive to backward-looking considerations. In contrast to Study 2, however, triggers related to export were also somewhat weaker predictors of ESC across Studies 3a and 3b.

Study 3 additionally extends beyond Studies 1-2 by suggesting that ESC and FSC are elicited under somewhat different conditions. In Study 3a, anticipated learning was a stronger predictor of ESC than of FSC, while surprise at the content of a question was a stronger predictor of FSC than of ESC. In Study 3b, surprise and learning potential were stronger predictors of ESC than of FSC, while regularity and future utility were stronger predictors of FSC than of ESC. These were not the differences that we predicted (and, in the case of surprise, were inconsistent across studies), but they nonetheless suggest that future research might benefit from treating curiosity targeting explanations and curiosity targeting facts as potentially distinct. Supporting this suggestion, the latent variable structure of the tested triggers varied across the explanation-seeking and fact-seeking question sets. These differences leave open several avenues for future research on the distinction (if any) between ESC and other forms of curiosity.

2.5 General Discussion

When and why do we ask why? Across three studies, we evaluated potential triggers of explanation-seeking curiosity: the state of curiosity about why something is the case. We proposed a functional approach to ESC premised on the idea that ESC should be triggered when pursuing explanations is likely to achieve core functions of explanation search: learning about the explanandum, learning for export, and learning from the domain-based expertise of others. Using this approach, we derived a set of potential triggers of ESC, including many that have not been tested in prior work on curiosity or information search. Notably, these potential triggers were forward-looking in nature—that is, concerning expectations about the downstream consequences of explanation search—in contrast to the backward-looking triggers (concerning fit with prior beliefs) that have dominated previous work on curiosity. Below, we review our empirical findings before turning to implications, open questions, and limitations.

First, we found that ratings of explanation-seeking curiosity predicted choices about which explanations to reveal (Study 1). This validates our measure of explanation-seeking curiosity, as well as the link between ESC and explanation search that our functional approach assumes. Second, we found that forward-looking potential triggers predicted ESC above and beyond backward-looking triggers (Studies 1-3). These forward-looking triggers included anticipated learning, learning potential, need for relevant expertise, anticipated complexity, whether the explanandum is anticipated to reflect a pattern or regularity, and anticipated breadth of application. Notably, these included the three triggers with the highest predictive value across all of our data sets (learning potential, anticipated learning, and future utility).

To move beyond a large, unstructured list of individual predictors, we also explored the possibility that ratings of the potential triggers of ESC reflect a smaller number of latent factors

(Studies 2-3). We identified three factors that were both highly interpretable and consistent across a diverse set of stimuli. These factors corresponded to (forward-looking) learning, (forward-looking) export, and backward-looking considerations. For fact-seeking questions, we found far less coherence and consistency (Study 3), which makes the coherence and consistency revealed for explanation-seeking questions all the more remarkable. Even for explanation-seeking questions, however, we did observe variation in the relative importance of each factor for predicting curiosity across studies, with the backward-looking factor playing the largest role when participants encountered novel explananda (Study 3b). These results simultaneously vindicate a role for backward-looking considerations from prior research (namely novelty or surprise) and suggest that forward-looking considerations (such as anticipated learning and future utility) deserve a great deal more attention.

Although our findings did support several candidate triggers of ESC drawn from past research, we found a minimal role for others. Our results do not support the role of negative valence in driving explanation seeking, as suggested in the literature on attribution (Weiner, 1985b; P. T. Wong & Weiner, 1981), though this may be explained by the nature of our experimental stimuli, which did not include particularly valenced or personally-relevant questions. Our results also challenge the descriptive adequacy of theories of need for explanation proposed in philosophy (Grimm, 2008; W. Wong & Yudell, 2015)—across all studies, participant ratings of "map mismatch" and "fact-and-foil" were fairly weak predictors of ESC and of explanation search. This likely reflects both limitations in our measurement (a single rating item per theory) and limitations in the theories as descriptive accounts of human behavior.

An important question raised by our results is how people evaluate forward-looking considerations when encountering a potential target of explanation, given that these

considerations are not reducible to the backward-looking triggers tested here. One reason backward-looking considerations may be imperfectly linked to forward-looking expectations is because the former are often inaccurate. Metacognitive calibration (i.e., how accurate a learner is in representing what they do and do not know) is often poor (e.g., Glenberg & Epstein, 1985; Lin et al., 2001). This has been shown for explanatory knowledge, in particular, with participants dramatically over-estimating their understanding of causal systems such as zippers and helicopters (Rozenblit & Keil, 2002). If people over-estimate how much they know, they are unlikely to experience the novelty, surprise, information gap, or map mismatch that would prompt curiosity, even in cases where they would benefit from acquiring further information. Notably, however, attempting to explain can improve metacognitive calibration: after Rozenblit and Keil (2002) prompted participants to generate an explanation, they produced lower (and more accurate) estimates of how well they actually understood the mechanism they were trying to explain. This suggests that better calibration (in particular, recognizing an information gap or map mismatch) may be a functional consequence of explanation search, and therefore an unreliable guide to whether explanation search is necessary in the first place. Similarly, some accounts of surprise suggest that surprise is the consequence of explanatory processes, not its cause: according to Foster and Keane (2015), surprise functions as a metacognitive signal reflecting the difficulty of explaining some observation (see also R. Maguire et al., 2011). These findings help explain why backward-looking considerations (such as information gap and surprise) may be so closely associated with the pursuit of explanations. However, we are left with the puzzle of how people form the judgments that actually trigger ESC, especially when these judgments concern expectations about the future. An important question for future

research, then, is how people arrive at estimates of future learning or export, and whether such estimates are in fact reliable guides to how much will be learned from an explanation.

Relatedly, future research must go beyond our correlational approach to test whether forward-looking and backward-looking triggers can be experimentally manipulated to bring about ESC and explanation-seeking behavior. While forward- and backward-looking considerations will often go hand in hand, the relevant computations are conceptually distinct, and potentially dissociable empirically.[14] For example, the magnitude of one's surprise can be captured by the change in one's beliefs after receiving new evidence (Itti & Baldi, 2009), while anticipated learning can be captured by the expected reduction in uncertainty after asking a particular question (Oaksford & Chater, 1994). Given that the latter depends not only on one's current beliefs but also on the particular answers one expects to receive, while the former depends only on current beliefs, these quantities may come apart under some circumstances (for example, a surprising piece of data may overturn a whole body of knowledge, while asking a single explanation-seeking question about that surprising data may only be expected to resolve a small amount of the resulting uncertainty, and only if particular answers are received). In Chapter 3, we identify one setting under which backward- and forward-looking considerations can come apart, and we test how ESC depends on each set of considerations.

Another question for future research concerns the role of instrumental considerations in driving explanation search. On many classic accounts of curiosity (e.g., Loewenstein, 1994), information search motivated by practical purposes (e.g., asking why a bicycle's brakes are

[14] For example, in an unpublished study described in Loewenstein (1994), participants reported higher levels of curiosity about the rule that tied together a longer list of states versus a shorter list of states, purportedly because they wanted to learn the rule that was expected to unify *more* information (Loewenstein et al., 1992). This manipulation seems to target forward-looking considerations with minimal influence on backward-looking considerations.

squeaky in order to fix them) does not constitute curiosity at all. The present research, however, demonstrates that people's curiosity— at least as reflected in self-report—*is* in fact associated with expectations about future utility (see also Dubey & Griffiths, 2020). Importantly, neither our functional approach nor our data necessarily imply that participants *explicitly* evaluate export or utility prior to experiencing ESC, nor that they experience ESC as being purely instrumental. That said, we find it quite plausible that ESC for practical purposes (to the extent that this can still be considered curiosity at all) will be predicted more strongly by export-related considerations regarding future utility than will ESC for purely epistemic purposes. We further explore how instrumental considerations shape curiosity in Chapter 3.

Several additional limitations of our studies are worth acknowledging. First, while we used a large and diverse set of questions drawn from naturalistic sources, our stimuli and task were nonetheless restricted and somewhat artificial. With the exception of Study 3b, the questions that participants rated were largely within the domain of science, and with the exception of the ConceptNet question set, they induced a sufficiently high level of ESC or FSC to be posed in an online forum. It remains to be tested whether the triggers of ESC and explanation-seeking behavior are consistent across different sets of explanation-seeking questions (e.g., queries posted on online forums, queries entered into search engines, queries directed towards experts), across types of explanation-seeking questions (e.g., questions seeking teleological versus mechanistic explanations, or how- versus why- explanations), for different domains of questioning (e.g., questions concerning religious phenomena versus scientific phenomena versus social phenomena), in different cultural contexts, and when participants are required to generate questions themselves, rather than evaluating presented questions. Relatedly, while our focus was on the epistemic goals of explanation search, other research emphasizes the

interpersonal or intrapersonal goals of explanation search (Malle, 2004) or the argumentative function of explanations (Mercier & Sperber, 2011). Future research might explore whether these functions suggest different potential triggers of explanation-seeking behavior.

Additionally, these studies leave open the questions of whether and to what extent ESC guides exploratory behavior more generally. For example, does ESC also predict when people seek information from experts, when people make or generate new observations, and when people attempt to generate explanations themselves? This final behavior is particularly important: some of the previous research we use to motivate the "functional" triggers of ESC are based on the functions of internal explanation search (i.e., explanation generation "in the mind"), while the behavior we use to validate our measure of ESC involves external explanation search (i.e., seeking the explanation from an outside source). It remains to be seen whether the same triggers predict "internal" versus "external" explanation search.

Broadly, this research provides a framework for thinking about the triggers of exploratory behavior, and of explanation search in particular. We find that explanation-seeking curiosity is aroused in a systematic and interpretable way. People are not only sensitive to the match or mismatch between a given stimulus and their current beliefs, but also to how they expect an explanation to affect their future selves. This future-directed orientation not only helps us understand curiosity in individuals, but also prompts new questions that can direct our future research as scientists.

3 CHAPTER 3: QUANTIFYING AND DISSOCIATING ELICITORS OF CURIOSITY[†]

3.1 Introduction

Curiosity is often viewed as a force for good: curiosity drives scientific progress, curiosity about science counteracts politically motivated reasoning (Kahan et al., 2017), and curiosity in educational settings motivates us to learn (e.g., Bernacki & Walkington, 2018; Von Stumm et al., 2011). However, the consequences of curiosity can also be quite costly: for example, individuals are willing to subject themselves to electric shocks to learn the secrets behind magic tricks (Lau et al., 2020; see also Hsee & Ruan, 2016; Wiggin et al., 2019).

Are these costs of curiosity worth it? If curiosity is to be a force for good, people must generally experience curiosity under the right circumstances, so that the potential costs of curiosity are outweighed by the potential benefits. As an analogy, imagine you are going fishing in a lake. If you use a sonar device, which detects fish below the boat, you can cast your fishing line in specific areas where fish are plentiful. As a result, you will be reasonably likely to catch a fish, even if it takes some time. In contrast, if you decide where to cast your fishing line using another technique—for example, going back to where you caught fish yesterday—success is less likely, so your time might be wasted. Curiosity's goal is not to motivate fish-finding, but rather to motivate information-finding: that is, learning. If we experience curiosity and seek information precisely when we detect the most opportunity for learning, curiosity will frequently

[†] Parts of this chapter were published in Liquin, E. G., Callaway, F., & Lombrozo, T. (2020). Quantifying curiosity: A formal approach to dissociating causes of curiosity. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 309–315). Cognitive Science Society. Components of this work were also presented in the following: Liquin, E. G., Callaway, F. & Lombrozo, T. (2020, July). Quantifying curiosity: A formal approach to dissociating causes of curiosity. Talk presented at the annual meeting of the Cognitive Science Society, virtual.

guide inquiry towards new information. However, if we frequently experience curiosity in moments when learning is less likely to occur, curiosity might come at a steep cost for little benefit. In the present research, we test whether people selectively experience curiosity in moments where learning is most likely. If so, this would comprise evidence that curiosity is optimally triggered for achieving the goal of motivating learning (what we will refer to as "epistemically optimal").

In addition, we test whether curiosity guides inquiry towards learning that will specifically be *useful*. To extend our analogy, you might want to catch only large fish, in which case you might tune your sonar device to detect larger masses at a greater depth. Curiosity might likewise shift to track learning that is useful for achieving some instrumental goal. Influential accounts of curiosity (e.g., Berlyne, 1954; Loewenstein, 1994) have defined curiosity as a desire for information with no obvious practical value, and thus have excluded the possibility that humans might experience curiosity in the service of instrumental goals. However, it is nonetheless possible that individuals experience curiosity even in such situations, and even that curiosity guides inquiry specifically towards useful information. In the present research, we test whether curiosity is "instrumentally optimal": if individuals selectively experience curiosity in moments where useful learning (not just any learning) is most likely to occur.

3.1.1 Prior work on epistemically and instrumentally optimal information search

A growing body of research has investigated the extent to which information search (which may or may not be motivated by curiosity) is epistemically or instrumentally optimal. People seek information in a way that is reasonably well-described by optimal models of expected learning (e.g., Markant & Gureckis, 2014b; Nelson et al., 2010; Oaksford & Chater, 1994; Steyvers et al., 2003; for a review, see Coenen et al., 2019). For example, adults can

identify the question that would best diagnose possible configurations of hidden objects (Rothe et al., 2018). However, the evidence that information search is sensitive to the instrumental utility of information is mixed (Abir et al., 2020; Bramley et al., 2015; Dubey & Griffiths, 2020; Markant & Gureckis, 2012; Meder & Nelson, 2012; Rothe et al., 2018). For example, Dubey and Griffiths (2020) found that individuals chose to endure a delay to reveal information that they expected to be quizzed on later, but Bramley et al. (2015) found that individuals designed interventions that approximated epistemic, but not instrumental, optimality. Together, this research suggests that exploration is aligned with the goal of learning, and perhaps the goal of useful learning.

This research leaves open several questions. First, it is unclear whether curiosity is guided by epistemic and instrumental goals. Behavioral measures like information search might index curiosity at least in part, but they may be contaminated by other considerations—for example, curiosity, expected learning, and the expected instrumental value of information might be *separate* inputs into decisions to explore. Thus, it is an open question whether curiosity is epistemically or instrumentally optimal. Second, while exploration appears to be well-described by optimal models of learning, several non-optimal models make similar predictions (Klayman & Ha, 1987; Markant et al., 2015; Markant & Gureckis, 2014a; Navarro & Perfors, 2011). In fact, research on curiosity has tended to focus on surprise (e.g., Berlyne, 1966; Gruber & Ranganath, 2019; Vogl et al., 2020) and uncertainty (e.g., Berlyne, 1962; Kang et al., 2009; Kobayashi et al., 2019) as triggers of curiosity, both of which could guide curiosity in a way that resembles optimal expected learning. When one is surprised by or uncertain about something, seeking information on the matter is reasonably likely to lead to learning. For example, climbing into the attic to investigate a loud noise is likely to reveal new information: the source of the

noise. However, some surprises or uncertainties nonetheless have little potential for learning—for example, one might be highly uncertain whether extraterrestrial life exists yet still expect to learn little by pursuing information search. Thus, surprise and uncertainty often overlap with the epistemically optimal, but they are dissociable.

Surprise and uncertainty instead might be thought of as "heuristic triggers": imperfect cues to expected learning that might be easier to compute or more robust to error. Forming accurate expectations about learning might be cognitively challenging: optimal models of expected learning require the learner to consider multiple possible futures (i.e., "If I explore, what might I find? How much would I learn if I find X rather than Y?"), then integrate them to determine the total potential for information. In fact, people are less likely to seek information that maximizes expected learning when under time pressure (Coenen et al., 2015), suggesting that tracking optimal expected learning is indeed difficult. Surprise and uncertainty, in contrast, depend on the congruity between the current situation and prior beliefs without consideration for possible futures, and thus might be easier to compute.

In Chapter 2, we found that unique variance in curiosity was explained by surprise, uncertainty, and expected learning, with the strongest role for expected learning—however, these potential triggers of curiosity were measured through self-report. If curiosity tracks perceptions of expected learning, but those perceptions are inaccurate, then curiosity falls short of being optimally triggered. Quantifying the actual potential for learning according to an ideal learner model will allow us to test whether curiosity is in fact optimally triggered. Additional related studies (Markant et al., 2015; Markant & Gureckis, 2014a; Poli et al., 2020) have used ideal learner models to quantify surprise, uncertainty, and/or expected learning, but they have considered only subsets of these features and have relied on behavioral measures of exploration.

In addition, some research has quantified and contrasted epistemically optimal and instrumentally optimal exploration (Bramley et al., 2015; Markant & Gureckis, 2012; Meder & Nelson, 2012), but has not considered heuristic epistemic alternatives (like surprise/uncertainty) or heuristic instrumental alternatives (which we introduce in Study 6).

Finally, prior work has largely neglected to disentangle two possible sources of uncertainty, and thus two corresponding targets for expected learning. Research on curiosity has focused on uncertainty or expected learning about the query of interest ("query uncertainty/expected learning"). For example, curiosity about trivia questions is associated with individuals' uncertainty/expected learning about the answer to the trivia question (Gruber et al., 2014; Kang et al., 2009). In contrast, research on information search has focused on uncertainty or expected learning about an ultimate learning goal ("global uncertainty/expected learning"), where each query provides only some information about that goal. For example, in the classic "20 questions" game (in which the player has 20 questions to identify a target object), each question provides only some information about the target object—and adults' and children's questions to identify a target object are aligned with optimal models of expected learning about the target (e.g., Nelson et al., 2014; Ruggeri et al., 2016, 2017). These two sources of uncertainty/expected learning have rarely been contrasted as possible drivers of information search (but see Markant & Gureckis, 2014a) and have not been contrasted as possible triggers of curiosity.

3.1.2 The present research

In the present research, we clarify whether curiosity is epistemically or instrumentally optimal, or whether curiosity instead tracks heuristic cues to learning or utility. In addition, we test whether curiosity is focused towards resolving uncertainty at the level of individual queries,

or whether curiosity tracks uncertainty or expected learning about the ultimate target of inquiry. In short, we ask the following questions: (1) Does curiosity track epistemically optimal features and/or heuristic epistemic alternatives (or neither)? (2) Does curiosity track uncertainty and expected learning about the question of interest and/or uncertainty and expected learning about the ultimate learning goal (or neither)? (3) Does curiosity track instrumentally optimal features and/or heuristic instrumental alternatives (or neither)?

Prior work has generally failed to address these questions, likely because disentangling these potential triggers is methodologically challenging. In the present research, we overcome these methodological challenges by modifying a paradigm developed by Dorfman et al. (2019), which was originally used to test how beliefs about causal structure influence reinforcement learning. This paradigm uses a modified multi-armed bandit task, a popular task for studying reinforcement learning in psychology and neuroscience (e.g., Daw et al., 2006; Lee et al., 2011; Speekenbrink & Konstantinidis, 2015; Steyvers et al., 2009; C. M. Wu et al., 2018). In this task, participants learn about several mines by digging for gold. Each mine has a fixed but unknown probability of producing gold as opposed to rocks, and these reward probabilities can be learned over time by choosing to dig in each mine. However, unlike a standard multi-armed bandit task, each individual outcome (gold or rocks) can be influenced by an intervening agent (e.g., a bandit, who takes away gold and leaves rocks behind). In our modified version of the task, we prompt participants to rate on each trial how curious they are about whether the intervening agent interfered. We develop a Bayesian reinforcement learning model to capture the optimal learner's beliefs on a trial-by-trial basis. This model also quantifies the ideal learner's surprise (how much the outcome on a given trial violates expectations), query uncertainty (uncertainty about whether an intervention occurred), query expected learning (expected learning about whether an

intervention occurred), global uncertainty (uncertainty about the chosen mine's reward probability), and global expected learning (expected learning about the chosen mine's reward probability). We test the extent to which curiosity tracks these model-estimated features. In addition, in some studies, we incentivize participants to perform a specified task after learning about the mines, allowing us to assess whether curiosity tracks the instrumental value of information (i.e., its potential to improve performance on the incentivized task).

Our method has several strengths that other studies of curiosity generally lack. First, we elicit curiosity by self-report, allowing us to measure when participants feel they are curious. This has an advantage over behavioral measures of curiosity because it minimizes contamination by external considerations (e.g., the costs or rewards associated with seeking information). Second, our method involves both query-level and global sources of uncertainty: learning the target information (whether or not an intervention occurred) is not only informative about that question, but sheds light on a broader question of interest (the reward probability of each mine). This is rare in the curiosity literature, where the predominant paradigm measures curiosity about trivia questions (e.g., Dubey & Griffiths, 2020; Gruber et al., 2014; Kang et al., 2009; Vogl et al., 2020), which by definition are largely trivial with regards to broader questions of interest. Finally, learning the query-level information is not necessary for completing the task, as participants can still learn about the mines by merely observing the outcomes of their choices (gold or rocks). As a result, curiosity is essentially optional, even if participants are motivated to learn useful information. This provides a maximally stringent test of the predictors of curiosity, as participants' desire to learn useful information in general (e.g., to learn the reward probability of each mine) can be divorced from participants' curiosity about a specific question at a specific moment (e.g., whether an intervention occurred on one particular trial).

In Studies 4 and 5, we contrast surprise, query uncertainty/expected learning, global uncertainty, and global expected learning, in a task where participants have no specified instrumental goal. In Studies 6 and 7, we additionally test whether curiosity is sensitive to optimal instrumental value or heuristic instrumental value. Across these studies, we find that curiosity is most closely correlated with uncertainty and/or expected learning about the query of interest, and the other potential triggers explain less variance in curiosity.

3.2 Study 4

In Study 4, we test four proposed triggers of curiosity: surprise, query uncertainty/expected learning (which are confounded in this task), global uncertainty, and global expected learning.

3.2.1 Method

Participants

One hundred sixty-four participants (ages 20 to 69; 67 women, 95 men; one non-binary, and one unspecified) were recruited from Amazon Mechanical Turk (MTurk). Participation was restricted to MTurk workers in the United States, who had completed at least 1000 tasks on MTurk with a minimum 99% approval rate. An additional six participants were excluded for failing to answer two multiple-choice attention check questions. All participants in Studies 4-7 provided informed consent under a protocol approved by the Institutional Review Board at Princeton University.

Procedure

The procedure was adapted from Study 1 of Dorfman et al. (2019). Participants completed a modified reinforcement learning task, modeled after a multi-armed bandit problem (see Figure 4). Participants were instructed that they were supervising a team of gold miners, who would dig in two different-colored mines that each had a fixed but unknown probability of

producing gold as opposed to rocks. On each trial, participants would select a mine in which their team of miners would dig, then they would receive either gold or rocks. However, there was also an outside influence on the mines: between-subjects, either a bandit (N = 55), a tycoon (N = 53), or a sheriff (N = 56).[15] Participants were instructed that on 30% of trials, regardless of which mine had been chosen, they would receive an outcome caused by the intervening agent. This varied by condition: an intervention deterministically caused rocks in the bandit condition, deterministically caused gold in the tycoon condition, and randomly caused rocks or gold in the sheriff condition.

Before proceeding to the main task, participants answered two multiple-choice practice questions: 1. How often does the [bandit/tycoon/sheriff] visit the mines? (correct response: 30% of the time) and 2. What does it mean when you see a mine with a different color? (correct response: it has a different probability of giving me gold). If participants answered a question incorrectly, they continued attempting to answer until they selected the correct response.

The main task consisted of 50 trials. Participants were instructed that their goal was to maximize the amount of gold received during this task, but they were not incentivized for their performance. On each trial, participants clicked to select one of two mines, then they received either rocks or gold. The outcomes on each trial were sampled from the generative model (i.e., on 70% of trials, the outcome was determined by the chosen mine's probability of producing gold; on 30% of trials, the outcome was determined by the intervening agent). Unbeknownst to participants, one mine produced gold with 30% probability when an intervention did not occur; the other mine produced gold with 70% probability. When the outcome (rocks or gold) was revealed, participants were prompted: "How curious are you about why this outcome occurred

[15] These conditions were introduced to increase the variability in surprise, uncertainty, and expected learning, and we do not analyze any condition differences.

(i.e., about whether or not the [bandit/tycoon/sheriff] intervened)?". Participants rated their curiosity by dragging a slider, which was initially anchored at the midpoint. The ends of the slider were labeled "Not at all curious" and "Very curious." Slider position was converted to a scale measure ranging from 0 to 100. After rating their curiosity, participants proceeded to the next trial, without learning whether or not an intervention had occurred.

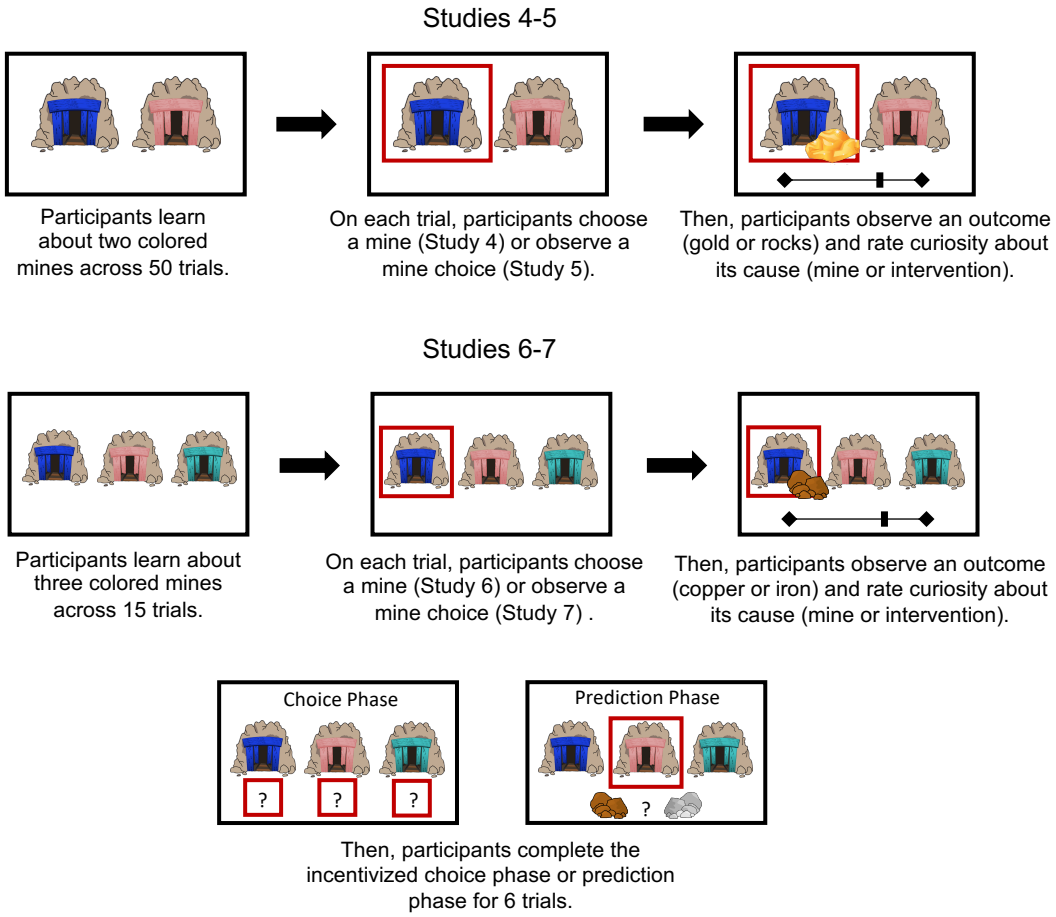


Figure 4 Procedure for Studies 4-7.

After 50 trials, participants answered two attention check questions: 1. Who controlled 30% of the outcomes you saw? (correct response: bandit/sheriff/tycoon, depending on condition), and 2. Which of the following did you never receive from the mines? (correct

response: emeralds). Those who did not answer both questions correctly were excluded from all analyses. Finally, participants provided their gender and age.

Computational model

We fit a Bayesian reinforcement learning model to the choices and outcomes observed by each participant. The model is similar to that presented in Dorfman et al. (2019), but it represents exact posterior distributions over each mine's reward probability, rather than an approximate posterior mean. For ease of exposition, we present the model for a single mine, noting that it is extended to two mines by separately tracking the evidence about each mine with a copy of the single-mine model. The generative model is as follows. On each trial, t, the selected mine produces gold with probability θ and rocks with probability $1 - \theta$. However, a hidden agent intervenes with probability ϵ. This intervention is modeled as a latent variable, $Z_t \sim$ Bernoulli(ϵ), where $Z_t = 1$ means an intervention occurred on trial t. When an intervention does occur, gold is instead produced with probability ζ, which is set to 0, 1, or 0.5 depending on the experimental condition (0 in bandit condition, 1 in tycoon condition, 0.5 in sheriff condition). As a result, the reward R_t for a trial is distributed: $R_t \sim$ Bernoulli$(\theta(1 - Z_t) + \zeta Z_t)$.

The model estimates the value of θ. We assume a uniform prior on θ between 0 and 1. Unfortunately, the standard Beta-Bernoulli process does not apply in this case because of the interventions, Z_t. However, the posterior can still be numerically computed by marginalizing over the sequence of interventions, $\vec{z}$,

$$p(\theta \mid r_{1:t}, \zeta, \varepsilon) \propto \sum_{\vec{z}} \prod_{t} p(z_t \mid \varepsilon) p(r_t \mid \theta, \zeta, z_t).$$

Although marginalizing over $\vec{z}$ directly is intractable, this quantity can be efficiently computed by noting that it depends only on the number of times each possible combination of (r, z) occurs. This derivation is presented in Appendix C. We compute the normalizing constant

by adaptive quadrature. We assume that ζ and ε are known. For concision, we omit the dependence on these parameters in later equations.

On each trial, in addition to estimating θ, the model computes several quantities, defined below and in Table 2.

Table 2 Model-estimated Features and Summary of Results, Studies 4-7

Feature	Goal	Optimal/Heuristic	Definition	Results
Unsigned reward prediction error (uRPE)	Epistemic	Heuristic	Extent to which outcome violates expectations	Related to curiosity when considered in isolation, inconsistent results when controlling for other features
Query Entropy/Expected Information Gain (EIG)	Epistemic	Optimal/Heuristic	Uncertainty/expected learning about whether intervention occurred	Strongly related to curiosity, when considered in isolation and when controlling for other features
Global Entropy	Epistemic	Heuristic	Uncertainty about mine reward probability	Related to curiosity when considered in isolation, inconsistent results when controlling for other features
Global Expected Information Gain (EIG)	Epistemic	Optimal	Expected learning about mine reward probability	Strongly related to curiosity when considered in isolation, inconsistent results when controlling for other features
Value of Information (VOI)	Instrumental	Optimal	Value of learning whether intervention occurred (for choice/prediction task)	Related to curiosity when considered in isolation, inconsistent results when controlling for other features
Value of Perfect Information (VPI)	Instrumental	Optimal	Value of learning mine reward probability (for choice/prediction task)	Only related to curiosity for prediction task when considered in isolation, unrelated to curiosity when controlling for other features
Heuristic Value (HV)	Instrumental	Heuristic	Approximate value of learning mine reward probability (for choice/prediction task)	Only related to curiosity for prediction task when considered in isolation, inconsistent results when controlling for other features

1. Surprise encodes the extent to which the observed reward on a given trial differs from the expected reward. It is defined as the unsigned reward prediction error (uRPE): $|R_t - \overline{r}_t|$,

where the expected reward $\overline{r}_t$ is the posterior mean estimate of θ prior to the current reward, $\int_0^1 p\left(\theta \mid r_{1:t-1}\right)\theta \, d\theta$. We also considered several additional definitions of surprise (uRPE where expected reward is the MAP estimate of θ, uRPE where expected reward weighs the posterior mean estimate of θ against the possibility of intervention, and information-theoretic surprisal). These alternative operationalizations of surprise produced similar results, and our main definition of uRPE accounted for the most variance in curiosity when averaged across all studies (see Appendix D for brief descriptions of alternative operationalizations and results).

2. Uncertainty features describe uncertainty about model estimates, using the information theory measure of entropy (Shannon, 1948). The first feature, *query entropy*, captures uncertainty about whether an intervention occurred (i.e., the target query about which participants rate their curiosity). This is defined as the entropy of the intervention's predictive distribution, $H(Z_t \mid r_{1:t}) = -\overline{z}_t \log \overline{z}_t - (1-\overline{z}_t)\log(1-\overline{z}_t)$ where $\overline{z}_t = p(Z_t = 1 \mid r_{1:t})$ is the conditional probability of the agent having intervened on this trial given all observed rewards,

$$\overline{z}_t = \int_0^1 p\left(\theta \mid r_{1:t}\right) \frac{p(r_t \mid \theta, Z_t = 1)p(Z_t = 1)}{\sum_{z_t' \in \{0,1\}} p(r_t \mid \theta, z_t')p(z_t')} d\theta.$$

The second feature, *global entropy*, captures uncertainty about the estimated probability of receiving gold from the chosen mine, θ. This is defined $H(\theta \mid r_{1:t}) = -\int_0^1 p\left(\theta \mid r_{1:t}\right)\log\left(p(\theta \mid r_{1:t})\right)d\theta$.

3. Expected learning features encode how much learning is expected to occur if the target of one's curiosity (whether an intervention occurred) were to be revealed, operationalized using "expected information gain" (EIG; Lindley, 1956). First, expected learning about whether an intervention occurred, or *query EIG*, is the expected reduction in entropy in the predictive distribution, $H(Z_t \mid r_{1:t})$, after observing the value of Z_t. Because observing the value of Z_t

would reduce entropy to zero, query EIG is equal to query entropy, and we refer to this as "query entropy/EIG." Second, expected learning about the chosen mine's reward probability, or *global EIG*, is the expected reduction in entropy of the posterior distribution of θ after observing the value of Z_t, $\sum_{z\in\{0,1\}} p\,(Z_t = z \mid r_{1:t}) H(\theta \mid r_{1:t}, Z_t = z) - H(\theta \mid r_{1:t})$. Discovering whether an intervention occurred on a given trial only provides some information about the mine's reward probability, so global entropy and global EIG are distinct.

A relatively large body of research (e.g., Coenen et al., 2015; Nelson et al., 2010; Rothe et al., 2018; Steyvers et al., 2003; for a review, see Coenen et al., 2019) has described exploration using a measure similar to or equivalent to our measure of global EIG. In addition, one line of research has attempted to disentangle global EIG from query entropy/EIG (or slight variations), finding that exploration more closely aligns with the latter (Markant et al., 2015; Markant & Gureckis, 2014a). Finally, other research has considered uRPE as a driver of exploration, learning, and memory (e.g., Rouhani et al., 2018; Rouhani & Niv, 2021; Stojić et al., 2020). However, these cues have not been associated with self-report curiosity, and they have not been quantified in a single study.

3.2.2 Results

Below we present analyses addressing the following questions: Are epistemically optimal and heuristic features associated with curiosity? And do epistemically optimal and heuristic features each explain unique variance in curiosity?

Prior to all analyses, we z-score curiosity ratings within participants. This allows us to correct for variation in participants' use of the 100-point scale (e.g., one participant might reserve high values for only extraordinary levels of curiosity, while another might distribute their

modest levels of curiosity across the full scale). All model-estimated features were z-scored across the full dataset. This allows for a standard interpretation of regression coefficients.

We fit mixed-effects regression models using the "lme4" package in R (Bates et al., 2015). We estimate all regression coefficients using restricted maximum likelihood estimation, while maximum likelihood estimation is used for model comparison. In each reported regression model, we include by-participant random intercepts, as well as by-participant random slopes for any within-participant predictors. We additionally fit generalized additive mixed models (GAMMs) using the "mgcv" R package (Wood, 2011), and report p-values estimated from this package using the approach detailed by Wood (2013). GAMMs flexibly allow for virtually any non-linear relationship between predictor variables and response variables and have been used in prior research to assess the correlates of attention during learning (Poli et al., 2020). In each GAMM, we use penalized cubic regression splines to estimate smooth functions, with 10 knots spaced by quantile. Between each knot, the GAMM estimates a cubic polynomial, resulting in a smooth function over the full dataset. Notably, the overall smooth function can take any number of forms, ranging from linear to "wiggly." We also include by-participant random effects for each parameter, allowing for participant-level variation in the extent to which each model-estimated feature is related to curiosity. We estimate the degree of smoothness (i.e., the complexity of each non-linear association) as part of the fitting process using restricted maximum likelihood estimation. This approach balances complexity and goodness-of-fit, resulting in non-linear functions that approximate the data without overfitting.

Are epistemically optimal and heuristic features associated with curiosity?

First, we tested the association between each model-estimated feature and curiosity, considering each feature in isolation. We fit four mixed-effects regression models, each

predicting curiosity as a function of one model-estimated feature (uRPE, query entropy/EIG, global entropy, and global EIG). Each model-estimated feature was significantly associated with curiosity (*p*s < .001; see Table 3). The coefficient estimates for each model-estimated feature are in Figure 5. Query entropy/EIG ($\beta = 0.52$) was strongly associated with curiosity, while uRPE ($\beta = 0.24$), global entropy ($\beta = 0.18$), and global EIG ($\beta = 0.34$) were more weakly associated with curiosity. Curiosity increased as each model-estimated feature increased (see Figure 6).

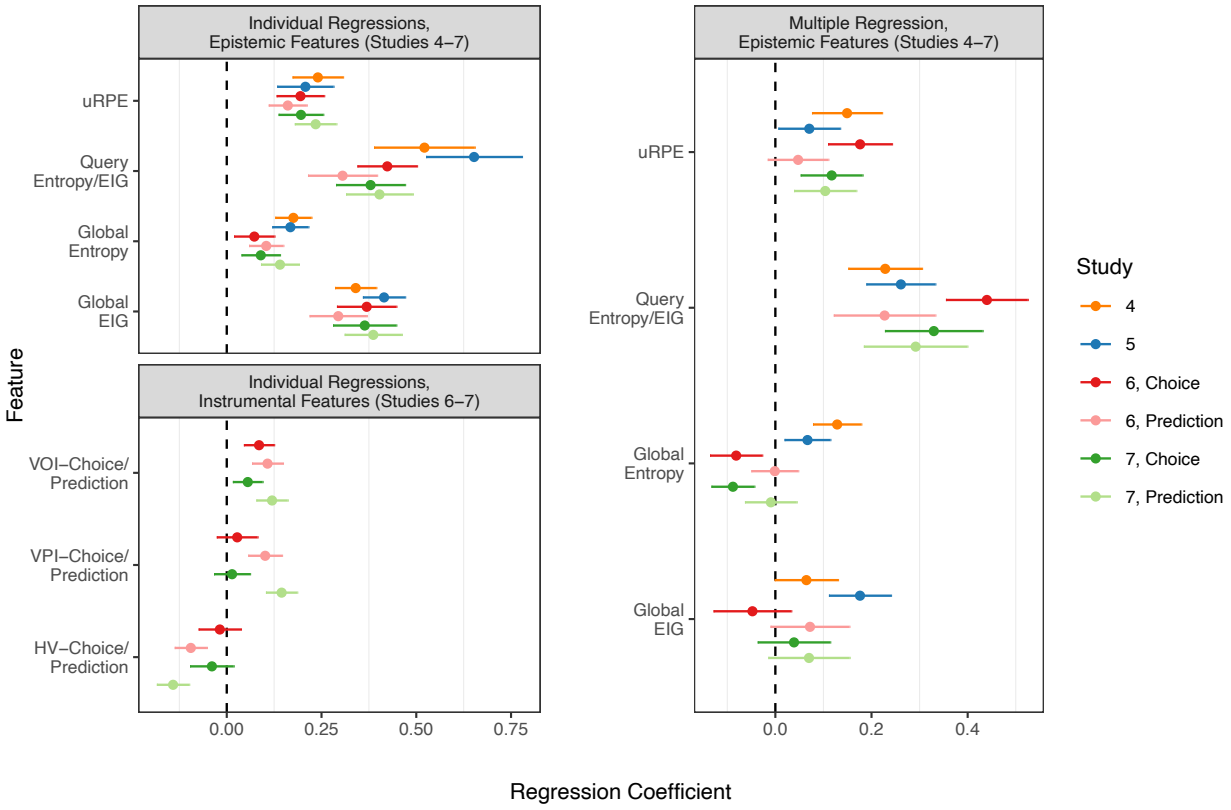


Figure 5 Regression Analysis Results, Studies 4-7.
Left/top: Results of individual regression models relating each epistemic feature to curiosity in Studies 4-7. **Left/bottom:** Results of individual regression models relating each instrumental feature to curiosity in Studies 6-7. **Right:** Results of multiple regression models relating epistemic features to curiosity in Studies 4-7. In all plots, Points indicate estimated regression coefficients; error bars indicate 95% CIs.

We repeated this analysis using GAMMs rather than linear models, allowing us to capture non-linear associations between each feature and curiosity. Each model-estimated feature was significantly related to curiosity (ps < .001). The estimated associations between each feature and curiosity are presented in Figure 6.

Do epistemically optimal and heuristic features each explain unique variance in curiosity?

While these results suggest that query entropy/EIG is the best predictor of curiosity when considered in isolation, it remains possible that these results mask the true determinants of curiosity. Because the model-estimated features were moderately correlated (r ranging from 0.05 and 0.66, see Figure 7), it is possible that the pairwise association between curiosity and any particular feature can be explained away by other features. To disentangle these multiple candidate elicitors of curiosity, we next fit a mixed-effects multiple regression model including all four model-estimated features as fixed effects. In this model, uRPE, query entropy/EIG, and global entropy explained unique variance in curiosity (ps < .001), while global EIG was not significant (p = .05; see Table 3). The coefficient estimates for each fixed effect are presented in Figure 5, and the predicted associations are presented in Figure 6. Query entropy/EIG (β = 0.23) was again the strongest predictor of curiosity, and uRPE (β = 0.15), global entropy (β = 0.13), and global EIG (β = 0.06) were weaker.

We additionally fit an equivalent GAMM, predicting curiosity as a function of all four model-estimated features. All four features explained unique variance in curiosity: uRPE (p = .01), query entropy/EIG (p < .001), global entropy (p < .001), and global EIG (p = .02). The predicted associations between each feature and curiosity are presented in Figure 6.

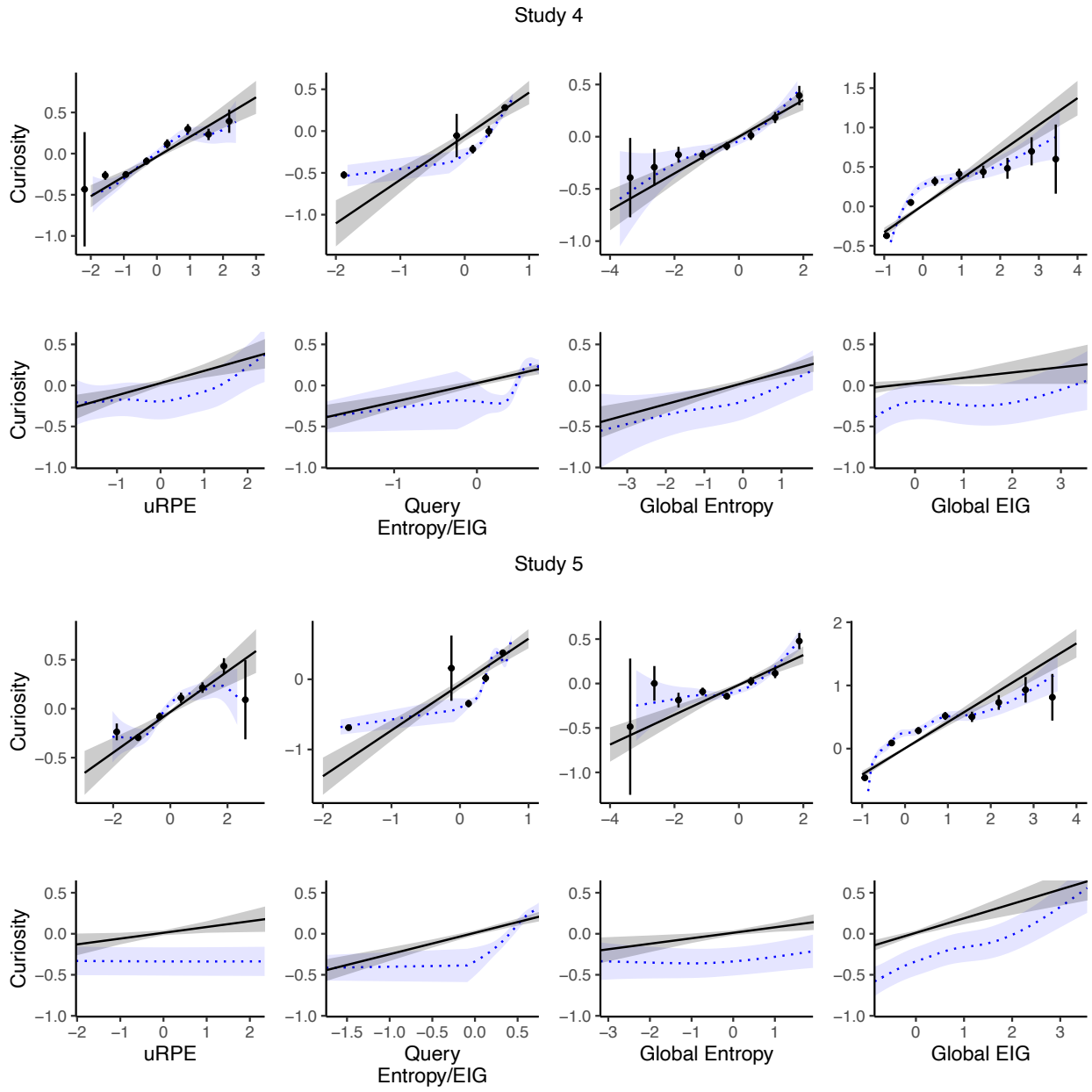


Figure 6 Associations Between Epistemic Features and Curiosity, Studies 4-5.
Fitted associations with 95% CIs between each feature and curiosity in individual regression models (first and third row) and multiple regression models (second and fourth row). Black solid lines with gray ribbons are the results of generalized linear mixed-effects models; blue dotted lines with blue ribbons are the results of generalized additive mixed models. On first and third row, points are average curiosity ratings at binned levels of the relevant feature; error bars indicate 95% CIs. All fitted associations assume random effects are equal to zero.

Table 3 Individual Regression Analyses Significance Tests, Studies 4-7

Likelihood ratio tests for individual regression models predicting curiosity, with a separate model for each epistemic feature. Each model was compared to a reduced model, excluding the fixed effect for the relevant feature (but retaining all random effects).

Predictor	Study 4	Study 5	Study 6, Prediction	Study 6, Choice	Study 7, Prediction	Study 7, Choice
uRPE	$\chi^2(1) = 44.01$, $p < .001$	$\chi^2(1) = 27.50$, $p < .001$	$\chi^2(1) = 30.87$, $p < .001$	$\chi^2(1) = 29.99$, $p < .001$	$\chi^2(1) = 48.55$, $p < .001$	$\chi^2(1) = 33.82$, $p < .001$
Query Entropy/EIG	$\chi^2(1) = 51.01$, $p < .001$	$\chi^2(1) = 77.26$, $p < .001$	$\chi^2(1) = 32.35$, $p < .001$	$\chi^2(1) = 58.01$, $p < .001$	$\chi^2(1) = 48.30$, $p < .001$	$\chi^2(1) = 44.44$, $p < .001$
Global Entropy	$\chi^2(1) = 45.41$, $p < .001$	$\chi^2(1) = 42.28$, $p < .001$	$\chi^2(1) = 17.40$, $p < .001$	$\chi^2(1) = 7.05$, $p = .008$	$\chi^2(1) = 25.76$, $p < .001$	$\chi^2(1) = 11.26$, $p < .001$
Global EIG	$\chi^2(1) = 107.20$, $p < .001$	$\chi^2(1) = 135.79$, $p < .001$	$\chi^2(1) = 43.69$, $p < .001$	$\chi^2(1) = 57.85$, $p < .001$	$\chi^2(1) = 63.56$, $p < .001$	$\chi^2(1) = 52.38$, $p < .001$
VOI (prediction/ choice)			$\chi^2(1) = 24.86$, $p < .001$	$\chi^2(1) = 16.95$, $p < .001$	$\chi^2(1) = 29.43$, $p < .001$	$\chi^2(1) = 7.59$, $p = .006$
VPI (prediction/ choice)			$\chi^2(1) = 17.90$, $p < .001$	$\chi^2(1) = 1.02$, $p = .31$	$\chi^2(1) = 36.61$, $p < .001$	$\chi^2(1) = 0.34$, $p = .56$
HV (prediction/ choice)			$\chi^2(1) = 16.79$, $p < .001$	$\chi^2(1) = 0.42$, $p = .52$	$\chi^2(1) = 32.77$, $p < .001$	$\chi^2(1) = 1.78$, $p = .18$

Table 4 Multiple Regression Analyses Significance Tests, Studies 4-7

Significance tests for multiple regression models predicting curiosity, with all four epistemic features in a single model. Each model was compared to a reduced model, excluding the fixed effect for the relevant feature (but retaining all random effects).

Predictor	Study 4	Study 5	Study 6, Prediction	Study 6, Choice	Study 7, Prediction	Study 7, Choice
uRPE	$\chi^2(1) = 15.47$, $p < .001$	$\chi^2(1) = 4.56$, $p = .03$	$\chi^2(1) = 2.19$, $p = .14$	$\chi^2(1) = 24.91$, $p < .001$	$\chi^2(1) = 9.55$, $p = .002$	$\chi^2(1) = 12.44$, $p < .001$
Query Entropy/EIG	$\chi^2(1) = 28.27$, $p < .001$	$\chi^2(1) = 39.29$, $p < .001$	$\chi^2(1) = 15.25$, $p < .001$	$\chi^2(1) = 56.52$, $p < .001$	$\chi^2(1) = 21.04$, $p < .001$	$\chi^2(1) = 29.98$, $p < .001$
Global Entropy	$\chi^2(1) = 23.45$, $p < .001$	$\chi^2(1) = 7.41$, $p = .007$	$\chi^2(1) = .002$, $p = .96$	$\chi^2(1) = 8.25$, $p = .004$	$\chi^2(1) = 0.12$, $p = .73$	$\chi^2(1) = 13.63$, $p < .001$
Global EIG	$\chi^2(1) = 3.69$, $p = .05$	$\chi^2(1) = 26.45$, $p < .001$	$\chi^2(1) = 2.89$, $p = .09$	$\chi^2(1) = 1.34$, $p = .25$	$\chi^2(1) = 2.60$, $p = .11$	$\chi^2(1) = 1.01$, $p = .32$

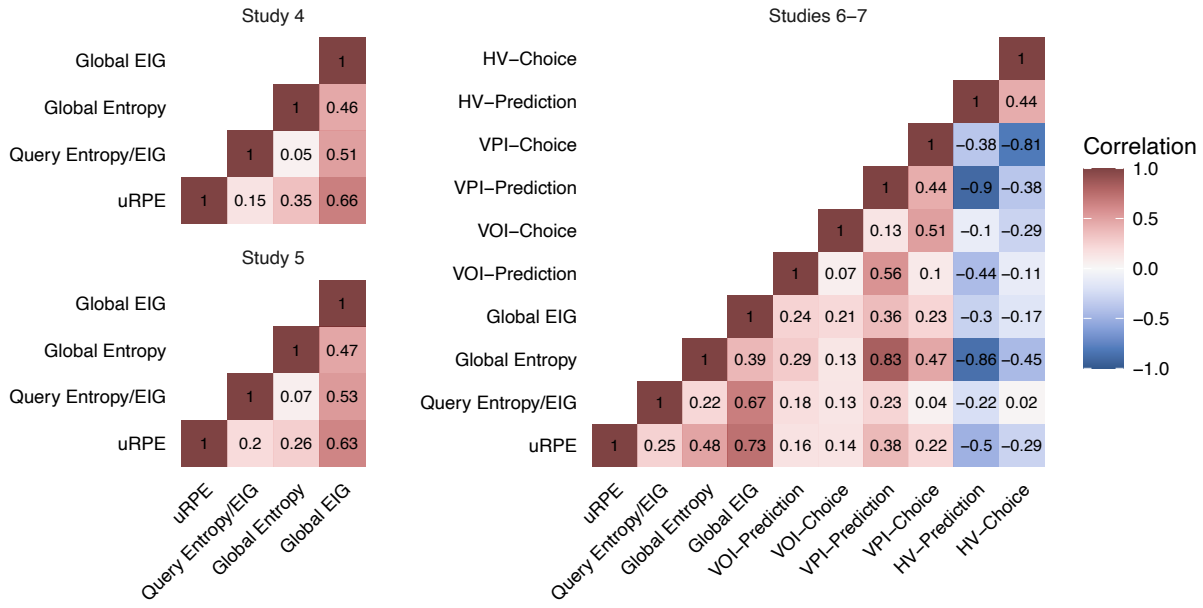


Figure 7 Model-Estimated Feature Correlations, Studies 4-7
Pairwise Pearson correlations between trial-by-trial estimates of each model-estimated feature for Study 4 (left/top), Study 5 (right/bottom), and Studies 6-7 combined (right).

3.2.3 Discussion

In Study 4, we tested whether epistemically optimal and heuristic features were associated with curiosity, as well as which explained unique variance. We operationalized these proposed triggers using a Bayesian reinforcement learning model, resulting in estimates of trial-by-trial surprise (uRPE), uncertainty/expected learning about the target of curiosity (query entropy/EIG), uncertainty about the chosen mine (global entropy), and expected learning about the chosen mine (global EIG). We found that all of the potential triggers were significantly related to curiosity in isolation. However, when estimating the unique variance explained by each feature in a multiple regression model, query entropy/EIG was the strongest predictor of curiosity, while the other features explained somewhat less variance in curiosity, with global EIG failing to reach significance in the multiple regression analysis. Notably, this raises the question

of whether query entropy or query EIG is in fact driving the strong association between query entropy/EIG and curiosity—we return to this question in the General Discussion.

Aside from query entropy/EIG (which conflates heuristic and optimal triggers), heuristic triggers—uRPE and global entropy—explained unique variance in curiosity, while an epistemically optimal trigger—global EIG—did not. These findings do not support the hypothesis that curiosity is epistemically optimal at a global level: expected reduction in uncertainty about the true reward probability of the mine did not predict curiosity above and beyond other features. One possible explanation for this null result is that our task captures only low levels of global EIG: trial-by-trial interventions are somewhat informative about the true reward probability of the mines, but it is possible to learn about the mines simply by observing trial-by-trial rewards. In a task where global EIG were generally higher (e.g., where learning trial-by-trial interventions was critical for learning mine reward probabilities), curiosity might track this feature more closely. We return to this possibility in the General Discussion. A related concern is that participants' reliance on heuristic triggers results from their focus on choosing between the two mines. As participants' goal was to maximize gold by choosing between the mines, they may have (1) prioritized choice between the mines as a means for seeking information (e.g., choosing the mine on the left to gain more information about it), and (2) prioritized finding rewards as the focus of their cognitive effort (e.g., quickly rating curiosity so that they could move onto the next trial to find gold). Both possibilities would likely shift participants' attention away from curiosity. If tracking optimal features requires sufficient attention or cognitive resources (see Coenen et al., 2015), we might expect curiosity to track optimal features more closely when participants' sole task is rating curiosity. We test this possibility in Study 5.

3.3 Study 5

In Study 5, we attempt to replicate the results of Study 4 with a modified method. Instead of allowing participants to choose between the mines on each trial, we present participants with a sequence of fixed choices between the two mines. In addition, participants are not instructed to maximize reward. If curiosity is more responsive to optimal triggers when participants' attention and cognitive resources are solely directed towards curiosity ratings, we would expect global EIG (and perhaps query entropy/EIG) to predict curiosity more strongly in this experiment.

3.3.1 Method

Participants

One hundred sixty-one participants (ages 18 to 78; 60 women, 97 men; three non-binary, and one unspecified gender) were recruited from MTurk, restricted to the same population as Study 4. An additional four participants were excluded for providing the same curiosity rating on all 50 trials, and an additional seven participants were excluded for failing to answer two multiple-choice attention check questions.

Procedure

The procedure was nearly identical to that of Study 4 (see Figure 4). Participants were told that their team of miners was responsible for deciding where to dig, so they observed all choices and outcomes as a third party. As in Study 1, participants were randomly assigned to the bandit condition (N = 54), the tycoon condition (N = 55), or the sheriff condition (N = 52). The selected mines and outcomes on each trial were generated by randomly yoking participants to a condition-matched participant in Dorfman et al.'s (2019) Study 1, in which participants made their own selections. Participants were also not instructed to seek gold; to motivate them to pay

attention, participants were simply instructed, “You may be asked questions about what you have seen at the end of the task.”

3.3.2 Results

We again address the following questions: Are epistemically optimal and heuristic features associated with curiosity? And do epistemically optimal and heuristic features each explain unique variance in curiosity? In addition, we quantitatively compare Studies 4 and 5. We use the same analytic approach as described in Study 4, and we again z-score curiosity ratings within participants and z-score all model-estimated features across the full dataset.

Are epistemically optimal and heuristic features associated with curiosity?

First, we fit four mixed-effects regression models, each predicting curiosity ratings as a function of one model-estimated feature (uRPE, query entropy/EIG, global entropy, and global EIG). Each feature was significantly associated with curiosity (ps < .001; see Table 3). Like Study 4, query entropy/EIG ($\beta = 0.65$) was most strongly associated with curiosity, followed by global EIG ($\beta = 0.42$), uRPE ($\beta = 0.21$), and global entropy ($\beta = 0.17$; see Figure 5). Curiosity increased as each model-estimated feature increased (see Figure 6). Like Study 4, the GAMM analysis also revealed significant associations between each feature and curiosity ($ps < .001$; see Figure 6).

Do epistemically optimal and heuristic features each explain unique variance in curiosity?

To disentangle the candidate elicitors of curiosity, we next fit a multiple regression model including all four model-estimated features as fixed effects. All four features explained unique variance in curiosity ($ps < .05$; see Table 4). While query entropy/EIG ($\beta = 0.26$) was again the strongest predictor of curiosity, global EIG ($\beta = 0.18$) also explained unique variance in curiosity, unlike in Study 4. uRPE ($\beta = 0.07$) and global entropy ($\beta = 0.07$) were somewhat

weaker predictors (see Figure 5). The GAMM analysis revealed slightly different results, with uRPE failing to reach significance ($p = .54$). In the GAMM, only query entropy/EIG ($p < .001$), global entropy ($p = .03$), and global EIG ($p < .001$) were significant (see Figure 6).

Was curiosity differentially associated with epistemically optimal or heuristic features in Study 5 versus Study 4?

Finally, we compared Studies 4 and 5 directly. We fit additional individual regression models (with each feature in a separate model), adding study and the interaction between study and the relevant feature as fixed effects. There was no evidence for an interaction between task condition and uRPE, $\chi^2(1) = 0.43$, $p = .51$, query entropy/EIG, $\chi^2(1) = 1.60$, $p = .21$, global entropy, $\chi^2(1) = 0.05$, $p = .82$, or global EIG, $\chi^2(1) = 3.52$, $p = .06$. That is, there was no evidence that the four features were differentially associated with curiosity across studies when considered in isolation.

To compare the multiple regression analyses, we fit an additional multiple regression model, adding study and the interactions between study and each feature as fixed effects. We compared this model to a set of reduced models, which each excluded the interaction term between study and one feature. There was no evidence for a significant interaction between study and uRPE, $\chi^2(1) = 2.81$, $p = .09$, query entropy/EIG, $\chi^2(1) = 1.26$, $p = .26$, or global entropy, $\chi^2(1) = 2.63$, $p = .10$. However, there was a significant interaction between study and global EIG, $\chi^2(1) = 5.21$, $p = .02$, indicating that curiosity was differentially associated with global EIG across studies when all other epistemic features were held fixed. While these studies were conducted at different times and participants were not randomly assigned to studies, it is possible that the variation in participants' task between Studies 4 and 5 influenced the sensitivity of curiosity to global EIG.

3.3.3 Discussion

In Study 5, we attempted to replicate the results of Study 4 with a methodological modification: rather than choosing between the mines themselves to maximize reward, participants were third-party observers to the reinforcement learning task. If participants in Study 4 focused on choosing mines to gain information and maximize reward at the expense of attending to their curiosity, it is possible that we underestimated the extent to which curiosity tracks optimal triggers. Consistent with this possibility, in Study 5, we found a stronger positive association between global EIG and curiosity when controlling for the other model-estimated features. However, across both studies, query entropy/EIG explained the most variance in curiosity, and heuristic features (uRPE and global entropy) additionally explained unique variance in curiosity. These results suggest that curiosity most strongly tracks uncertainty or expected learning about the question of interest (which are confounded in this task) and might only additionally track global expected learning to the extent that cognitive resources are readily available (see Coenen et al., 2015).

3.4 Study 6

In Study 6, we attempt to replicate our Study 4-5 results, testing again whether curiosity tracks surprise, uncertainty, and/or expected learning. In addition, we specify and manipulate participants' instrumental goal, in order to test whether curiosity additionally tracks the expected value of information.

To do so, we modify our experimental task so that participants alternate between an "observation phase" and a "test phase." In the observation phase, participants observe outcomes from several mines in order to learn each mine's reward probability, rating their curiosity after each outcome. This observation is in the service of achieving one of two incentivized goals

during the test phase: either predicting the outcomes from randomly selected mines (prediction goal) or choosing between the mines to maximize one outcome (choice goal). If curiosity is sensitive to instrumental value, curiosity should differ systematically as a function of participants' incentivized goal. In particular, participants with a prediction goal should be curious to the extent that information would be revealing about mines with difficult-to-predict outcomes (i.e., those with a reward probability close to 50%). In contrast, participants with a choice goal should be curious to the extent that information would be revealing about mines that are potential candidates for choice (i.e., high reward probability mines, especially those with other high reward probability competitors). To test these predictions, we quantify the instrumental value of information for achieving participants' ultimate goals on a trial-by-trial basis. In addition, participants learn about three mines, where one mine has a true reward probability close to 50% (and thus is valuable to learn about in the prediction condition), and two mines are high reward probability competitors (and thus are valuable to learn about in the choice condition).

Like Study 4, we allow participants to make their own choices between the mines during the observation phase. Though it is possible that the ability to choose between the mines will reduce participants' reliance on epistemically and/or instrumentally optimal features (e.g., by directing their attention away from curiosity), this allows us to test whether our task manipulation is salient enough to shift participants' behavior. Following this validation of our manipulation, we test in Study 7 whether curiosity tracks instrumentally optimal features when participants cannot also choose between the mines.

3.4.1 Method

Participants

One hundred fifty-four participants (ages 18 to 68; 68 women, 85 men; one non-binary) were recruited from Prolific. Participants were required to reside in the United States and to have completed at least 100 previous studies on Prolific with a minimum 95% approval rate. An additional six participants were excluded for failing to answer two multiple-choice attention check questions, and an additional four participants were excluded for providing the same curiosity rating on all trials.

Procedure

The procedure was similar to that of Studies 4-5 (see Figure 4). Participants were instructed that they and a team of miners would be digging in sets of three mines at different mining sites. One mining site was used to introduce participants to the task, while the second and third mining sites comprised the "main task." We increased the number of mines from two to three to allow us to maximally differentiate between the two incentivized task conditions. Rather than providing rocks and gold (as in the previous studies) the mines produced copper and iron, so that both outcomes appeared reasonably valuable. Participants were instructed that different colored mines had different probabilities of producing copper as opposed to iron, and three different colored mines appeared at each mining site. In actuality, the mines at each site had a 20%, 65% and 80% probability of producing copper as opposed to iron, with the position of those reward probabilities randomly assigned to the three mines at each site.

Participants' goal was to complete an incentivized "prediction phase" (in the prediction condition, N = 77) or "choice phase" (in the choice condition, N = 77) at each mining site. In the incentivized prediction phase, participants were tasked with predicting the outcomes (copper or

iron) of fixed choices between the three mines. Each time a participant made a correct prediction, their payment increased by five cents. During each prediction phase, the three mines were chosen on two trials each (six total trials); outcomes were generated according to the mines' reward probabilities. In the choice phase, participants were tasked with choosing between the three mines on six trials in order to gain as much copper as possible. Each time copper was received, a participant's payment increased by five cents. Outcomes of participants' choices were generated according to the mines' reward probabilities.

Prior to the choice or prediction phase at a given mining site, participants completed the 15-trial "observation phase." In the observation phase, participants chose a mine, then observed the outcome (copper or iron). On the first six observation trials, choices were fixed so that participants were required to choose each mine twice. On these six trials, outcomes were fixed so that copper was received after both 80% mine choices, iron was received after both 20% mine choices, and copper and iron were each received once after the two 65% mine choices (so that initial evidence was consistent with the true mine values). On the remainder of trials, participants chose freely between the three mines and rewards were sampled from the generative model. As in Studies 4-5, the sampled outcomes were influenced by an intervening agent during the observation phase: the copper bandit (N = 47), the copper tycoon (N = 52), or the sheriff (N = 52). The hidden agent intervened on 30% of trials; the copper bandit always caused iron, the copper tycoon always caused copper, and the sheriff caused copper or iron at random. During the observation phase, participants rated their curiosity about whether an intervention had occurred after each outcome on a 100-point slider. We occasionally sated participants' curiosity in an attempt to increase engagement: whether or not an intervention actually occurred was revealed on 20% of observation-phase trials (after participants rated their curiosity).

As an introduction to the task, participants completed a three-trial choice/prediction phase, followed by a six-trial observation phase (with all choices determined by the participant and no curiosity ratings), followed by an additional three-trial choice/prediction phase. Following this introduction, participants completed the main task: a 15-trial observation phase followed by a six-trial choice/prediction phase, repeated at two mining sites.

As in Studies 4-5, participants answered two attention check questions at the end of the task: 1. Who controlled 30% of the outcomes you saw? (correct response: copper bandit/copper tycoon/sheriff, depending on condition), and 2. Which of the following did you never receive from the mines? (correct response: gold). Those who did not answer both questions correctly were excluded from all analyses. Finally, participants provided their gender and age.

Computational model

We use the same model described in Study 4 (extended to three mines), with several modifications and additions. To model the occasional observation of the latent agent, we introduce a new variable $\tilde{z}_t$ that is equal to z_t on trials where the latent agent was observed and takes a null value, $\varnothing$, otherwise:

$$p(\tilde{z}_t \mid z_t) = \begin{cases} 0.8 & \text{if } \tilde{z}_t = \varnothing \\ 0.2 & \text{if } \tilde{z}_t = z_t \\ 0 & \text{otherwise} \end{cases}$$

This results in a new posterior,

$$p(\theta \mid r_{1:t}, \tilde{z}_{1:t}, \zeta, \varepsilon) \propto \sum_{z_{1:t}} \prod_t p\,(z_t \mid \varepsilon) p(\tilde{z}_t \mid z_t) p(r_t \mid \theta, \zeta, z_t).$$

where the final $\tilde{z}_t$ is fixed to be $\varnothing$ because the intervention is never observed before querying curiosity.

Because of this change to the definition of the posterior, all model estimates that condition on $r_{1:t}$ as described in Study 4 also condition on $\tilde{z}_{1:t-1}$ in the revised model. In addition to estimating uRPE, query entropy/EIG, value entropy, and value EIG on each trial, we estimate *instrumental features*, which capture the extent to which learning about the intervention on a given trial is likely to increase rewards during the choice or prediction phase. These features are described below.

Objective Functions To model the potential influence of instrumental value, we define for each task an objective function f, which gives the expected performance on the test task given the current state of knowledge.

For the prediction task, this is defined as the probability of correctly guessing whether a reward is received on the test trial, a random variable we denote R_T. The probability of a correct guess is

$$f^{\text{pred}}(r_{1:t}, \tilde{z}_{1:t}) = p(R_T = \hat{r}_T) = \max\{\hat{p}, 1 - \hat{p}\} \text{ where } \hat{p} = p(\hat{r}_T \mid r_{1:t}, \tilde{z}_{1:t}).$$

For the choice task, the objective function is the probability of receiving a reward from the chosen mine. One important property of the choice objective is that it depends on all three mines. That is, we can no longer treat the mines independently. We thus introduce the superscript notation $r^{(i)}$ to denote a reward from mine i. We can now define the chosen mine as the one with maximal reward probability,

$$\text{chosen} = \text{argmax}_i \hat{p}^{(i)} \text{ where } \hat{p}^{(i)} = p(\hat{r}_T^{(i)} \mid r_{1:t}^{(i)}, \tilde{z}_{1:t}^{(i)}).$$

The choice objective is then the probability of actually receiving reward from that mine,

$$f^{\text{choice}}\left(\{r_{1:t}, \tilde{z}_{1:t}\}^{(i)}\right) = \hat{p}^{(\text{chosen})} = \max_i \hat{p}^{(i)}.$$

Note that this objective is a function of the data from all the mines because the mines cannot be treated independently in this case. For consistency, we can redefine the prediction

objective as a function of all the mines by simply taking the average of each single-mine objective. Secondly, note that we have defined these objective functions for single trials. The true objective function for the task scales each of these by the number of test trials; however, because we z-score all the features before the regression, this scaling constant can be ignored.

Value of Information The *value of information (VOI)* is the expected increase in the objective function the learner will achieve if they learn a new piece of information. The VOI depends on three factors: the incentivized task (prediction or choice), the previously seen data, and the random variable that will be observed. For example, for a participant in the prediction condition, the value of learning whether the agent intervened on trial t is defined

$$\mathrm{VOI}^{\mathrm{pred}}(Z_t; r_{1:t}, \tilde{z}_{1:t-1}) = \mathrm{E}_{z_t}\left[\mathrm{f}^{\mathrm{pred}}(r_{1:t}, \tilde{z}_{1:t-1}, z_t)\right] - \mathrm{f}^{\mathrm{pred}}(r_{1:t}, \tilde{z}_{1:t-1}).$$

That is, the VOI is the expected accuracy when the agent makes predictions based on all the previous data and also knowledge of the intervention on the current trial, minus the prediction accuracy given only the previous data. Note that $\mathrm{f}^{\mathrm{pred}}(r_{1:t}, \tilde{z}_{1:t-1}, z_t)$ should be interpreted as $\mathrm{f}^{\mathrm{pred}}(r_{1:t}, \tilde{z}_{1:t})$ where $\tilde{z}_t := z_t$.

Choice VOI is defined similarly, though, as before, we must take into account information about all the mines, not just the one selected on the current trial. In the choice condition, the value of learning whether the agent intervened on trial t when mine j was chosen is

$$\mathrm{VOI}^{\mathrm{choice}}\left(Z_t^{(j)}; \{r_{1:t}, \tilde{z}_{1:t-1}\}^{(i)}\right) = \mathrm{E}_{z_t}\left[\mathrm{f}^{\mathrm{choice}}\left(\{r_{1:t}, \tilde{z}_{1:t-1}\}^{(i)}, z_t^{(j)}\right)\right] - \mathrm{f}^{\mathrm{choice}}\left(\{r_{1:t}, \tilde{z}_{1:t-1}\}^{(i)}\right).$$

Note that in both cases we take an expectation over the possible values of the unobserved variable, just as we did for EIG. Indeed, EIG can be understood as a special case of VOI where the objective function is the negative entropy of the posterior.

One potential problem with the above VOI features is that they are myopic. That is, they implicitly assume that this is the last piece of information that will be gained before the test trial, and thus do not account for the possibility that future information will make the current information more useful. As a result, these features systematically underestimate the true value of observing the respective variables (Hay et al., 2012). In particular, when observations are limited to a small set of discrete values, the myopic VOI will often be exactly zero.[16]

To address this deficiency of the myopic VOI estimate, previous work in artificial intelligence (Callaway et al., 2018; Dearden et al., 1998; Howard, 1966; Matheson, 1968) and psychology (Callaway et al., 2021) has supplemented the myopic VOI with an upper bound on the VOI, the value of having perfect information. In our experiment, we can define the *value of perfect information (VPI)* as the value of observing θ. That is, how much would we expect our predictions or choices to improve if we knew the true reward probability of one of the mines? This feature provides a (loose) upper bound on the true value of observing an intervention for the specified mine.

The VPI for the prediction task is defined

$$\mathrm{VPI}^{\mathrm{pred}}(\theta; r_{1:t}, \tilde{z}_{1:t-1}) = \mathrm{E}_{\theta}[\max\{\theta, 1-\theta\}] - \max\{\hat{p}, 1-\hat{p}\},$$

where $\hat{p} = p(\hat{r}_T \mid r_{1:t}, \tilde{z}_{1:t-1})$ as before. Note that the second term can also be written as $\mathrm{f}^{\mathrm{pred}}(r_{1:t}, \tilde{z}_{1:t-1})$.

[16] To see why, consider the case where no interventions occur ($\varepsilon = 0$), and imagine that one mine has produced reward on six out of eight trials, resulting in a posterior mean of 70% (the expectation of $\mathrm{Beta}(1+6, 1+2)$). Critically, observing a ninth reward cannot possibly change your prediction in the test trial. After observing that reward, your posterior mean would be either 64% or 73%; in both cases you will predict reward. Given that the information can't possibly change your decision, it has no value. A similar line of reasoning can be applied in the choice condition and in the presence of the latent agent.

The choice VPI feature is defined similarly. Again, we need to take into account the data for all the mines when computing the value of learning about any given mine. Thus, we have

$$\mathrm{VPI}^{\mathrm{choice}}\left(\theta^{(j)}; \{r_{1:t}, \tilde{z}_{1:t-1}\}^{(i)}\right) = \mathrm{E}_{\theta^{(j)}}\left[\max\left\{\theta^{(j)}, \max_{i \neq j} \hat{p}^{(i)}\right\}\right] - \max_i \{\hat{p}^{(i)}\},$$

Note that the term inside the expectation could also be written

$$\max_i \{\hat{p}^{(i)}\} \text{ where } p^{(j)} := \theta^{(j)}.$$

For both features, the expectations are taken with respect to the current posterior, $p(\theta \mid r_{1:t}, \tilde{z}_{1:t-1})$.

Heuristic Instrumental Features The defined features (*VOI-prediction, VOI-choice, VPI-prediction, VPI-choice*) reflect the optimal value of information based on the incentivized task (prediction or choice), the previously seen data, and the expected information about agent intervention on trial t. Prior research has not typically considered heuristic alternatives to these instrumentally optimal features, but such alternatives are possible. We construct a set of heuristic features that consider the incentivized task and the previously seen data, but do not take into account the expected information about the agent intervention (just as epistemic heuristic features take into account the previously seen data but not expected information).

Intuitively, in the prediction condition, the probability of a correct guess will be lowest for mines with θ near 50%. It is therefore instrumentally valuable for the prediction task (in general) to learn about mines with θ near 50%. Thus, we define *heuristic value (HV)* for the prediction condition *HV-prediction* $= |0.5 - \theta|$. In contrast, in the choice condition, the probability of choosing the mine with the maximal reward probability will be lowest for mines with θ near other high-θ competitors (for example, it is more difficult to determine which of two mines is better when $\theta_1 = 0.7$ and $\theta_2 = 0.75$, compared to when $\theta_1 = 0.3$ and $\theta_2 = 0.75$). It is therefore instrumentally valuable for the choice task (in general) to learn about mines with θ near

a high-θ competitor. Thus, we define $\textit{HV-choice} = |\theta^{(j)} - \max_{i \neq j} \hat{p}^{(i)}|$. Notably, the predicted associations between curiosity and the HV features are negative: curiosity should be higher to the extent that HV-prediction or HV-choice is lower.

The HV features were constructed on the basis of intuition. To validate that these features are reasonable heuristic guides to the instrumentally optimal features, we can examine the correlations between the HV feature, the VOI features, and the VPI features. (see Figure 7). Collapsing across Studies 6 and 7, HV-prediction was only modestly correlated with VOI-prediction (r = -0.44), and HV-choice was only modestly correlated with VOI-choice (r = -0.29). However, HV-prediction was strongly correlated with VPI-prediction (r = -0.90), and HV-choice was strongly correlated with VPI-choice (r = -0.81). This validates that the heuristic features are reliable guides to the VPI features (but perhaps not to the VOI features).

Mine Choices To analyze whether participants' choices between the mines were influenced by instrumental value, we also estimated VPI-prediction, VPI-choice, HV-choice, and HV-prediction for all three mines before a mine was selected on each trial. To do so, we simply omit r_t from the conditioning data, as it has not yet been observed. VOI-prediction and VOI-choice specifically concern the value of learning whether an intervention occurred on a given trial and are therefore not relevant for predicting choices between the three mines.

3.4.2 Results

We present analyses addressing the following questions: (1) Are mine choices affected by the instrumental value of information? (2) Is curiosity affected by the instrumental value of information? (3) Does curiosity track optimal or heuristic epistemic features? And (4) Does curiosity track instrumental features (above and beyond epistemic features)? The first set of analyses validates our task manipulation, and the second set of analyses tests whether this

manipulation influences curiosity. The third set of analyses attempts to replicate our results from Studies 4-5, and the final set of analyses addresses whether instrumental features account for variance in curiosity that is not captured by epistemic features.

We use the same analytic approach as described in Study 4, and we again z-score curiosity ratings within participants and z-score all model-estimated features across the full dataset. Study 6 was preregistered (https://osf.io/pua8r/?view_only=2d0790fafc6b43b091aba139b227bc46), but we deviate from our preregistered analysis plan in three ways. First, rather than using statistically inappropriate models (i.e., classic linear regression models), we deviate from our preregistration to use mixed-effects models as used in Studies 4-5, which allow us to account for the clustered nature of our data with random slopes and intercepts. Second, we preregistered analyses for several other model estimates, including alternative operationalizations of surprise, uncertainty, expected learning, and instrumental value. We report the features that best account for variation in curiosity across the studies, and we omit the remaining features for conciseness (see Appendix D for brief descriptions and results). Finally, because of these deviations, we default to more conservative two-tailed tests, despite preregistering several directional hypotheses.

Are mine choices affected by the instrumental value of information?

First, we tested whether participants' mine choices were sensitive to task condition, in order to validate our manipulation. For each participant, we calculated the proportion of trials on which they chose each mine (20%, 65%, 80%) during the 18 free-choice observation phase trials, collapsing across the two mining sites. We analyzed whether these proportions differed across conditions and across mines using non-parametric tests (Mann-Whitney U tests for between-subjects comparisons; Wilcoxon signed-rank tests for within-subjects comparisons). We

predicted that participants in the prediction condition would preferentially select the mine with true reward probability closest to 50% (i.e., the 65% mine), while participants in the choice condition would preferentially select the two mines with high true reward probability (i.e., the 65% mine and the 80% mine). To test this, we compared the proportion of trials on which participants chose each mine within each condition. Our predictions were generally supported by the data (see Figure 8), though the contrast between the 80% and 65% mines was not significant in the prediction condition. Comparing the relative frequency of choices between the two conditions, participants in the choice condition chose the 80% mine on more trials than participants in the prediction condition, $W = 4116$, $p < .001$ effect size $r = 0.34$, while participants in the prediction condition chose the 20% mine on more trials than participants in the choice condition, $W = 1640$, $p < .001$, $r = 0.39$. There was no difference in the proportion of trials on which participants chose the 65% mine by condition, $W = 2818$, $p = .59$, $r = 0.04$.

While this provides preliminary evidence that participants' mine choices were sensitive to the instrumental value of information, we next conducted a more fine-grained test of our manipulation: on a trial-by-trial basis, was the probability of choosing a given mine predicted by its model-estimated instrumental value? We fit a multinomial logistic regression model (using the "mlogit" R package; Croissant, 2020) predicting mine choices (20%, 65%, or 80%) as a function of VPI-prediction and VPI-choice (with no intercept). The model also included interaction terms between task condition and each feature. As predicted, the interaction terms were significant: VPI-prediction was more strongly predictive of mine choices in the prediction condition compared to the choice condition, $b = 0.20$, 95% CI = [0.11, 0.29], $z = 4.15$, $p < .001$, while VPI-choice was less predictive of mine choices in the prediction condition compared to the choice condition, $b = -0.25$, 95% CI = [-0.38, -0.12], $z = -3.89$, $p < .001$. We repeated these

analyses with HV-prediction and HV-choice. Again, the interaction terms were significant: HV-prediction was more strongly predictive of mine choices in the prediction condition compared to the choice condition, $b = -0.22$, 95% CI = [-0.31, -0.13], $z = -4.64$, $p < .001$, while HV-choice was less predictive of mine choices in the prediction condition compared to the choice condition, $b = 0.35$, 95% CI = [0.24, 0.47], $z = 6.02$, $p < .001$. The AIC of the VPI model was 6047, while the AIC of the heuristic model was 6014, suggesting that heuristic features predicted participants' choices better than optimal features.

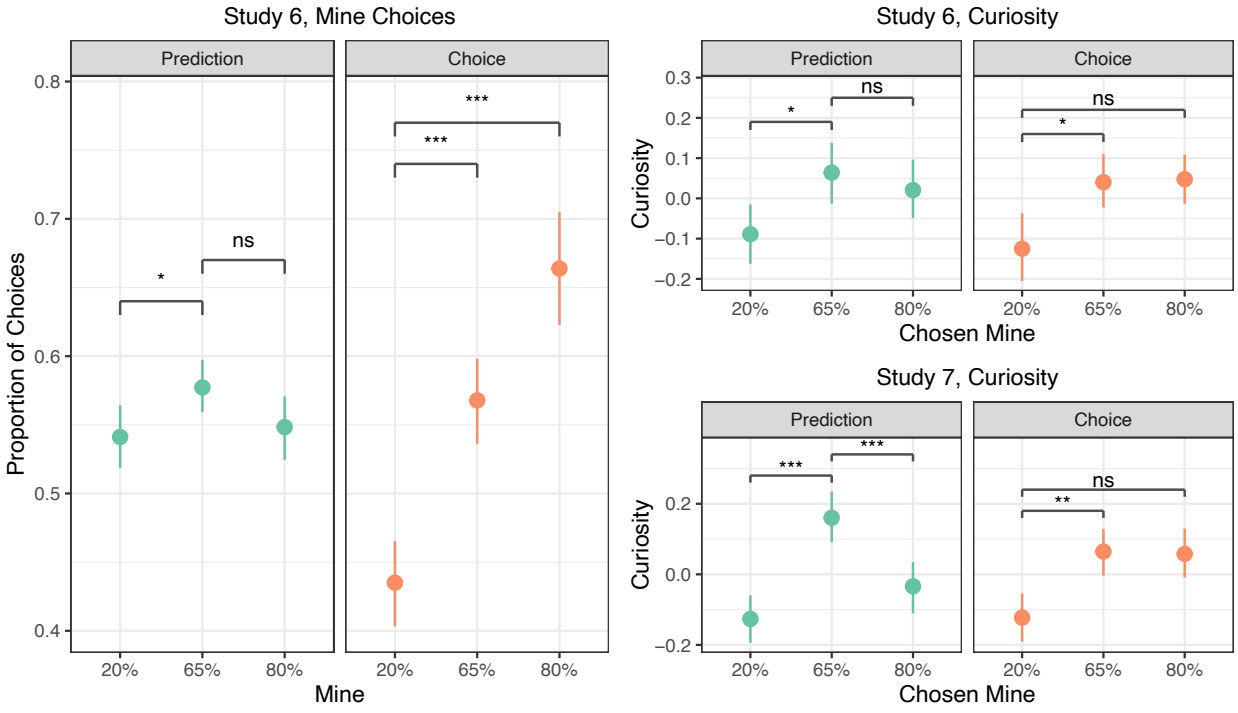


Figure 8 Instrumental Value, Choice, and Curiosity, Studies 6-7
Left: Study 6 mine choices during the observation phase, as a function of mine and task condition (prediction/choice). The significance of each contrast according to Wilcoxon-signed rank tests is indicated. **Right/top:** Study 6 curiosity ratings during the observation phase, as a function of mine and task condition (prediction/choice). The significance of each contrast according to t-tests with Satterthwaite's estimated degrees of freedom (within a mixed-effects regression model) is indicated. **Right/bottom:** Study 7 curiosity ratings during the observation phase, as a function of mine and task condition (prediction/choice). The significance of each contrast according to t-tests with Satterthwaite's estimated degrees of freedom (within a mixed-effects regression model) is indicated. *** p < .001; ** p < .01; * p < .05; ns = not significant.

Notably, these models do not account for the clustered nature of our data (observations nested within participants), so results should be taken as tentative. However, these results suggest that participants' exploratory choices are influenced by instrumental value: participants chose to explore mines that were more likely to provide information relevant for achieving their goals, both according to the true values of the mines and according to optimal and heuristic model estimates on a trial-by-trial basis.

Is curiosity affected by the instrumental value of information?

Next, we tested whether curiosity was similarly influenced by instrumental value, conducting similar analyses to those reported above. First, we tested whether participants were differentially curious about outcomes received as a function of the chosen mine (20%, 65%, 80%) and task condition. In a mixed-effects regression model, the interaction between mine and task condition was not significant, $\chi^2(2) = 0.41$, $p = .82$. We also investigated specific preregistered contrasts, predicting curiosity as a function of mine within each task condition. We estimated the significance of specific contrasts (e.g., between 20% and 65%) using the "lmerTest" package in R (Kuznetsova et al., 2017), which uses Satterthwaite's method to estimate degrees of freedom. We predicted that participants would be more curious about outcomes from the 65% mine and the 80% mine relative to the 20% mine in the choice condition, and we predicted that participants would be less curious about outcomes from the 20% mine and the 80% mine relative to the 65% mine in the prediction condition. In fact, participants were more curious about outcomes from the 65% mine relative to the 20% mine in both conditions, but the other contrasts were not significant (see Figure 8).

Though participants' curiosity did not appear to be affected by instrumental value at the level of mines, it remains possible that participants' curiosity differed across task conditions on a

trial-by-trial basis. However, this was not the case: in a model with both VOI features, the interaction between task condition and VOI-prediction was not significant, $\chi^2(1) = 0.74$, $p = .39$, nor was the interaction between task condition and VOI-choice, $\chi^2(1) = 0.08$, $p = .78$. We repeated these analyses with VPI and HV features. In a model with both VPI features, the interaction between task condition and VPI-prediction was not significant, $\chi^2(1) = 0.48$, $p = .49$, nor was the interaction between task condition and VPI-choice, $\chi^2(1) = 0.21$, $p = .65$. In addition, in a model with both heuristic instrumental features, the interaction between task condition and HV-prediction was not significant, $\chi^2(1) = 0.12$, $p = .73$, nor was the interaction between task condition and HV-choice, $\chi^2(1) = 0.43$, $p = .51$. Together, these results do not provide evidence that curiosity was influenced by participants' incentivized goal.

Does curiosity track optimal or heuristic epistemic features?

We next tested whether curiosity tracked epistemic (non-instrumental) features, replicating our analyses from Studies 4-5. As preregistered, we conducted this analysis separately within each task condition. In individual regression models, all epistemic features were significant predictors of curiosity in both the prediction condition and the choice condition (ps < .01; see Table 3), and the estimated regression coefficients were comparable to those in Studies 4-5 (see Figure 5). In particular, query entropy/EIG ($\beta_{prediction} = 0.31$, $\beta_{choice} = 0.42$) and global EIG ($\beta_{prediction} = 0.29$, $\beta_{choice} = 0.37$) were the strongest predictors of curiosity, while uRPE ($\beta_{prediction} = 0.16$, $\beta_{choice} = 0.19$) and global entropy ($\beta_{prediction} = 0.10$, $\beta_{choice} = 0.07$) were somewhat weaker. Next, we tested which features explained unique variance in curiosity by fitting a multiple regression model with all four features as fixed effects, as again, many features were modestly correlated (see Figure 7). Only query entropy/EIG ($\beta = 0.23$) explained unique variance in curiosity in the prediction condition ($p < .001$; see Table 4). In the choice condition,

uRPE (β = 0.17), query entropy/EIG (β = 0.44), and global entropy (β = -0.08) explained unique variance in curiosity (ps < .01; see Table 4). The estimated regression coefficients appeared to differ somewhat from those in Studies 4-5 (see Figure 5): for example, global entropy was negatively associated with curiosity in the choice condition, while it was positively related to curiosity in Studies 4-5. However, query entropy/EIG was most strongly related to curiosity across both task conditions, as it was in Studies 4-5.

Does curiosity track instrumental features (above and beyond epistemic features)?

Though the instrumental features were not differentially predictive of curiosity as a function of task condition (see above), it is possible that they nonetheless predict curiosity within each condition. First, we tested whether each instrumental feature predicted curiosity in isolation (considering only condition-appropriate features, e.g., only prediction features in the prediction condition). VOI-prediction (β = 0.11), VPI-prediction (β = 0.10), and HV-prediction (β = -0.09) were significantly associated with curiosity in the prediction condition, and VOI-choice (β = 0.09) was significantly associated with curiosity in the choice condition (ps < .01; see Table 3). The estimated regression coefficients were generally small, especially compared to query entropy/EIG and global EIG (see Figure 5).

In addition, we tested whether the instrumental features explained additional variance in curiosity above and beyond the epistemic features. We fit an additional multiple regression model for each instrumental feature, including fixed effects for all epistemic features and for one instrumental feature. We tested the unique variance explained by the added instrumental feature by comparing the full model to a reduced model excluding the relevant fixed effect. Only VOI-prediction explained unique variance in curiosity beyond the epistemic features (in the prediction

condition), but this effect was quite small ($\beta = 0.04$). In the choice condition, none of the instrumental features were significant (see Figure 9).

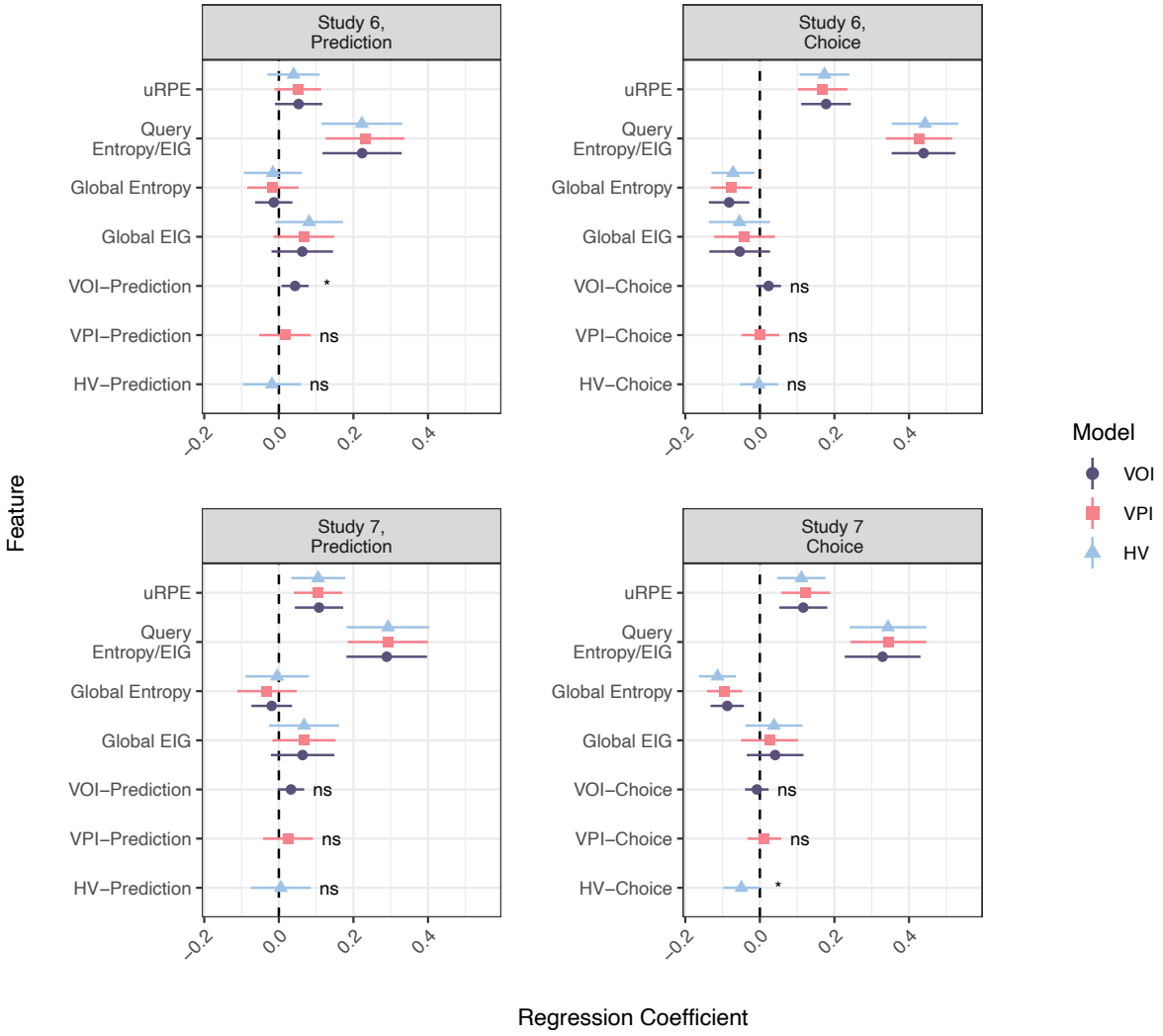


Figure 9 Added Variance Explained by Instrumental Features, Studies 6-7
Results of regression models assessing the additional variance explained by each instrumental feature in Study 6 (top) and Study 7 (bottom). Each model included all four epistemic features (uRPE, query entropy/EIG, global entropy, and global EIG) and one instrumental feature (VOI-prediction/choice, VPI-prediction/choice, or HV-prediction/choice). The significance of each instrumental feature is indicated. *** p < .001; ** p < .01; * p < .05; ns = not significant.

3.4.3 Discussion

In Study 6, we extended the paradigm developed in Studies 4-5 to also test whether curiosity was sensitive to instrumental value. Validating our manipulation, participants' choices between the mines shifted as a function of their incentivized task condition. However, there was little evidence that curiosity was influenced by instrumental value, and curiosity was only weakly related to instrumental features.

As we acknowledged in introducing Study 6, allowing participants to choose between the mines may have shifted participants' cognitive resources away from curiosity ratings. We therefore attempt to replicate our results in Study 7, modifying the method so that participants are only responsible for rating curiosity (and not choosing between the mines) during the observation phase. In addition, we attempt to replicate our results concerning the association between epistemic features and curiosity, which were qualitatively different between Studies 4-5 and Study 6. In particular, while query entropy/EIG was the strongest predictor of curiosity across conditions and studies, the role of uRPE, global entropy, and global EIG appeared to be more variable. We test these associations again in Study 7, and we return to a discussion of these cross-study differences in light of the Study 7 results.

3.5 Study 7

In Study 7, we again test whether instrumental value influences curiosity, in a task in which participants do not also choose between the mines. We also attempt to replicate our results regarding the association between epistemic features and curiosity.

3.5.1 Method

Participants

One hundred fifty-one participants (ages 18 to 66; 84 women, 64 men; two non-binary, and one unspecified) were recruited from Prolific. To participate, participants were required to reside in the United States and to have completed at least 100 previous studies on Prolific with a minimum 95% approval rate. An additional nine participants were excluded for failing to answer two multiple-choice attention check questions, and an additional four participants were excluded for providing the same curiosity rating on all trials.

Procedure

The method was identical to that of Study 6 (see Figure 4), with the following exception: during the observation phase, participants were instructed that they would be observing their miners dig in the mines, rather than digging themselves. Thus, in the observation phase, after the first six fixed-choice trials (two per mine), there were nine additional fixed-choice trials (three per mine). Eighty participants were randomly assigned to the choice condition and 71 were randomly assigned to the prediction condition. In addition, 47 were randomly assigned to the copper bandit, 52 were randomly assigned to the copper tycoon, and 52 were randomly assigned to the sheriff.

3.5.2 Results

As in Study 6, we present analyses addressing the following questions: (1) Is curiosity affected by the instrumental value of information? (2) Does curiosity track optimal or heuristic epistemic features? And (3) Does curiosity track instrumental features (above and beyond epistemic features)? In addition, we quantitatively compare Studies 4-5 and Studies 6-7.

We use the same analytic approach as described in Studies 4-6. We again z-score curiosity ratings within participants and z-score all model-estimated features across the full dataset. Study 7 was preregistered (https://osf.io/z6p5g/?view_only=540fecfdc8c747c48d75f4277f9efc0d), but we deviate from our preregistered analysis plan as described in Study 6.

Is curiosity affected by the instrumental value of information?

First, we tested whether participants were differentially curious about outcomes received from the three mines as a function of condition, repeating the analyses reported in Study 6. We initially tested whether participants' curiosity about the three mines differed according to task condition. Unlike Study 6, the interaction between task condition and chosen mine was significant, $\chi^2(2) = 6.20$, $p = 0.045$. To follow up on this interaction, we investigated whether participants were differentially curious about outcomes from the three mines within each task condition (see Figure 8). As predicted, participants were less curious about outcomes received from the 20% mine and the 80% than the 65% mine in the prediction condition, and participants were more curious about outcomes received from the 65% mine than the 20% mine in the choice condition. However, the predicted difference between the 80% mine and the 20% mine in the choice condition was not significant.

In addition, we tested whether curiosity was sensitive to instrumental value when using trial-by-trial estimates of VOI, VPI and HV. First, we tested whether curiosity was associated with VOI-prediction and VOI-choice differentially across conditions. The interaction between task condition and VOI-prediction was not significant, $\chi^2(1) = 0.49$, $p = .48$, nor was the interaction between task condition and VOI-choice, $\chi^2(1) = 1.13$, $p = .29$. We repeated this analysis for VPI-prediction and VPI-choice. Similarly, the interaction between task condition and VPI-prediction was not significant, $\chi^2(1) = 1.77$, $p = .18$, nor was the interaction between

task condition and VPI-choice, $\chi^2(1) = 0.004$, $p = .95$. Finally, we tested whether curiosity was associated with HV-prediction and HV-choice differentially across conditions. The interaction between task condition and HV-prediction was significant, $\chi^2(1) = 5.41$, $p = .02$, but the interaction between task condition and HV-choice was not, $\chi^2(1) = 2.14$, $p = .14$.

To follow up on the significant interaction, we fit a model including HV-choice and HV-prediction as fixed effects within each task condition. In the prediction condition, HV-prediction was related to curiosity, $\beta = -0.16$, 95% CI [-0.21 -0.12], $\chi^2(1) = 35.98$, $p < 0.001$, and HV-choice did not explain additional variance in curiosity, $\beta = 0.06$, 95% CI [-0.004, 0.12], $\chi^2(1) = 3.41$, $p = 0.06$. In the choice condition, HV-prediction was weakly related to curiosity, $\beta = -0.08$, 95% CI [-0.13, -0.03], $\chi^2(1) = 9.59$, $p = 0.002$, while HV-choice did not explain additional unique variance in curiosity, $\beta = -0.002$, 95% CI [-0.06, 0.06], $\chi^2(1) = 0.002$, $p = 0.96$. In sum, HV-prediction explained variance in curiosity in both conditions, but it was more strongly related to curiosity in the prediction condition than the choice condition. When controlling for HV-prediction, HV-choice did not account for variance in curiosity in either condition.

Together, these results provide weak evidence that curiosity is sensitive to instrumental value: participants were selectively curious about outcomes from mines with moderate reward probability in the prediction condition and with close high-reward probability competitors in the choice condition. However, on a trial-by-trial basis, evidence for condition-specific sensitivity to instrumental value was weak—only HV-prediction was differentially predictive of curiosity in the prediction condition relative to the choice condition.

Does curiosity track optimal or heuristic epistemic features?

We next sought to replicate our analyses regarding the association between curiosity and epistemic features. As in Study 6, we conducted this analysis separately within each task

condition. In individual regression models, each epistemic feature was significantly related to curiosity in both conditions (ps < .001; see Table 3, Figure 5), replicating Studies 4-6. As in the previous studies, query entropy/EIG was the strongest predictor of curiosity ($\beta_{prediction}$ = 0.40, β_{choice} = 0.38), followed by global EIG ($\beta_{prediction}$ = 0.39, β_{choice} = 0.36), uRPE ($\beta_{prediction}$ = 0.23, β_{choice} = 0.20), and global entropy ($\beta_{prediction}$ = 0.14, β_{choice} = 0.09). In multiple regression models (estimating the unique variance explained by each model-estimated feature), the results were largely consistent with those of Study 6. In the prediction condition, only uRPE (β = 0.10) and query entropy/EIG (β = 0.29) explained unique variance in curiosity, while in the choice condition, uRPE (β = 0.12), query entropy/EIG (β = 0.33), and global entropy (β = -0.09) explained unique variance in curiosity (ps < .01; see Table 4, Figure 5). Like Studies 4-6, query entropy/EIG was the strongest predictor of curiosity across both task conditions.

Does curiosity track instrumental features (above and beyond epistemic features)?

Finally, we replicated our Study 6 analyses testing whether task-specific instrumental features were associated with curiosity within each condition. Like Study 6, VOI-prediction (β = 0.12), VPI-prediction (β = 0.14), and HV-prediction (β = -0.14) were significantly associated with curiosity in the prediction condition, while VOI-choice (β = 0.05) was significantly associated with curiosity in the choice condition (ps < .01; see Table 3, Figure 5). We additionally tested whether the instrumental features explained additional variance in curiosity beyond the epistemic features (see Figure 9). Unlike Study 6, only HV-choice (β = -0.05) explained unique variance in curiosity beyond the epistemic features (in the choice condition). In the prediction condition, no instrumental features explained unique variance in curiosity.

Do associations between curiosity and epistemic features differ between Studies 4-5 and Studies 6-7?

The epistemic correlates of curiosity appeared similar in Studies 6 and 7, but they appeared to differ slightly compared to the results of Studies 4 and 5 (see Figure 5). We quantitatively compared the two sets of studies by fitting a multiple regression model with fixed effects for all four non-instrumental features, study set (Studies 4-5 vs. Studies 6-7), and two-way interactions between each feature and study set. We compared this model to a set of reduced models, which each excluded one interaction term. This exploratory analysis was conducted separately for each task condition in Studies 6-7. Comparing Studies 4-5 and the prediction condition of Studies 6-7, there was evidence for an interaction between study set and global entropy, $\chi^2(1) = 9.48$, $p = .002$, but not uRPE, $\chi^2(1) = 0.75$, $p = .39$, query entropy/EIG, $\chi^2(1) = 0.83$, $p = .36$, or global EIG, $\chi^2(1) = 1.81$, $p = .18$. Participants in the prediction condition of Studies 6-7 tracked global entropy less closely than participants in Studies 4-5 (see Figure 5). Comparing Studies 4-5 and the choice condition of Studies 6-7, there was evidence for an interaction between study set and query entropy/EIG, $\chi^2(1) = 13.27$, $p < .001$, global entropy, $\chi^2(1) = 33.50$, $p < .001$, and global EIG, $\chi^2(1) = 10.70$, $p = .001$, but not uRPE, $\chi^2(1) = 1.02$, $p = .31$. Participants in the choice condition of Studies 6-7 tracked query entropy/EIG more closely and global EIG less closely than participants in Studies 4-5, while global entropy flipped sign in its association with curiosity (from positive in Studies 4-5 to negative in Studies 6-7; see Figure 5).

3.5.3 Discussion

In Study 7, we tested whether curiosity was sensitive to the instrumental value of information, in a task variant where participants' sole task (during the observation phase) was

rating curiosity. There was weak evidence that curiosity was influenced by instrumental value: participants were differentially curious about the three mines dependent upon task condition, HV-prediction predicted curiosity more strongly in the prediction condition than the choice condition, and HV-choice was associated with curiosity in the choice condition when controlling for all other epistemic features. However, in several other analyses, there was no evidence for sensitivity to the instrumental value of information: for example, there was no evidence for a significant interaction between task condition and VOI-prediction, VOI-choice, VPI-prediction, VPI-choice, or HV-choice. In addition, most of the variance in curiosity was accounted for by epistemic features, especially query entropy/EIG. Thus, while instrumental value may impact curiosity, this effect appears to be small and fragile.

We found qualitative similarity in the epistemic correlates of curiosity between Study 6 and Study 7. However, this makes the differences between Studies 4-5 and Studies 6-7 all the more puzzling: for example, global entropy was positively related to curiosity in Studies 4-5 and negatively related to curiosity in the choice condition of Studies 6-7. One possibility is that the presence of a test condition or the presence of three mines (rather than two) pulled participants' curiosity towards different features. Another possibility is that curiosity is related to different features over the course of learning. In Studies 4-5, participants had 50 trials to learn about two mines, while in Studies 6-7, participants had only 15 trials to learn about three mines. As a result, participants in Studies 6-7 had far less information on average about the mines. In fact, when we restrict the multiple regression analysis in Study 5 to the set of consecutive trials for each participant in which each mine was chosen no more than five times (equating the amount of evidence to that received by participants in Studies 6-7), only query entropy/EIG, $\beta = 0.49$, 95% CI [0.33, 0.64], $\chi^2(1) = 37.66$, $p < .001$, and uRPE, $\beta = 0.12$, 95% CI [0.01, 0.24], $\chi^2(1) = 4.42$, p

= .04, explained unique variance in curiosity, resembling the results in the prediction condition of Studies 6-7. This analysis was exploratory, but hints at one possible explanation for the conflicting results between Studies 4-5 and Studies 6-7.

3.6 General Discussion

In the present research, we investigated what cues guide the selectivity of curiosity during learning, contrasting heuristic epistemic features, epistemically optimal features, heuristic instrumental features, and instrumentally optimal features (see Table 2). To the extent that curiosity tracks epistemically and instrumentally optimal features, learners will desire information precisely when the information is most likely to lead to useful learning. However, heuristic cues like surprise and uncertainty might approximate expected useful learning reasonably well, allowing a learner to track simpler cues yet still achieve epistemic and instrumental goals much of the time. To test when curiosity is in fact elicited, we asked three questions: Does curiosity track epistemically optimal features or heuristic epistemic alternatives? Does curiosity track uncertainty and expected learning about the question of interest, or about the ultimate learning goal? And does curiosity track instrumentally optimal features or heuristic instrumental alternatives?

To answer these questions, we developed a paradigm and computational model that disentangle heuristic and optimal triggers for epistemic and instrumental goals. In particular, we operationalized surprise, uncertainty, expected learning, optimal instrumental value, and heuristic instrumental value in the context of a Bayesian reinforcement learning model, using mathematical definitions of these features that have been used in prior research on learning, exploration, and decision making. In addition, we estimated uncertainty and expected learning about two components of the task: the question being asked and the ultimate learning goal.

Having operationalized these features, we tested their association with participants' reported curiosity during a learning task. This is the first empirical investigation that has simultaneously investigated this breadth of possible triggers of curiosity using a computational approach.

In Studies 4-5, we contrasted heuristic epistemic features with epistemically optimal features. In Study 4, curiosity was most strongly predicted by query entropy/EIG (which conflates heuristic and epistemically optimal features). However, uRPE and global entropy (heuristic features) also explained unique variance in curiosity when controlling for all features in a single model. In Study 5, we modified the method so that participants' only task was to rate their curiosity, with no other specified goal. In this study, curiosity was again most strongly predicted by query entropy/EIG, but all three additional epistemic features—uRPE, global entropy, and global EIG—explained unique variance in curiosity. On the basis of these results, we suggested that curiosity might have the potential to track global epistemically optimal features, but this appears to be fragile: only when participants were solely responsible for reporting curiosity did curiosity track global expected learning.

In Studies 6-7, we contrasted all four sets of features (heuristic epistemic features, epistemically optimal features, heuristic instrumental features, and instrumentally optimal features), by incentivizing participants to complete either a prediction task or a choice task. In Study 6, we found little evidence that curiosity tracked instrumental features, but we validated our task manipulation using an independent behavioral measure. In Study 7, we modified the task so that participants were solely responsible for reporting curiosity (like Study 5). We found weak evidence that curiosity was sensitive to heuristic instrumental triggers: HV-prediction was more strongly related to curiosity in the prediction condition than the choice condition, and HV-choice explained unique variance in curiosity above and beyond the epistemic features

investigated in Studies 4-5. However, in both the choice and prediction conditions, query entropy/EIG was again the strongest predictor of curiosity.

3.6.1 Implications and open questions

Below, we discuss answers to our motivating questions: Does curiosity track epistemically optimal features or heuristic epistemic alternatives? Does curiosity track uncertainty and expected learning about the question of interest, or about the ultimate learning goal? And does curiosity track instrumentally optimal features or heuristic instrumental alternatives? In doing so, we highlight implications and open questions.

Questions 1 and 2: Epistemically optimal or heuristic; query or global?

In response to our first question (does curiosity track epistemically optimal features or heuristic epistemic alternatives?), we cannot completely adjudicate between heuristic triggers and optimal triggers. Across all studies, the strongest predictor of curiosity was query entropy/EIG, considered both in isolation and controlling for other features. Unfortunately, this is the sole feature that confounds heuristic and optimal triggers, as the amount of uncertainty that exists about whether or not an intervention occurred on a given trial is equivalent to the amount of uncertainty that would be resolved if that question were to be answered. One open question for future research is whether the heuristic or optimal component of query entropy/EIG best explains variance in curiosity. This might be tested by manipulating the reliability of information about the intervention on each trial (i.e., how likely it is to be accurate; see Markant & Gureckis, 2014a). If query EIG shapes curiosity more than query entropy, individuals should be less curious when the information they expect to receive will be unreliable rather than reliable, even if they are quite uncertain. Prior work suggests that information search is sensitive to the reliability of expected information: even young children prefer to seek information from reliable

informants over unreliable informants (e.g., Corriveau et al., 2009; Corriveau & Harris, 2009; Corriveau & Kurkul, 2014; Koenig & Harris, 2005; Mills et al., 2010). However, it remains to be seen whether curiosity is sensitive to this cue to expected learning.

In response to our second question (does curiosity track uncertainty and/or expected learning about the question of interest, or about the ultimate learning goal?), we have a much clearer answer. Across all four studies, query-level uncertainty/expected learning was the strongest predictor of curiosity in both individual and multiple regression analyses. While global expected learning was also a reasonably strong predictor of curiosity in individual regression analyses, neither global uncertainty nor global expected learning explained unique variance in curiosity across all studies and conditions. These findings are consistent with prior research on information search that has disentangled query-level and global features (Markant et al., 2015; Markant & Gureckis, 2014a). However, it is possible that our task underestimates the importance of global features as cues to curiosity. In our task, local information (whether or not an intervention occurred) is not necessary for learning global information (the reward probability of the mines): it is possible to infer each mine's reward probability by merely observing the rewards associated with each mine. As discussed in Section 3.1, this divorces curiosity from the desire to perform the task well, and thus provides a stringent test of the triggers of curiosity. However, we might expect curiosity to be more strongly associated with global features when local information is a more important source of global information. One such situation is in fact represented in the present research: in a shorter task (or in early trials of a longer task), global expected learning is much higher than in a longer task (particularly during late trials). Put simply, when little is known about the global target of inquiry, any relevant local information will have a fairly large impact on one's global knowledge. However, in the shorter tasks used in

Studies 6-7, global EIG was not a stronger predictor of curiosity than in the longer tasks used in Studies 4-5. In addition, global EIG did not explain unique variance in curiosity during the early trials of Study 5 (see Section 3.5.3), though it did across the full task (see Section 3.3.2). While this provides preliminary evidence that curiosity is not more closely related to global features in settings where local information is more important for learning global information, future research might further evaluate this possibility.

Our results on the epistemic triggers of curiosity raise several additional questions for future research. We found across two pairs of studies that curiosity appeared to be less optimal when participants were responsible for choosing between the mines in addition to rating curiosity (failing to track global EIG in Study 4 and failing to track instrumental value in Study 6). One possibility advanced in the Section 3.2.3 is that optimally triggered curiosity requires a sufficient amount of attention and cognitive resources. If this is the case, we would expect that curiosity would track expected learning and/or expected instrumental value to the extent that the learning task was simple, while curiosity would track heuristic features to the extent that the learning task was complex. In addition, we might expect individual differences in the triggers of curiosity, so that individuals with a tendency towards more reflective deliberation (e.g., Toplak et al., 2011) would be more optimal in their curiosity. This is consistent with prior research (Coenen et al., 2015), in which participants sought information in a way that was less consistent with optimal expected learning when under time pressure. However, as we did not design the present research to specifically test these predictions, it remains to be tested whether the triggers of curiosity shift depending upon task complexity or individual characteristics.

In addition, because there were many differences between Studies 4-5 and Studies 6-7, it is unclear why the epistemic correlates of curiosity differed across these sets of studies. In the

Discussion to Study 7, we proposed one possible explanation for these results: participants in Studies 6-7 received evidence covering a different distribution of surprise, uncertainty, and expected learning because they made fewer observations per mine. This could impact the results in two ways: first, the triggers of curiosity might be non-linearly related to curiosity (e.g., global EIG might be more predictive of curiosity when it is low or moderate rather than high), or second, the triggers of curiosity might interact with each other (e.g., global EIG might only be predictive of curiosity when global uncertainty is low). Consistent with the first possibility, curiosity literature has theorized an "inverted u-shaped curve" between curiosity and surprise, uncertainty, or expected learning (Kang et al., 2009; Kidd et al., 2012; Loewenstein, 1994), such that curiosity is highest when surprise, uncertainty, or expected learning are moderate, and lower when surprise, uncertainty, or expected learning are low or high. We explored non-linear associations through the GAMM analyses in Studies 4-5, and trends appeared to be reasonably well-approximated by linear regression lines (though query entropy/EIG was more strongly related to curiosity at high levels). Given that we did not find "inverted u-shaped curves" in the present research, future research might investigate the conditions under which there are non-linear associations between curiosity and epistemic triggers. In addition, further research is needed to determine whether the triggers of curiosity interact with each other.

Question 3: Instrumentally optimal or heuristic?

Finally, in response to our third question (does curiosity track instrumentally optimal features or heuristic instrumental alternatives?), we find mixed results. At a coarse level, participants were on average more curious about outcomes from mines that were more valuable for the condition-specific incentivized task (e.g., participants were more curious about the 65% mine relative to the 20% or 80% mine in the prediction condition), but only in Study 7. On a

more fine-grained level, however, curiosity did not adjust to track trial-by-trial estimates of optimal instrumental value in Study 6 or 7, while evidence for heuristic instrumental triggers was only present in Study 7. Together, these results suggest that curiosity sometimes shifts to track the instrumental value of information, but potentially only in crude ways or according to heuristic cues. Notably, the prior work that has found an effect of instrumental value on information search (Abir et al., 2020; Dubey & Griffiths, 2020) has operationalized instrumental value at a fairly coarse level: for example, people are more likely to seek information about questions that they expect to be quizzed on later (Dubey & Griffiths, 2020), and people are more likely to seek information about COVID-19 to the extent that they are more concerned about the pandemic (Abir et al., 2020). In contrast, the prior work that has failed to find an effect of instrumental value on information search (Bramley et al., 2015; Markant & Gureckis, 2012; Meder & Nelson, 2012) has operationalized instrumental value on a fine-grained, trial-by-trial basis, where each piece of information contributes to an ultimate learning goal. Combining this evidence with our findings, it is possible that curiosity might only track instrumental value at a certain level of granularity—however, this proposal requires further investigation.

While many questions remain open, the present research provides new insight into the cues that might trigger curiosity. Specifically, curiosity co-occurs with query uncertainty or query expected learning, but evidence was mixed for an association between curiosity and surprise, global uncertainty, global expected learning, or instrumental value.

3.6.2 Limitations

Several limitations of this work must be highlighted. First, it was not possible to manipulate the epistemic features independently because, while the correlations between these variables were moderate, the variables are non-linearly related. Thus, we cannot make strong

claims about the causal influence of any single epistemic variable. It also remains an open question whether and when these potential triggers curiosity come apart in real-world contexts. The task used in the present research has features of many naturalistic information-seeking contexts. For example, observations could be caused by several hidden causal forces (e.g., a plant wilting can be caused by too much sun or by not enough water) and inquiry increased knowledge incrementally (e.g., checking my basil plant each day provides only a bit of information about whether I am a good gardener). However, our method is nonetheless artificial, and it would be beneficial to disentangle the triggers of curiosity in a real-world learning context.

In addition, we have not considered all possible precursors of curiosity in the present research. For example, researchers have advanced several definitions of surprise (for a review, see Munnich et al., 2019): surprise could be determined by the presence of a low probability event (Reisenzein, 2000), the extent to which one has difficulty explaining an event (Foster & Keane, 2015; R. Maguire et al., 2011), or by the extent to which one sees patterns where they expect noise (P. Maguire et al., 2018). Likewise, EIG is only one possible operationalization of optimal expected learning (see Coenen et al., 2019; Nelson, 2005). Future research is needed to fully explore alternative operationalizations of these constructs. Finally, we have not captured all definitions of curiosity. While we focus on state curiosity in response to external stimuli, other research has explored individual differences in trait curiosity (e.g., Litman, 2008). It remains an open question whether participants in our task vary in trait-level curiosity, and whether this variation affects the determinants of state-level curiosity.

3.6.3 Conclusion

Despite these limitations and open questions, the present research introduces a novel paradigm for studying curiosity and a computational model that allows surprise, query/global

uncertainty, query/global expected learning, and instrumental value to be partially disentangled. Across four studies, curiosity was most strongly associated with query uncertainty/expected learning, such that curiosity was higher when uncertainty and expected learning about the question of interest was also higher. Surprise, global uncertainty, and global expected learning explained variance beyond query uncertainty/expected learning in some studies but not others. Likewise, the instrumental value of information had a weak and variable influence on curiosity. Together, this research provides new insight into the cues that shape our experience of curiosity, opening many questions that promise to deepen our understanding of when and why humans are motivated to seek information.

4 CHAPTER 4: IN SEARCH OF SATISFYING EXPLANATIONS‡

4.1 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 at last: 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 B.C. 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

‡ A version of this chapter is currently under review. Liquin, E. G. & Lombrozo, T. (under review). In search of satisfying explanations. In addition, parts of this work are published Liquin, E. G., & Lombrozo, T. (2019). Inquiry, theory-formation, and the phenomenology of explanation. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 664–670). Cognitive Science Society. Components of this work were also presented in the following: Liquin, E. G., & Lombrozo, T. (2021, June). Explanatory satisfaction tracks epistemic success and motivates inquiry. Talk presented at the annual meeting of the Society for Philosophy and Psychology, virtual. Liquin, E. G. & Lombrozo, T. (2019, July). Inquiry, theory-formation, and the phenomenology of explanation. Talk presented at the annual meeting of the Cognitive Science Society, Montreal, QB.

behaviors that support reproduction, so explanatory satisfaction motivates humans to engage in inquiry that supports learning about the causal structure of the world.

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, 2020). For instance, in Chapter 2, we 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 four experiments we go on to report.

4.1.1 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 useful and new that is relevant to the posed query. The most direct test of this prediction—what we call the *actual learning prediction*— would be to link the experience of explanatory satisfaction to some objective measure of query-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 objective learning (Glenberg & Epstein, 1985; Lin & Zabrucky, 1998; Maki, 1998). Researchers have argued that judgments of one's own 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). If judgments of one's own learning 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 objective 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.

4.1.2 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.

Perceptions of useful learning could be based on direct cues (e.g., how much information an explanation is perceived to contain), or on indirect cues, such as an explanation's simplicity and breadth. As argued in prior work, properties such as simplicity and breadth could indicate that an explanation is likely to support generalization beyond the specific case being explained (T. Blanchard, Vasilyeva, et al., 2018; Friedman, 1974; Kitcher, 1989; Lombrozo, 2016; Strevens, 2004; Thagard, 1978), and they could be valued in explanations precisely because they support effective inquiry, leading learners to seek and identify the generalizable or "portable" features of the world (Lombrozo, 2011, 2016; Lombrozo & Carey, 2006).

Prior work offers mixed support for the perceived learning prediction. 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. This suggests that if direct measures of perceived learning predict explanatory satisfaction, the relationship is not strong. The evidence for indirect cues is more promising: 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 as well (T. Blanchard, Lombrozo, et al., 2018; Lombrozo, 2007; Pacer & Lombrozo, 2017; but see Johnson et al., 2019; Zemla et al., 2017, 2020; Lim & Oppenheimer, 2020). In light of this prior work, it is unclear whether to expect a robust association between explanatory satisfaction and properties such as perceived learning, utility, simplicity, or breadth.

4.1.3 Aligned motivation: Selective reinforcement

A final prediction of the aligned motivation account is that receiving a satisfying explanation should motivate 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, 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 increase 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*: a satisfying explanation should (a) depress curiosity concerning the original query, (b) increase curiosity about the answers to related questions, and (c) have no effect on curiosity about the answers to unrelated questions. 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 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.

4.1.4 Overview of Experiments

In the present research, we test the aligned motivation account of explanatory satisfaction. The aligned motivation account predicts that satisfaction is related to 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 foster 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 5).

Table 5 Predictions of Aligned and Brute Motivation Accounts, Studies 8-11

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

Prediction	Aligned Motivation Account	Brute Motivation Account	Evidence
Actual Learning Prediction	Explanations should be more satisfying to the extent that more is learned from them.	No connection between satisfaction and actual learning.	Studies 9-11: little evidence for a connection between satisfaction and actual learning (free recall/multiple-choice performance).
Perceived Learning Prediction	Explanations should be more satisfying to the extent that one believes the explanation supports learning (direct) and/or has explanatory virtues (indirect).	No connection between satisfaction and perceived learning.	Studies 8, 9, 11: evidence for association between satisfaction and direct/indirect cues to learning; direct cues explain preference for reductive explanations.
Selective Reinforcement Prediction	Satisfaction increases the perceived value of subsequent inquiry on related questions, decreases the perceived value of subsequent inquiry on the initial target question, and does not change the 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 8-9: 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.

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 9-11), the perceived learning prediction (Studies 8, 9, and 11), and the selective reinforcement prediction (Studies 8-9). 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. In addition, we find some evidence for the selective reinforcement prediction: receiving a satisfying 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 11, 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 provide 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 inquiry, guiding learners to seek explanations and—sometimes—to achieve a better understanding of the world.

4.2 Study 8

In Study 8, 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 both direct measures of learning (e.g., "To what extent has the answer to this question taught you

something new?"), and indirect measures in the form of explanatory features that could signal generalizability, such as simplicity and breadth (e.g., "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)?"). All measures are included in Table 6.

Our indirect measures of learning include one measure of simplicity, one measure of breadth, and one measure of the extent to which the explanation identifies a general pattern or regularity (referred to as "regularity"). In addition, we include a measure of expertise. 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; 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). This selective reliance on experts provides a potential means for people to judge whether an explanation has epistemic merit: even if an individual doesn't have the domain knowledge to evaluate an explanation themselves, the knowledge that the explanation was written or endorsed by experts may provide indirect evidence for the explanation's quality. Moreover, 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).

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.

Table 6 Items, Studies 8-9

Items rated (each on a seven-point scale) in response to explanations in Studies 8-9

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)
Direct measures of (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)
Indirect measures of (useful) learning	
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)

4.2.1 Method

Participants

Participants were 184 adults (106 male and 78 female, ages 21-69) recruited from Amazon Mechanical Turk (MTurk). Sixteen additional participants completed Study 8 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 at https://osf.io/gyuqa/ (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 1,000 prior tasks with a minimum approval rating of 99%, and participation was restricted so that participants could not participate in more than one study. All

participants in Studies 8-11 provided informed consent under a protocol approved by the Institutional Review Board at Princeton University.

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 MTurk participants 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 at https://osf.io/gyuqa/ 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/gyuqa/.

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, be more accurate in their expectations about epistemic success, 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 (2020), 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 at https://osf.io/gyuqa/.[17]

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 arithmetic 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

[17] In the interest of brevity, we omit a full analysis of calibration in Studies 8-9, 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).

subsequent problems. Those who did not correctly respond to at least five items were excluded from all analyses.[18]

In the answer phase, participants saw the same questions 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?") and seven items directly and indirectly assessing perceived learning (see Table 6), 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.

4.2.2 Results

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

[18] 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.

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 using a likelihood ratio test.

Testing the perceived learning prediction

First, we investigate the association between explanatory satisfaction and perceived learning, measured both directly and indirectly (Figure 10). Ratings on each learning measure were entered into a multiple regression model predicting explanatory satisfaction. We also included topic-level ratings of interest and knowledge as fixed effects to control for these possible confounds. Together, the learning 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 perceived learning—most notably, direct measures of how much has been learned, how much information an explanation contains, and how useful that information is, as well as indirect measures tracking the explanation's generality and perceptions of expertise. Furthermore, as we report in the supplementary material at https://osf.io/gyuqa/ and as depicted in Figure 10, all learning measures were related to explanatory satisfaction in isolation.

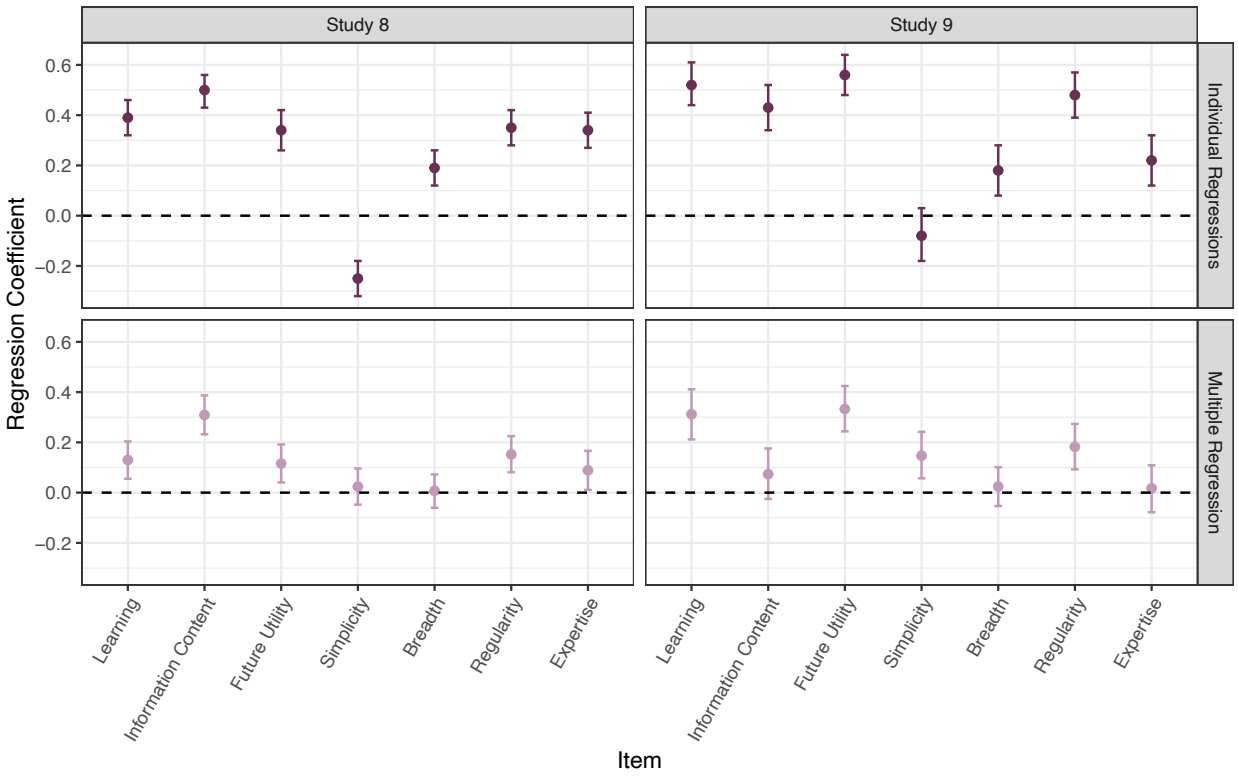


Figure 10 Explanatory Satisfaction and Perceived Learning, Studies 8-9
Association between each measure of learning and rated explanatory satisfaction, in Studies 8 and 9. For individual regressions (top panels), each learning measure was entered into a separate regression model predicting satisfaction, with random intercepts for participant (in Study 8) and question-answer pair (in Studies 8-9). For multiple regressions, each learning measure was entered as a fixed effect in a simultaneous regression model, with random intercepts for participant (in Study 8) and question-answer pair (in Studies 8-9). Topic-level interest and knowledge were included as control measures in both individual and multiple regressions.

Testing the selective reinforcement prediction

We next tested whether explanatory satisfaction motivates further inquiry on related questions, which we assessed through 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 this 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], $\chi^2(1) = 48.38$, $p < .001$.

4.2.3 Discussion

In Study 8, we find 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 learning—not only how much was learned from the explanation, but also how much information it was judged to contain, how useful it was, and how much it appealed to a general pattern or regularity. In addition, the perceived learning measures that did not explain unique variance in satisfaction were nonetheless related to satisfaction when considered in isolation.[19] Second, consistent with the selective reinforcement 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?".

While these results are consistent with the aligned motivation account, 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). Does explanatory satisfaction track actual learning of explanatory information, or only *perceived* learning? To address this question, the remaining studies include objective

[19] 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, 2020; 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 (T. 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 learnable information contained in an explanation than the causal structure of that explanation.

measures 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 9.

4.3 Study 9

In Study 9, we attempt to replicate evidence for the perceived learning prediction, and 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? In addition, we explore whether satisfaction (to the extent that it is related to actual learning at all) is specifically related to explanatorily relevant learning, as opposed to learning of any content contained within an answer to a question. If explanatory satisfaction is tuned to support and motivate learning through explanatory inquiry, satisfaction should track whether information has been learned that answers the explanation-seeking question.

Additionally, we followed up on the results of Study 8 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 9 was preregistered (see https://osf.io/gyuqa/); any departures from our preregistered analysis plan are detailed in the Results section.

4.3.1 Method

Participants

Participants were 396 adults (221 male and 175 female, ages 20-78) recruited from MTurk. 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,[20] and was based on a power analysis using the R package simr (Green & MacLeod, 2016) based on pilot data.

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 8, 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 included three related follow-up questions, and four multiple-choice comprehension questions. Two of the multiple-choice questions for each

[20] 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.

question-answer pair concerned details that were explanatorily relevant, 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/gyuqa/.

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 8. 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.

4.3.2 Results

Analytic approach

Our analytic approach was identical to that of Study 8. In Study 9, all regression analyses included random intercepts for question-answer pair.

Testing the perceived learning prediction

First, we replicated the results of Study 8, investigating the association between perceived learning and explanatory satisfaction for a new set of question-answer pairs (Figure 10). The seven learning measures together explained significant variance in satisfaction beyond interest and knowledge, $\chi^2(1) = 234.14$, $p < .001$. 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 8, these results suggest that judgments of how much has been learned and how useful/generalizable this information is expected to be are associated with explanatory satisfaction. In addition, like Study 8, 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 at https://osf.io/gyuqa/ and as shown in Figure 10, an exploratory (non-

preregistered) analysis revealed that nearly all measures of learning 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).

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, β = -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 explanatorily relevant questions and performance on explanatorily irrelevant questions (each scored out of two). In a regression model including both scores and controlling for interest and knowledge, there was no evidence for an association between explanatory satisfaction and accurate responses concerning explanatorily relevant details, β = -0.10, 95% CI [-0.21, 0.004], $\chi^2(1) = 3.58$, $p = .06$, or explanatorily irrelevant details, β = 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, β = -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 at https://osf.io/gyuqa/).

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 Figure 11). 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 Figure 11). As in Study 8, 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.

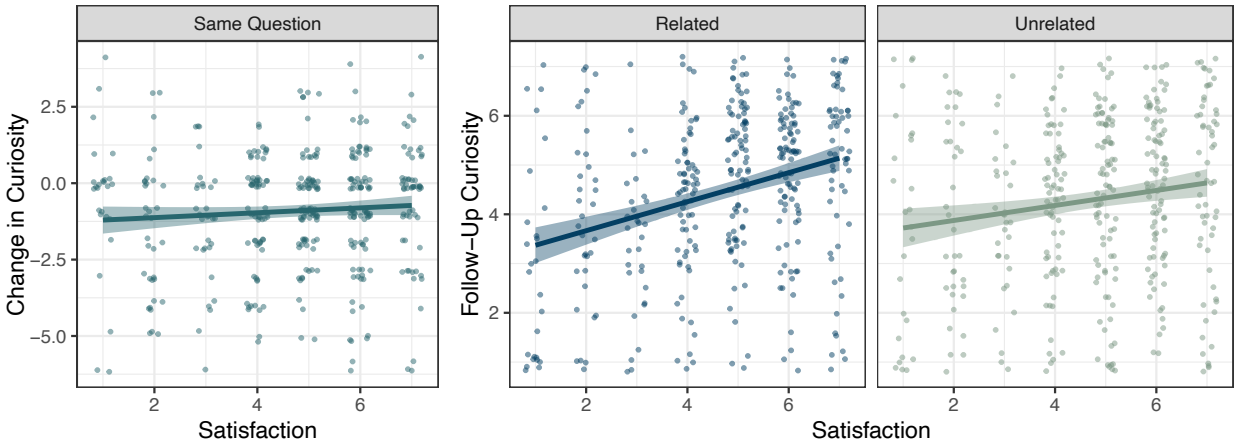


Figure 11 Explanatory Satisfaction and Subsequent Curiosity, Study 9
The association between satisfaction and (1) participants' change in curiosity about the target question, (2) participants' curiosity about related follow-up questions, and (3) participants' curiosity about unrelated follow-up questions. Each jittered point corresponds to one participant's satisfaction rating for one explanation.

4.3.3 Discussion

In Study 9, we replicated the results of Study 8 in support of the perceived learning prediction: explanations that were perceived to support (useful) learning were more likely to be found satisfying. However, there were some differences between Studies 8 and 9 in which specific measures of perceived learning 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—brought different measures of perceived learning to the forefront. We further discuss the need to study satisfaction across a broader set of explanations in the General Discussion.

The results of Study 9 also go beyond those of Study 8 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. Given these concerns, we adopt a different measure of objective learning in Study 10, and we return to the link between perceived and objective learning in Study 11.

A second way in which Study 9 went beyond Study 8 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. In tension with 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.

In support of prediction (b), we replicated the positive association from Study 8 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.

4.4 Study 10

In Study 9, we found no association between an objective measure of learning and explanatory satisfaction, but this measure had several weaknesses. In Study 10, we sought 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 contents 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 contents of the explanation and integrated these contents with prior knowledge. Thus, In Study 10, we conduct a better test of the actual learning prediction: we attempt to replicate our results with the same measure of multiple-choice understanding, but we also add an additional measure, where participants are asked to reproduce the explanation from memory to the best of their ability.

4.4.1 Method

Participants

Participants were 320 adults (158 male, 159 female, 2 other, and 1 prefer not to specify, ages 21-70) recruited from Amazon MTurk. 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 at https://osf.io/gyuqa/ (Study S4).

Materials

Study 10 used the same set of question-answer pairs as Study 9.

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 8-9), participants completed two measures of objective 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 explanation-seeking 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 9. Finally, participants provided their age and gender, and answered an attention check question that required them to identify one of the explanation-seeking questions they had previously rated from a list of four distractors.

4.4.2 Results

Analytic approach

Our analytic approach was again identical to that of Study 8. Study 10 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.

Testing the actual learning prediction: Multiple-choice performance

First, replicating Study 9, we tested whether multiple-choice performance predicted explanatory satisfaction, controlling for topic-level interest and knowledge. Overall multiple-choice performance (including performance on explanatorily relevant and explanatorily 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 explanatorily irrelevant details, $\beta = 0.03$, 95% CI [-0.04, 0.11], $\chi^2(1) = 0.80$, $p = .37$, or explanatorily relevant details, $\beta = 0.04$, 95% CI [-0.04, 0.11], $\chi^2(1) = 0.99$, $p = .32$, controlling for interest and knowledge. As in Study 9, these results provide no evidence for an association between learning—as measured by multiple-choice performance—and explanatory satisfaction.

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 (Figure 12). 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.

First, 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], $\chi^2(1) = 8.03$, $p = .005$. 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 objective learning, in particular to how well an explanation has been integrated with prior knowledge and can be reproduced in linguistic form.[21]

[21] 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, $\chi^2(1) = 13.38$, $p < .001$, 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 objective learning is associated with satisfaction selectively for the learned-from explanation.

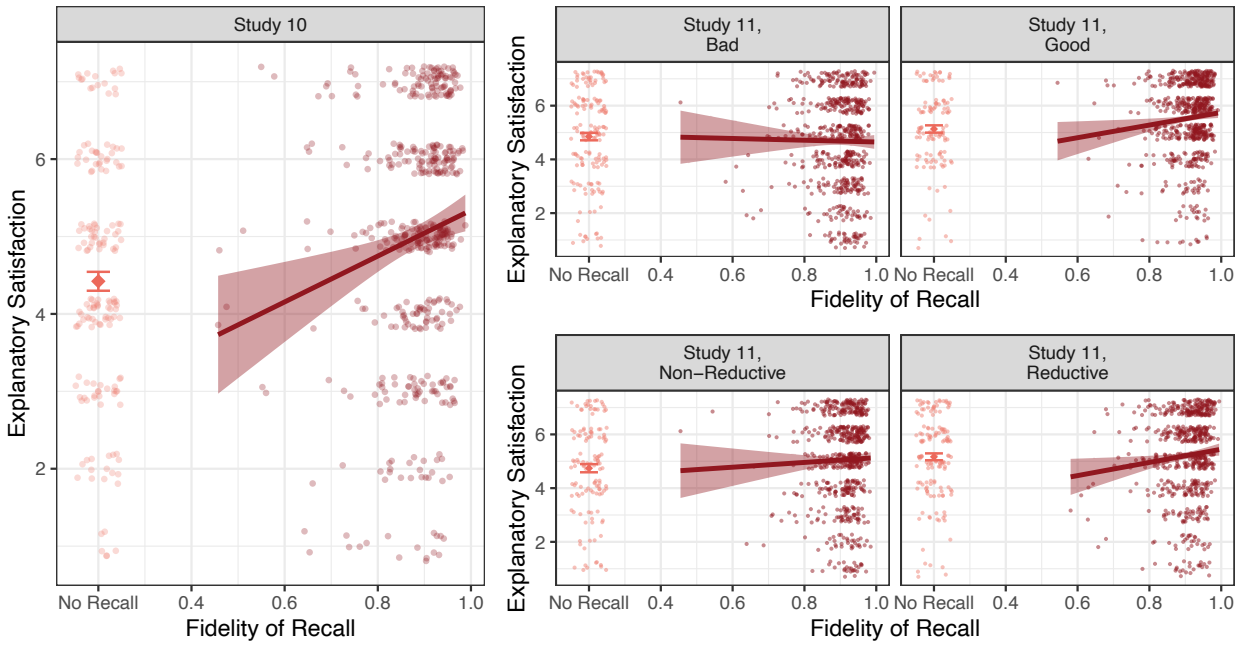


Figure 12 Explanatory Satisfaction and Actual Learning, Studies 10-11
Association between recall fidelity (either "no recall" or cosine similarity between recalled explanation and original explanation) and explanatory satisfaction in Study 10 (textbook explanations) and Study 11 (scientific explanations from Hopkins et al., 2016). For Study 11, data is plotted separately for bad explanations and good explanations (with each panel containing data for both non-reductive and reductive explanations), as well as for non-reductive and reductive explanations (with each panel containing data for both good and bad explanations).

4.4.3 Discussion

While Study 10, like Study 9, failed to find evidence for an association between multiple-choice performance and explanatory satisfaction, we did find 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 recalled accurately. More generally, this finding supports the aligned motivation account: while satisfaction may not track some measures of objective learning (e.g., multiple-choice performance), other measures (e.g., explanation recall) are in fact associated with satisfaction.

That said, the association between explanatory satisfaction and objective learning was relatively small, 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 this 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 objective learning using a broader range of good and bad explanations.

4.5 Study 11

In Study 11, we had three main aims. First, we attempt to replicate the evidence found in Study 10 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 10 were from the same reputable source and all contained explanatorily relevant information. In Study 11, we attempt to replicate the association between objective 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 objective learning. If, as we suggest, objective learning in part determines satisfaction, then an independent manipulation of satisfaction that provides no new explanatory content should have little or no effect on objective learning. If, on the other hand, satisfaction has a causal influence on later recall (e.g., because satisfying explanations are encoded with more attention), then a manipulation of satisfaction should also affect objective 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 new explanatorily 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 explanatorily relevant information, even if this is not objectively the case, and this judgment of perceived learning should mediate the effect of the reductive information on satisfaction. Further, this should be the case specifically for explanatorily relevant perceived learning, not just general 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 11 was preregistered (see https://osf.io/gyuqa/); any departures from our preregistered analysis plan are detailed in the Results section.

4.5.1 Method

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 10 effect sizes (for the association between recall fidelity and satisfaction), using the R package simr (Green & MacLeod, 2016).

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: three psychology phenomena and three physics phenomena. The full set of materials can be found at https://osf.io/gyuqa/.

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 participants' ratings reflected their evaluation of the explanation and not 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: *explanatorily 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, excluded from analyses with fewer than four correct answers). Subsequently, participants completed the recall task from Study 10 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.

4.5.2 Results

Analytic approach

We followed the same analytic approach as in Studies 8-10. 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.

Recall fidelity was calculated as in Study 10. 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 is to ensure that participants were not penalized for their inability to remember technical words, nor were recall fidelity scores inflated when they succeeded in remembering those words. Both measures of recall fidelity produced comparable results, so we report recall fidelity scores for comparisons against the read explanation below.

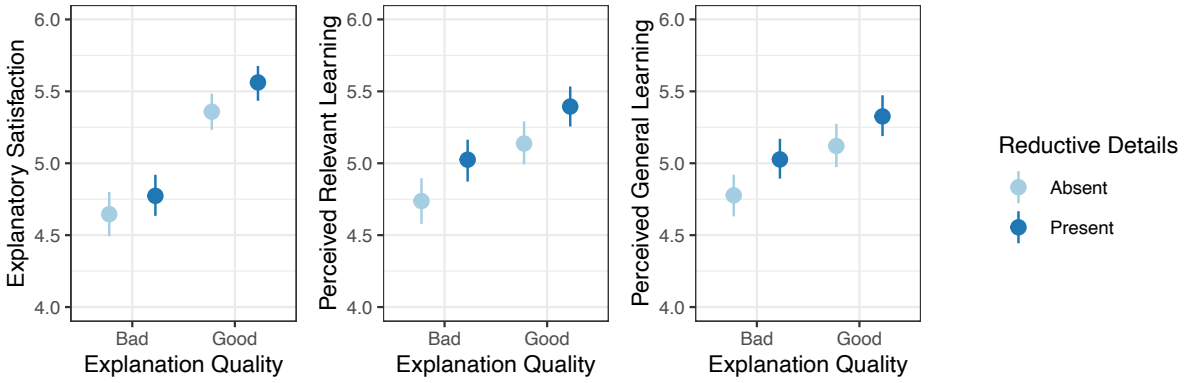


Figure 13 Manipulating Perceived Learning and Satisfaction, Study 11
Study 11 effect of explanation quality and reductive details on explanatory satisfaction, perceived explanatorily relevant learning, and perceived general learning. Points indicate the average rating on scale ranging from 1 to 7; error bars indicate 95% confidence intervals.

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.[22] Specifically, the best fitting model included explanation

[22] Our findings only partially replicate Hopkins et al.'s (2016). 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

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 Figure 13). We thus moved on to test our new predictions.

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 Figure 12). 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 Figure 12).

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

improve the quality of all explanations. However, our failure to detect this interaction has little bearing on the subsequently reported results.

details and explanation type). Our intention was to repeat the analyses above, and we report those results here. There was a 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 Figure 12). 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 significantly 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 (see Figure 12). Together, these results provide some support for the actual learning prediction, but only within high-quality explanations.

What could explain these results? As in Study 9, 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 explanatorily 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, β = 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, β = 0.08, 95% CI [0.01, 0.16], $\chi^2(1) = 4.66$, p = .03. Among bad explanations, there was no evidence for an association between recall

fidelity and perceived learning, β = -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.[23]

One potential concern is that participants recalled details from the descriptive passage 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], $\chi^2(1) = 74.58$, $p < .001$, suggesting that participants' recalled explanations were more similar to the explanations than the descriptive passages.

Explanatory vices and objective learning

Next, we tested whether explanation quality or the presence of reductive information influenced objective learning. If objective learning 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 objective learning. If, on the other hand, the causal relationship is reversed—from satisfaction to objective learning—then manipulations of satisfaction should also influence objective learning, (e.g., learning from reductive explanations should be better than learning from non-reductive

[23] 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.

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 (i.e., that reductive details have no effect on recall fidelity) was strong, BF = 28.35. Thus, our manipulation of reductive details did not influence objective learning, suggesting that objective learning is not determined by satisfaction. However, these results are difficult to interpret, given the lack of overall correlation between objective learning and satisfaction.

Does perceived learning explain explanatory vices?

Finally, we tested one possible implication of the perceived learning prediction: that individuals fall prey to explanatory vices when they believe they have learned from the misleading information. As noted already, 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 explanatorily 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 Figure 13).

Next, we fit two mediation models using the R package lavaan (Rosseel, 2012), to test whether perceived explanatorily relevant learning mediated the association between explanation quality and satisfaction, and, more critically, inclusion of reductive details and satisfaction. 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. For explanation quality, the indirect effect of perceived learning was significant, $\beta = 0.20$, 95% CI [0.12, 0.28]. The remaining direct effect was also significant, $\beta = 0.55$, 95% CI [0.44, 0.67]. In other words, perceived learning partially mediated the effect of explanation quality on satisfaction. For inclusion of reductive details, the indirect effect of perceived learning was significant, $\beta = 0.14$, 95% CI [0.07, 0.22], and the remaining direct effect was not significant, $\beta = 0.02$, 95% CI [-0.10, 0.14]. 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 explanatorily 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 Figure 13). 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 explanatorily 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 Figure 14). For explanation quality, the indirect effect of explanatorily relevant learning was significant, $\beta = 0.18$, 95% CI [0.11, 0.26], but the indirect effect of general learning was not, $\beta = 0.02$, 95% CI [-0.01, 0.04]. Moreover, the contrast between these two indirect effects was significant, $\beta = 0.16$, 95% CI [0.09, 0.24]. The remaining direct effect of explanation quality on satisfaction was also significant, $\beta = 0.55$, 95% CI [0.44, 0.67], indicating partial mediation. Likewise, for reductive details, the indirect effect of explanatorily relevant learning was significant, $\beta = 0.13$, 95% CI [0.06, 0.21], but the indirect effect of general learning was not, $\beta = 0.01$, 95% CI [-0.01, 0.03]. Moreover, the contrast between these two indirect effects was significant, $\beta = 0.12$, 95% CI [0.05, 0.20]. The remaining direct effect of reductive details on satisfaction was not significant, $\beta = 0.02$, 95% CI [-0.10, 0.14], 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.

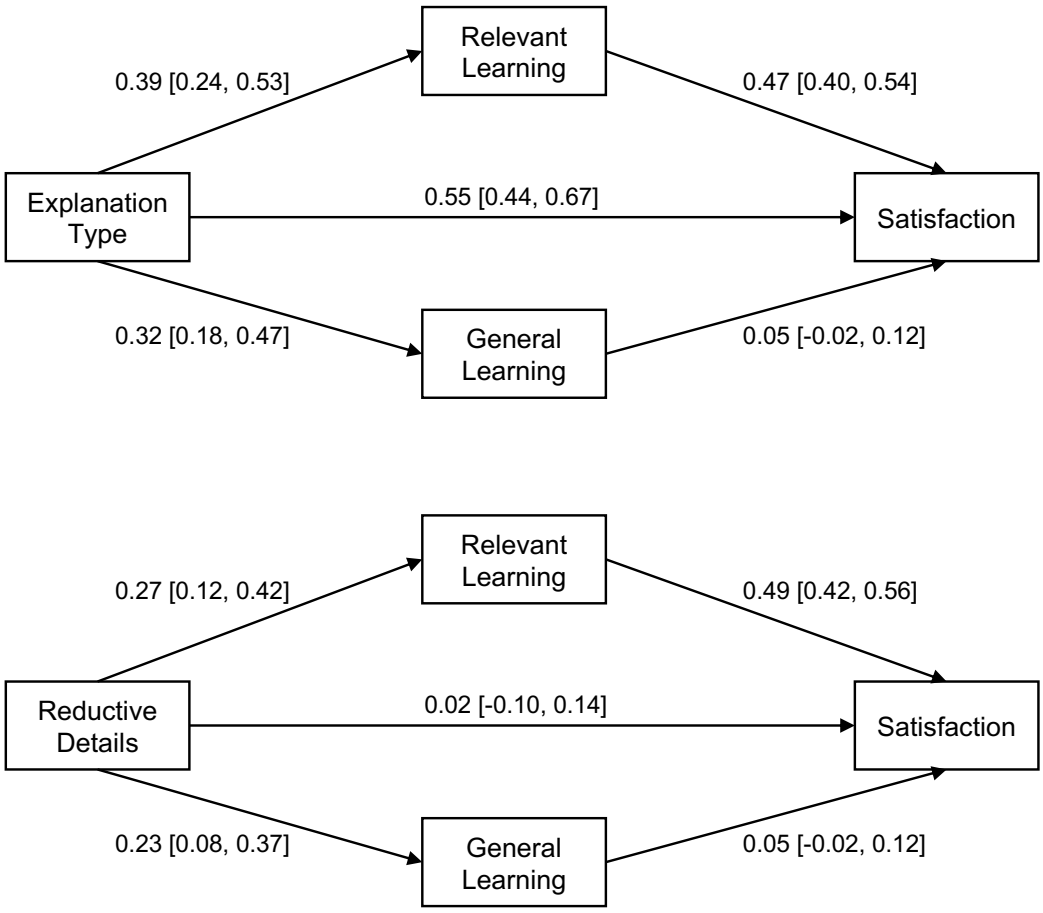


Figure 14 Mediation Analyses, Study 11

Study 11 mediation models (fit using SEM) testing mediation by perceived explanatorily 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 explanatorily relevant learning rather than perceived general learning.

4.5.3 Discussion

In Study 11, 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 set of results is exploratory and thus should be treated as tentative, these results suggest potential limits on when satisfaction tracks objective learning. Specifically, satisfaction might only track objective 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 objective learning: though satisfaction was influenced by the inclusion of reductive details, objective learning was not. Consequently, it is likely that in cases where objective 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 explanatorily 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, explanatorily relevant information. This goes beyond prior work in suggesting a possible mechanism behind the preference for reductive explanations.

4.6 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 10 that explanatory satisfaction was related to the fidelity with which an explanation was recalled, and we found in Study 11 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 9-10, and we did not replicate the association between recall fidelity and satisfaction across both good and bad explanations in Study 11. 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 9 and among bad explanations in Study 11, we also failed to find associations between perceived learning and actual learning. In contrast, among good explanations in Study 11, there was an association between perceived learning and actual learning.

Support for the perceived learning prediction was relatively strong. We found in Studies 8 and 9 that explanatory satisfaction was predicted by direct and indirect judgments of useful learning. In particular, controlling for other measures, significant variance in explanatory satisfaction was explained by judgments of how much was learned from the explanation, how useful the explanation would be in the future, and the extent to which the explanation picked out a general pattern or regularity, as well as how much information the explanation contained and

whether it seemed to be expert-produced (in Study 8 only) and how simple the explanation was (in Study 9 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, through both direct and indirect means.

Support for the selective reinforcement prediction was mixed. In Studies 8-9, 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 9, we did not find evidence that satisfaction was related to decreased curiosity about the initial question. We suspect that the generally strong associations between satisfaction and curiosity observed in Studies 8-9 were driven by the two inferences that participants made in parallel: one about the value of topical inquiry (e.g., "How valuable is it to pursue inquiry about dinosaur digestion?"), and the other about the value of available informants (e.g., "How good are explanations from this source?"). While we predicted selective reinforcement for topical inquiry, it's plausible that a satisfying explanation offered more general evidence concerning the value of the source (see also Corriveau & Kurkul, 2014; Koenig et al., 2004; Koenig & Harris, 2005; Landrum et al., 2015).

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 11, we found evidence in support of this prediction: perceived explanatorily relevant learning (above and beyond perceived general learning) fully mediated the effect of irrelevant reductive information on satisfaction.

4.6.1 Implications

The results from these studies provide evidence that explanatory satisfaction is partially aligned with epistemic success. On the one hand, our results suggest that satisfaction tracks objective 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). 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: 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? Conversely, what environments might take advantage of an imperfectly aligned agent (e.g., echo chambers; see Nguyen, 2021)? In general, a more detailed characterization of the links between the environment, phenomenology, inquiry, and 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 new 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 an identical unnamed condition ("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 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 or pragmatic 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 found in Chapters 2 and 3 that curiosity is (sometimes) 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 pragmatic 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 pragmatic motives. By combining self-report measures of curiosity and satisfaction—as we have done in this and other research (Liquin & Lombrozo, 2020)—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.

Finally, our findings have implications for understanding and predicting inquiry. Prior research has shown that, when provided with a given target of inquiry (e.g., identifying a target object from an array of possibilities, like the children's game "Guess Who?"), both children and adults choose queries that are well-described by optimal models of expected information gain (Bramley et al., 2015; McCormack et al., 2016; Ruggeri et al., 2016, 2017; Ruggeri & Lombrozo, 2015). But how is the target of inquiry itself selected? Our results hint at one mechanism by which learners narrow the vast space of possible targets: by considering the topics in which they have found satisfying explanations in the past (see also Oudeyer et al., 2007; Schmidhuber, 2010). This is consistent with our finding that satisfaction was especially

predictive of subsequent curiosity about related follow-up questions, as opposed to unrelated questions.

4.6.2 Limitations

Several limitations of these studies must be acknowledged. First, these studies are limited by their predominant use of correlational methods. While our account implies that satisfaction is determined by subjective and/or actual learning, and that satisfaction causes subsequent inquiry, the present research only explored the potential causal link between actual learning and satisfaction (in Study 11). It remains unclear, for example, whether satisfaction causally influences subsequent inquiry, or whether satisfaction and subsequent curiosity are both caused by some third variable (e.g., judgments of learning). Thus, while our theory of explanatory phenomenology provides one causal pathway relating these variables, and our results are consistent with this theory, future research is needed to test the posited causal links between individual variables.

Another limitation is that participants received a single explanation in response to each provided question. In contrast, much of the previous research on explanatory preferences (Cimpian & Steinberg, 2014; Johnson et al., 2016, 2019; Johnston et al., 2017, 2018, 2019; Lombrozo, 2007; Pacer & Lombrozo, 2017) has asked participants to contrast the quality of two competing explanations for the same event or phenomenon. The quality of competing alternative explanations has an influence on whether participants are willing to infer that the best explanation is true (Douven & Mirabile, 2018), and we might expect the presence of competing alternatives to influence explanation evaluation, as well. For example, satisfaction judgments in the present research (and in any context where explanations are evaluated in isolation) may reflect satisfaction that an explanation exists, rather than satisfaction that this explanation fulfills

certain explanatory desiderata relative to other possible explanations. Future work should explore whether different criteria are used to evaluate explanations presented simultaneously versus in isolation, and whether this influences the association between satisfaction and epistemic success (either subjective or objective).

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 8), targeted towards teaching academic material (Studies 9-10), or targeted towards explaining scientific phenomena (Study 11). With the exception of the "bad" explanations in Study 11, the explanations we provided to participants were hand-crafted to teach specific information in a satisfying way. Future research must explore a broader range of questions and explanations, especially those drawn from sources that do not include exclusively satisfying and true 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.

4.6.3 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.

5 CHAPTER 5: CONCLUSION

This dissertation sought to characterize the phenomenological states that guide our search for explanations: explanation-seeking curiosity and explanatory satisfaction. Explanation-seeking curiosity motivates us to initiate explanatory inquiry, and explanatory satisfaction signals our search for explanations has succeeded. However, these phenomenological states do not necessarily guide *successful* inquiry, which results in learning a useful explanation. In particular, if curiosity and satisfaction occur when useful learning is unlikely, inquiry will frequently fail to result in learning useful explanations. In contrast, if curiosity and satisfaction occur when learning is likely, inquiry will be more successful.

In Chapter 1, I motivated competing accounts of explanatory phenomenology: the indiscriminate account, the goal-misaligned account, the goal-aligned account, and several imperfectly goal-aligned accounts. The indiscriminate account, which predicts no selectivity in explanation-seeking curiosity or explanatory satisfaction, was immediately rejected—curiosity and satisfaction are clearly experienced for some questions and explanations more than others (see Gottlieb et al., 2013; Kidd & Hayden, 2015; Lombrozo, 2016). However, prior research did not clearly differentiate between the goal-aligned account, which predicts that curiosity and satisfaction are selectively experienced when useful learning is likely to occur or has occurred, and the imperfectly goal-aligned account, which predicts that curiosity and satisfaction are experienced selectively, but according to heuristic cues to useful learning or subjective perceptions of useful learning that sometimes lead inquiry astray. In addition, there was some evidence for the goal-misaligned account, which predicts that curiosity and satisfaction are experienced selectively but independently of useful learning.

In 11 studies across Chapters 2-4, I asked whether curiosity and satisfaction were aligned, misaligned, or imperfectly aligned with useful learning. In general, results supported an imperfectly goal-aligned account: curiosity and satisfaction were responsive to individuals' perceptions of useful learning, curiosity was additionally sensitive to heuristic cues to learning, and satisfaction was additionally sensitive to indirect cues to an explanation's quality. However, we also found weak evidence that curiosity and satisfaction were aligned with actual useful learning under certain conditions. In the following section, I review these results in more detail, providing answers to the six questions I sought to address.

5.1 Summary of empirical work

5.1.1 Curiosity: Summary of Chapters 2-3

First, I posed the following three questions: (1) When expected useful learning and alternative heuristic cues are disentangled, is explanation-seeking curiosity aligned with the goal of useful learning? (2) To the extent that curiosity is aligned with the goal of useful learning, is it aligned with the actual likelihood of learning, or subjective expectations about learning? and (3) Is curiosity aligned with epistemic goals (learning in general) or instrumental goals (learning useful information)? Chapters 2 and 3 together addressed these questions, using a diverse range of methods to investigate when explanation-seeking curiosity is experienced.

Chapter 2 tested whether curiosity co-occurs with subjective expectations about useful learning, above and beyond heuristic cues to learning. Study 1 revealed that "forward-looking" potential triggers of curiosity concerning the goals of inquiry (such as perceived expected learning) explain unique variance in curiosity above and beyond backward-looking heuristic triggers (such as surprise). Study 2 explored the underlying structure of forward- and backward-looking triggers of curiosity, revealing that participants' ratings clustered into three latent

factors: forward-looking expectations about learning, forward-looking expectations about export to new situations, and backward-looking heuristic cues to learning. Each cluster of triggers explained unique variance in curiosity ratings, suggesting that curiosity independently tracks expectations about learning, expectations about utility, and heuristic cues to learning. Studies 3a and 3b sought to replicate these results in new sets of questions (including fact-seeking in addition to explanation-seeking questions). The latent structure of the triggers of curiosity replicated across both new sets of explanation-seeking questions (but not fact-seeking questions), though the three clusters were differentially predictive of curiosity in different sets of questions.

Chapter 3 tested whether curiosity co-occurs with actual expected learning and expected useful learning according to an optimal model, above and beyond heuristic cues as captured in the same model. Study 4 revealed that curiosity was most strongly associated with uncertainty and/or expected learning about the question of interest. Additional variance in curiosity was explained by other heuristic cues to learning but not optimal expected learning about the global target of inquiry. Study 5 reduced demands on participants' attention and found that a small amount of variance in curiosity was accounted for by global expected learning. Studies 6 and 7 manipulated participants' incentivized goal to investigate whether curiosity shifted to track the instrumental value of information. Though the manipulation influenced participants' behavior, curiosity only shifted in one study and primarily according to heuristic cues to instrumental value.

Together, the studies in Chapters 2 and 3 support the imperfectly goal-aligned account of curiosity, according to which curiosity is selectively experienced in a way that approximates alignment with useful learning but falls short in predictable ways. In particular, when expected useful learning and alternative heuristic cues (e.g., surprise, novelty) were disentangled,

explanation-seeking curiosity was aligned with both perceptions of learning, perceptions of the utility of information, and perceptions of heuristic cues (Chapter 2). However, the evidence that curiosity tracked actual expected learning about the ultimate target of inquiry (as operationalized by an optimal model) was relatively weak, and curiosity only tracked heuristic cues to instrumental value (Chapter 3).

Because subjective expectations about useful learning and heuristic cues to learning are fallible, explanation-seeking curiosity appears to fall short of being goal-aligned. However, Chapter 3 raises many open questions about when curiosity might in fact track useful learning. In the studies reported in Chapter 3, curiosity was most strongly associated with uncertainty and/or expected learning about the query of interest according to an optimal model. Uncertainty and expected learning were confounded in this analysis—but if this association is driven by expectations about learning rather than uncertainty, this would constitute evidence for goal-alignment with actual learning. In addition, curiosity was associated with actual expected learning about the global target of inquiry in Study 5, suggesting that curiosity might be goal-aligned with actual learning under certain circumstances. Together, these findings could vindicate the goal-aligned account—but further research is needed to test when curiosity might be goal-aligned and when it falls short.

5.1.2 Satisfaction: Summary of Chapter 4

Second, I posed the following three questions: (1) Is explanatory satisfaction aligned with the goal of useful learning? (2) To the extent that satisfaction is aligned with the goal of useful learning, is it aligned with actual useful learning, or perceived useful learning? and (3) Is satisfaction aligned with epistemic goals (learning in general) or instrumental goals (learning useful information)?

Chapter 4 addressed these questions, investigating whether explanatory satisfaction is selectively experienced when useful learning (actual or perceived) has occurred. Studies 8 and 9 found that explanatory satisfaction was associated with direct judgments of learning (e.g., how much was learned) and indirect judgments of explanation quality (e.g., how simple the explanation was). Studies 9 through 11 found that explanatory satisfaction was unrelated or weakly related to measures of how much a participant actually learned from the explanation. In particular, satisfaction was associated with the fidelity of explanation recall, but only when perceived learning and recall fidelity were associated (likely indicating that perceived learning is the more reliable predictor of satisfaction). Finally, Study 11 found that perceptions of learning can fully explain a cue to satisfaction that previously was cast as misaligned with learning: the preference for explanations that contain logically irrelevant reductive information (Hopkins et al., 2016; Weisberg et al., 2008).

Together, the studies in Chapter 4 support the imperfectly goal-aligned account of satisfaction, according to which satisfaction is selectively experienced in a way that approximates alignment with useful learning but falls short in particular ways. In particular, satisfaction was associated with perceptions of learning, perceptions of the utility of information, and indirect assessments of an explanation's quality. However, satisfaction was only associated with actual learning when perceptions of learning were accurate. In other words, satisfaction is as aligned with learning as possible given our limited ability to accurately assess our own learning. Nonetheless, our assessments of learning will often lead us astray (Glenberg & Epstein, 1985; Lin & Zabrucky, 1998; Maki, 1998), and thus, satisfaction falls short of being goal-aligned.

5.2 Limitations

Though I discussed several limitations of each set of studies in Chapters 2-4, I reiterate some of the most important limitations here. First, nearly all of the studies presented in this dissertation were correlational, and thus we are unable to establish causal links between actual or perceived useful learning, heuristic cues, and explanatory phenomenology. In Chapter 3, we found in one study that the instrumental value of information affected curiosity by manipulating participants' goals. In addition, in Chapter 4, we found that actual learning was unlikely to be causally influenced by satisfaction. However, many other causal links remain to be explored. For example, it is possible that perceived learning is a trigger of curiosity and satisfaction, but it is also possible that perceived learning and curiosity/satisfaction have a common cause. Future research is needed to assess these possibilities.

Second, I focused in this dissertation on *explanation-seeking* curiosity and *explanatory* satisfaction. Explanations help us make sense of the world, and they inform reasoning (e.g., Ahn et al., 2002; Karmiloff-Smith & Inhelder, 1974), inference (e.g., Lombrozo, 2016), and action (Bonawitz et al., 2012; Legare, 2012). An account of when and why we seek explanations is therefore likely to have broad implications, along with practical applications in educational contexts. However, a natural question is whether the imperfectly goal-aligned account of curiosity and satisfaction extends beyond an explanation-seeking context. On the one hand, many of the studies on curiosity that motivated this research (for reviews, see Gottlieb et al., 2013; Kidd & Hayden, 2015) concern information search broadly, rather than explanation search specifically. On the other hand, explanations have unique epistemic and practical benefits, which could shape the triggers of curiosity and satisfaction (e.g., simplicity and breadth might inform inquiry towards explanations but not facts). In Chapter 2, we found that the potential triggers of

curiosity formed a less coherent structure when rated in response to fact-seeking questions as opposed to explanation-seeking questions, though we found few consistent differences in the potential triggers' associations with curiosity. On the basis of these results, it is unclear whether explanation-seeking curiosity and fact-seeking curiosity should be characterized as the same phenomenology directed towards different targets, as opposed to qualitatively different phenomenologies (see Schwitzgebel, 1999). Future research is needed to answer this question.

Third, I investigated explanatory inquiry in constrained tasks that do not approach the richness of real-world inquiry. In the real world, inquiry is extended in time, iterative, and often social. That is, we might spend many hours researching a single question (or even many years, as scientists); each piece of new information might open new questions or lead us back to old questions; and information from other people can answer our questions (e.g., Aguiar et al., 2012; Bromme & Thomm, 2016; Keil et al., 2008; Lutz & Keil, 2002; Mills et al., 2010) and inform when we experience curiosity in the first place (Dubey et al., 2021; Y. Wu & Gweon, 2020). In the present research, in contrast, curiosity was only briefly elicited or satisfied, participants did not have the opportunity to pursue further questions of interest, and participants had no access to social information. The richness of real-world inquiry is likely to shape the experience of curiosity and satisfaction—for example, the affective experience associated with curiosity shifts depending on whether curiosity will be resolved quickly or later in time (Noordewier & Dijk, 2017). Thus, it will be important to examine explanatory phenomenology in more complex inquiry tasks.

Finally, my main questions were addressed using self-report measures of curiosity and satisfaction. The use of self-report measures has some advantages over the use of behavioral measures, which I described in Chapter 1—for example, decisions to seek information might be

partly informed by curiosity but also incorporate other considerations. However, individuals might not have perfect introspective access to their own curiosity or satisfaction, or they might have difficulty mapping these phenomenological states onto a Likert scale. Research on emotion, which faces similar methodological challenges, has accompanied self-report measures with both behavioral and physiological measures (Mauss & Robinson, 2009). It is possible that research on curiosity and satisfaction would also benefit from this approach, for example by measuring pupil dilation (Brod & Breitwieser, 2019; Kang et al., 2009) or neural responses (Jepma et al., 2012; Kang et al., 2009; Ligneul et al., 2018; van Lieshout et al., 2018; for a review, see Gottlieb & Oudeyer, 2018), along with curiosity and information search. In lieu of using behavioral and physiological measures, we attempted to validate our self-report measures in Chapters 2 and 4. In Chapter 2, curiosity was associated with later information search decisions, as predicted. However, in Chapter 4, satisfaction was not related to decreased curiosity about the initial question. Further research combining self-report measures with behavioral measures could provide further insight into whether satisfaction halts inquiry, and more broadly, how curiosity and satisfaction motivate inquiry and are shaped by the goal of useful learning.

5.3 Open questions on imperfect goal-alignment

Despite these limitations, the work presented in this dissertation sheds light on when humans experience explanation-seeking curiosity and explanatory satisfaction. In particular, curiosity and satisfaction appear to be imperfectly aligned with the goal of learning useful explanations, and thus experienced according to perceptions of or heuristic cues to useful learning. However, further research is still needed to refine our understanding of the imperfectly goal-aligned account. Below, I discuss three open questions raised by this work.

5.3.1 Imperfectly aligned, or aligned with a different goal?

First, I assumed throughout this dissertation that the goal of explanatory inquiry is to achieve useful learning. Other goals beyond the epistemic and instrumental are certainly possible (e.g., Malle, 2004; Mercier, 2016). In addition, we might consider alternative formulations of the epistemic and instrumental goals of explanatory inquiry. According to the formulation assumed throughout this dissertation, perfectly goal-aligned inquiry should lead to new and useful information in every individual instance of inquiry. Consequently, I explained failures in explanatory inquiry as signals that explanatory phenomenology is imperfectly goal-aligned. However, another explanation is possible: these failures could be signals that explanatory phenomenology is perfectly goal-aligned, but with respect to an alternative formulation of useful learning. For example, perfectly goal-aligned inquiry might lead to the largest body of new and useful information in the long run. As explicated briefly in Chapter 4, these two formulations of what it means to be epistemically and instrumentally aligned can diverge: for example, many individual experiments in psychology fail to provide new and useful explanations for a given psychological phenomenon. However, these failed experiments often critically inform the next steps of our research, so that scientific inquiry succeeds in finding useful explanations in the long run.

This proposal is reminiscent of algorithms for reinforcement learning that balance exploration and exploitation (e.g., Cohen et al., 2007; Hills et al., 2015; Sutton & Barto, 1998). The best policy for maximizing reward in the long run is not to maximize reward in the short run, but rather to balance reward-seeking behavior with occasional exploration. Likewise, inquiry might more effectively lead to useful learning in the long run by occasionally appearing

suboptimal in the short run—allowing us to find information where it was not initially expected. These two possibilities must be differentiated in future research.

5.3.2 Epistemic environments

The imperfectly goal-aligned account also raises questions about the kinds of epistemic environments we inhabit. In Chapter 1, I argued that even indiscriminate or goal-misaligned inquiry can successfully achieve useful learning to the extent that we inhabit an "epistemically friendly" environment. That is, if good and useful explanations are plentiful, it might not matter when we pursue information, as we are likely to find good and useful explanations wherever we look. On the other hand, if good and useful explanations are sparse, goal-alignment might be necessary for good and useful explanations to be found.

The key findings presented in this dissertation raise many questions about our epistemic environment. For example, is imperfect goal-alignment adaptive in certain environments? Again, the reinforcement learning literature could provide some insight into this question: for a reinforcement learning agent, non-reward-maximizing exploration is especially useful when the environment changes over time, as exploration allows the agent to notice and later exploit these changes (Knox et al., 2012; Navarro et al., 2016; Speekenbrink & Konstantinidis, 2015). Likewise, we might predict that imperfect goal-alignment is most useful when our epistemic environment is changing, as it allows us to detect and respond to these changes. Future work is needed to examine whether our epistemic environment in fact changes, but this appears plausible. For example, educational settings are designed to provide us with maximally informative and useful explanations, yet we are faced with rapidly spreading misinformation online (Vicario et al., 2016; Vosoughi et al., 2018), which could threaten our ability to attain true and useful explanations. Thus, we might face relatively friendly environments in some settings

and relatively unfriendly environments in other settings. Further empirical and computational work is needed to determine whether imperfectly aligned curiosity and satisfaction are beneficial for navigating such environments.

Future research might also consider whether the cues that shape explanation-seeking curiosity and explanatory satisfaction change according to one's epistemic environment. In particular, curiosity and satisfaction might become less aligned with useful learning when informative and useful explanations are plentiful and more aligned with useful learning when informative and useful explanations are rare. Prior work shows that exploration is more closely aligned with optimal expected learning after receiving feedback that a simpler strategy leads to poor performance (Coenen et al., 2015), suggesting that people adjust their exploration depending on the viability of certain strategies in a given environment. However, it remains an open question whether this extends to curiosity and satisfaction. By considering the cues that shape curiosity and satisfaction in combination with the epistemic environment, we can gain further insight into when imperfectly aligned explanatory phenomenology helps us find new and useful explanations, as well as when it impedes us.

5.3.3 Development and evolution

Finally, there are many open questions about the evolutionary and developmental origins of explanatory phenomenology. Some aspects of adult-like curiosity and satisfaction appear to be present from an early age. For example, very early in development, infants notice stimuli that violate their expectations (Spelke et al., 1992; Xu & Garcia, 2008). Furthermore, attention to surprising events promotes subsequent exploration and learning in 10- to 12-month-olds (Stahl & Feigenson, 2015). There is also evidence that rhesus macaques sacrifice rewards to gain information, and that this preference for information scales with the extent to which it resolves

uncertainty (Wang & Hayden, 2019), suggesting that the desire to resolve uncertainty is not uniquely human. These studies do not differentiate between a goal-aligned account and an imperfectly goal-aligned account—infants' and monkeys' information search could be motivated by the actual potential for learning, the perceived potential for learning, or by surprise and uncertainty (i.e., heuristic cues). However, these results suggest that selectivity in inquiry, perhaps motivated by curiosity, could have partly evolutionary origins.

There is also clear evidence that selectivity in inquiry develops during childhood. For example, children's ability to select maximally informative questions or interventions improves across development (e.g., Nussenbaum et al., 2020; Ruggeri et al., 2016; Ruggeri & Lombrozo, 2015; for a review, see Jones et al., 2020). Children's curiosity might similarly become more responsive to expected learning (or imperfect cues to expected learning) with age. In addition, metacognition improves across development: for example, children's judgments about their own uncertainty become more accurate during the preschool years (Lyons & Ghetti, 2011). Thus, even if curiosity is sensitive to perceived uncertainty from an early age, the reliability of perceived uncertainty in accurately signaling learning might improve across development. Finally, children's ability to identify relevant and non-circular explanations improves across development (Baum et al., 2008; Corriveau & Kurkul, 2014; Johnston et al., 2019; Mills et al., 2017, 2019). Thus, children's perceptions of explanatory learning likely become more aligned with actual learning across development. Taken together, it is likely that the cues that shape curiosity and satisfaction change across development, and that the reliability of these cues as guides to learning changes as children's explanation evaluation and metacognition improves. Future research is needed to evaluate these possibilities.

It is additionally possible that the goal of inquiry changes across human development. For example, it has been suggested that adults privilege "exploitation"—taking advantage of learned information to guide efficient action—while children privilege "exploration"—seeking information (Gopnik, 2020). If adults' and children's explanatory phenomenology serves these distinct goals, we might expect adults' curiosity and satisfaction to more closely track the instrumental value of information (i.e., its potential for guiding efficient action) and children's curiosity and satisfaction to more generally track any potential for learning. In Chapters 2-4, I presented evidence that adults' curiosity and satisfaction are at least somewhat sensitive to the (perceived or actual) instrumental value of information, above and beyond general learning. In addition, there are shifts across development in how individuals explore in instrumental contexts (Blanco & Sloutsky, 2019, 2020; Liquin & Gopnik, 2020; Meder et al., 2021; E. Schulz et al., 2019; Somerville et al., 2017; Sumner et al., 2019). However, further research is needed to evaluate whether adults' and children's explanatory phenomenology is guided by these (or other) goals.

5.4 Explanatory phenomenology and useful learning

In this dissertation, I investigated whether and how the phenomenological states associated with explanatory inquiry—explanation-seeking curiosity and explanatory satisfaction—guide inquiry towards informative and useful explanations. Specifying the cues that shape our experience of curiosity and satisfaction has broad implications for understanding the process of explanatory inquiry and intervening to improve it. For example, insight into the cues that shape curiosity and satisfaction can inform how we construct better environments for learning. In classrooms, educators might design materials that trigger then satisfy curiosity, precisely tuned to the most important course content. Online, the cues of curiosity and

satisfaction might be harnessed to promote more effective inquiry on search engines or question-and-answer forums. Moreover, by differentiating the cues that are responsible for triggering curiosity and satisfaction, we know where to place blame when something goes wrong: when we seek information where none is available, or when we feel satisfied by an uninformative explanation. For example, if curiosity and satisfaction are triggered by perceived learning, then interventions can improve the efficacy of inquiry by improving the accuracy of these perceptions.

In 11 studies across Chapters 2-4, explanation-seeking curiosity and explanatory satisfaction were imperfectly aligned with the goal of useful learning: curiosity and satisfaction were generally sensitive to individuals' perceptions of useful learning and to heuristic cues like surprise, uncertainty, and novelty. There was some evidence that curiosity and satisfaction were aligned with actual learning—but this evidence was in some cases ambiguous (see Chapter 3) and in other cases was inconsistent across studies (see Chapters 3-4). While these results may indicate that humans fall short of perfect alignment with useful learning, I argued in this chapter that some degree of misalignment could be beneficial for certain goals or in certain environments, raising new questions for future research.

Taken together, this work demonstrates that explanation-seeking curiosity and explanatory satisfaction play a role in guiding inquiry towards informative and useful explanations, but they do so imperfectly. In the right environments, our imperfect inquiry will nonetheless succeed in helping us understand the world, predict future events, and intervene to change them.

6 APPENDICES

APPENDIX A

Table A1 Descriptive statistics for all potential trigger measures, ESC, and FSC in Studies 1-3 (Chapter 1)

Measure	Mean (SD), Study 1	Mean (SD) Study 2, ELI5	Mean (SD) Study 2, ConceptNet	Mean (SD) Study 3a, ESQs	Mean (SD) Study 3a, FSQs	Mean (SD) Study 3b, ESQs	Mean (SD) Study 3b, FSQs
Breadth	4.07 (1.71)	4.07 (1.83)	3.47 (1.82)	3.86 (1.78)	2.57 (1.73)	3.32 (1.92)	2.87 (1.99)
Anticipated Learning	4.93 (1.70)	5.03 (1.68)	4.29 (1.97)	5.26 (1.54)	4.62 (1.88)	3.99 (1.91)	4.34 (2.03)
Learning Potential	5.28 (1.56)	5.20 (1.65)	4.67 (1.87)	5.56 (1.41)	4.63 (1.91)	4.41 (1.88)	4.47 (1.96)
Regularity	4.71 (1.54)	4.64 (1.58)	4.23 (1.77)	4.76 (1.63)	3.02 (1.84)	3.75 (1.82)	3.33 (1.94)
Future Utility	4.19 (1.79)	4.00 (1.92)	2.87 (1.77)	4.35 (1.89)	3.11 (1.90)	3.08 (1.90)	3.05 (1.93)
Expertise	4.38 (1.75)	4.60 (1.78)	3.89 (1.97)	4.76 (1.75)	3.33 (2.01)	3.61 (1.94)	3.60 (2.07)
Complexity	4.07 (1.78)	4.30 (1.86)	3.34 (1.74)	4.29 (1.74)	2.43 (1.72)	3.28 (1.90)	2.81 (1.98)
Surprise	3.28 (1.92)	3.33 (2.08)	2.23 (1.73)	3.56 (1.83)	3.45 (2.05)	3.52 (1.87)	3.08 (1.99)
Novelty	3.18 (1.90)	3.32 (2.06)	2.62 (2.01)	3.93 (1.83)	3.67 (2.00)	3.42 (1.88)	3.40 (2.08)
Information Gap	4.49 (1.98)	4.84 (2.01)	4.15 (2.22)	4.96 (1.93)	5.43 (2.10)	4.94 (2.04)	5.61 (2.03)
Map Mismatch	3.02 (1.64)	3.29 (1.72)	2.72 (1.82)	3.35 (1.52)	4.08 (1.54)	3.75 (1.49)	3.54 (1.55)
Fact-and-Foil	4.34 (1.94)	4.25 (2.00)	3.70 (2.13)				
Negative Valence	3.87 (1.28)	3.86 (1.25)	3.59 (1.00)	3.97 (1.39)	3.55 (1.46)	4.28 (1.45)	3.50 (1.58)
Curiosity	4.67 (1.84)	4.42 (1.99)	3.52 (2.05)	4.94 (1.75)	4.55 (2.11)	4.12 (2.05)	4.09 (2.24)

APPENDIX B

Table B1 Standardized regression coefficients predicting curiosity (ESC in Studies 1-2 and ESC/FSC in Studies 3a-3b), with all measures entered simultaneously and with random intercepts for participant and item.

Measure	*β* [95% CI], Study 1	*β* [95% CI], Study 2	*β* [95% CI], Study 3a ESC	*β* [95% CI], Study 3b ESC	*β* [95% CI], Study 3a FSC	*β* [95% CI], Study 3b FSC
Anticipated Learning	0.28*** [0.23, 0.34]	0.22*** [0.16, 0.27]	0.39*** [0.32, 0.46]	0.18*** [0.11, 0.25]	0.18*** [0.12, 0.24]	0.27*** [0.20, 0.34]
Learning Potential	0.15*** [0.10, 0.19]	0.18*** [0.13, 0.24]	0.15*** [0.08, 0.22]	0.22*** [0.16, 0.29]	0.20*** [0.14, 0.27]	0.13*** [0.06, 0.20]
Breadth	0.01 [-0.03, 0.05]	0.03 [-0.004, 0.07]	0.02 [-0.02, 0.07]	0.03 [-0.02, 0.08]	0.06 [-0.01, 0.13]	-0.02 [-0.08, 0.04]
Regularity	0.01 [-0.03, 0.06]	0.09*** [0.05, 0.13]	-0.02 [-0.08, 0.04]	-0.04 [-0.10, 0.01]	0.04 [-0.03, 0.11]	0.07* [0.02, 0.13]
Future Utility	0.24*** [0.20, 0.29]	0.20*** [0.16, 0.25]	0.22*** [0.17, 0.27]	0.15*** [0.09, 0.21]	0.13*** [0.06, 0.20]	0.24*** [0.17, 0.30]
Expertise	0.05* [0.005, 0.10]	-0.01 [-0.06, 0.04]	0.02 [-0.04, 0.09]	0.01 [-0.05, 0.07]	-0.01 [-0.07, 0.05]	-0.03 [-0.09, 0.03]
Complexity	0.03 [-0.02, 0.08]	0.06** [0.02, 0.11]	0.00 [-0.06, 0.06]	0.02 [-0.04, 0.08]	-0.06 [-0.13, 0.02]	-0.02 [-0.08, 0.04]
Surprise	0.13*** [0.09, 0.18]	0.16*** [0.11, 0.21]	0.07** [0.02, 0.13]	0.15*** [0.09, 0.20]	0.26*** [0.19, 0.32]	0.05 [-0.01, 0.11]
Novelty	0.00 [-0.05, 0.05]	0.04 [-0.01, 0.09]	0.09** [0.03, 0.15]	0.18*** [0.12, 0.24]	0.13*** [0.06, 0.19]	0.14*** [0.09, 0.20]
Information Gap	0.10*** [0.05, 0.14]	0.09*** [0.04, 0.14]	0.05 [-0.01, 0.10]	0.06* [0.01, 0.11]	0.03 [-0.03, 0.09]	0.08** [0.02, 0.13]
Map Mismatch	-0.05* [-0.09, -0.01]	-0.07** [-0.11, -0.03]	-0.10*** [-0.15, -0.05]	-0.05* [-0.09, -0.003]	-0.06* [-0.12, -0.005]	-0.09*** [-0.14, -0.04]
Fact-and-Foil	0.02 [-0.02, 0.06]	-0.003 [-0.04, 0.03]	---	---	---	---
Negative Valence	0.01 [-0.02, 0.05]	0.02 [-0.02, 0.06]	-0.01 [-0.06, 0.03]	-0.01 [-0.06, 0.04]	-0.06* [-0.11, -0.01]	-0.04 [-0.09, 0.02]

* $p < .05$; ** $p < .01$; *** $p < .001$

APPENDIX C

Here we derive an efficient expression for the posterior over θ. Note that we rely on the formulation of the posterior described in Study 6, which introduces the variable $\tilde{z}_t$. For Studies 4-5, $\tilde{z}_t$is fixed as $\oslash$ for all trials (i.e., the latent intervention is never observed).

First, by Bayes rule, we have

$$p(\theta \mid r_{1:t}, \tilde{z}_{1:t}, \zeta, \varepsilon) = \frac{p(\theta)p(r_{1:t}, \tilde{z}_{1:t} \mid \theta, \zeta, \varepsilon)}{\int_0^1 p\,(\theta)p(r_{1:t}, \tilde{z}_{1:t} \mid \theta, \zeta, \varepsilon)}.$$

Note that we have used the fact that $p(\theta \mid \zeta, \varepsilon) = p(\theta)$. We assume a uniform prior on θ, thus reducing the posterior to the normalized likelihood,

$$p(\theta \mid r_{1:t}, \tilde{z}_{1:t}, \zeta, \varepsilon) \propto p(r_{1:t}, \tilde{z}_{1:t} \mid \theta, \zeta, \varepsilon).$$

The likelihood is then expressed by marginalizing over all possible sequences of latent interventions

$$\begin{aligned} p(r_{1:t}, \tilde{z}_{1:t} \mid \theta, \zeta, \varepsilon) &= \sum_{z_{1:t}} \prod_t p\,(z_t \mid \varepsilon)p(r_t, \tilde{z}_t \mid \theta, \zeta, \varepsilon, z_t) \\ &= \sum_{z_{1:t}} \prod_t p\,(z_t \mid \varepsilon)p(\tilde{z}_t \mid z_t)p(r_t \mid \theta, \zeta, z_t). \end{aligned}$$

where the second equality follows from the independence structure of the causal model.

The naïve approach of directly computing the summation over all possible $z_{1:t}$ is intractable because there are 2^t such sequences. Fortunately, such a summation is not necessary because the likelihood does not depend on the exact order of the previously seen data; it depends only on the number of times each possible (r, z) pair occurred. Here we show how such an expression can be derived. Note that we have numerically verified the accuracy of the derivation by comparing the derived likelihood to the simpler direct marginalization approach on short sequences.

We begin by replacing the $p(\tilde{z}_t \mid z_t)$ term with a boolean constraint on $z_{1:t}$. We first pull the term outside of the product:

$$p(r_{1:t}, \tilde{z}_{1:t} \mid \theta, \zeta, \varepsilon) = \sum_{z_{1:t}} p\left(\tilde{z}_{1:t} \mid z_{1:t}\right) \prod_t p\left(z_t \mid \varepsilon\right) p(r_t \mid \theta, \zeta, z_t).$$

Recall that $\tilde{z}_t$ is either equal to z_t or an unobserved value $\oslash$. Becuause the probability of observation does not depend on or influence θ, it will get absorbed in the normalizing constant. Thus, we have

$$p(r_{1:t}, \tilde{z}_{1:t} \mid \theta, \zeta, \varepsilon) \propto \sum_{z_{1:t}} \mathbf{1}\left(\forall t\colon (\tilde{z}_t = z_t) \vee (\tilde{z}_t = \oslash)\right) \prod_t p\left(z_t \mid \varepsilon\right) p(r_t \mid \theta, \zeta, z_t),$$

where $\mathbf{1}$ is the indicator function.

Next, we will simplify the product. We begin by re-expressing the vectors $r_{1:t}$ and $z_{1:t}$ as a list of t pairs, (r_t, z_t). Let $n_{r,z}$ denote the number of times the pair (r, z) occurs in this list, and let $p_{r,z} = p(z \mid \varepsilon) p(r \mid \theta, \zeta, z)$ be the probability of each pair type occurring. We can then rewrite the product as

$$\begin{aligned}
\prod_t p\left(z_t \mid \varepsilon\right) p(r_t \mid \theta, \zeta, z_t) &= \prod_t p_{r_t, z_t} \\
&= \prod_{i=1}^{n_{0,0}} p_{0,0} \cdot \prod_{i=1}^{n_{0,1}} p_{0,1} \cdot \prod_{i=1}^{n_{1,0}} p_{1,0} \cdot \prod_{i=1}^{n_{1,1}} p_{1,1} \\
&= p_{0,0}^{n_{0,0}} \cdot p_{0,1}^{n_{0,1}} \cdot p_{1,0}^{n_{1,0}} \cdot p_{1,1}^{n_{1,1}}
\end{aligned}$$

Next, we convert the constraint on $z_{1:t}$ (from $\tilde{z}_{1:t}$) to a constraint on $n_{r,z}$. Let $\tilde{n}_{r,\tilde{z}}$ be the number of times the pair $(r, \tilde{z})$ was observed (with $\tilde{z} \in \{0,1,\oslash\}$). Intuitively, the latent agent intervened on some of the unobserved trials, and not the rest. Let x_r be the number of times the agent intervened when reward r was received. We then have

$$\begin{aligned}
n_{r,0} &= \tilde{n}_{r,0} + \tilde{n}_{r,\oslash} - x_r \\
n_{r,1} &= \tilde{n}_{r,1} + x_r \\
0 &\leq x_r \leq \tilde{n}_{r,\oslash}
\end{aligned}$$

This gives us a constraint over $n_{r,z}$ given $\tilde{z}_{1:t}$. Finally, to account for the fact that there are multiple $z_{1:t}$ sequences consistent with each x_r, we will multiply the probability associated with each x_r by the number of such sequences, $\binom{\tilde{n}_{r,\emptyset}}{x_r}$. Putting it all together we arrive at

$$p(r_{1:t}, \tilde{z}_{1:t} \mid \theta, \zeta, \varepsilon) \propto \sum_{x_0=0}^{\tilde{n}_{0,\emptyset}} \binom{\tilde{n}_{0,\emptyset}}{x_0} p_{0,0}^{\tilde{n}_{0,0}+\tilde{n}_{0,\emptyset}-x_0} p_{0,1}^{\tilde{n}_{0,1}+x_0} \cdot \sum_{x_1=0}^{\tilde{n}_{1,\emptyset}} \binom{\tilde{n}_{1,\emptyset}}{x_1} p_{1,0}^{\tilde{n}_{1,0}+\tilde{n}_{1,\emptyset}-x_1} p_{1,1}^{\tilde{n}_{1,1}+x_1}.$$

Given this (unnormalized) likelihood, we estimate the normalizing constant by integrating the likelihood over $\theta \in [0,1]$. The posterior is simply the likelihood divided by the normalizing constant.

Note that we write $p(\theta \mid r_{1:t}, \tilde{z}_{1:t-1}, \zeta, \varepsilon)$ to denote the posterior before the latent agent has been revealed. This can be computed using the above equation by setting $\tilde{z}_t = \emptyset$.

APPENDIX D

Alternative Surprise Features:

- uRPE-MAP: unsigned reward prediction error, where the expected reward is the maximum a posteriori (MAP) estimate of θ. Note the MAP estimate is undefined on the first selection of each mine (as all values of θ are equally probable). On these trials, we set the expected reward to 0.5 (the mean posterior estimate of θ).
- uRPE-Full: unsigned reward prediction error, where the expected reward is the mean posterior estimate of θ weighed against the probability of intervention, $p(R_t = 1|r_{1:t-1})$.
- Surprisal: information-theoretic measure of the unlikeliness of the received reward (Shannon, 1948), used in prior research on infants' attention (Kidd et al., 2012; Poli et al., 2020). Defined as $-\log p(r_t|r_{1:t-1})$.

Alternative Global Features:

- Max entropy: uncertainty about which mine has the highest reward probability, measured by information-theoretic entropy (Shannon, 1948).
- Max EIG: expected reduction in max entropy after observing whether an intervention occurred on the current trial.

Alternative Value Features:

- Value of complete information (VCI): value of observing θ for all mines for choice task (VCI-choice) or prediction task (VCI-prediction)

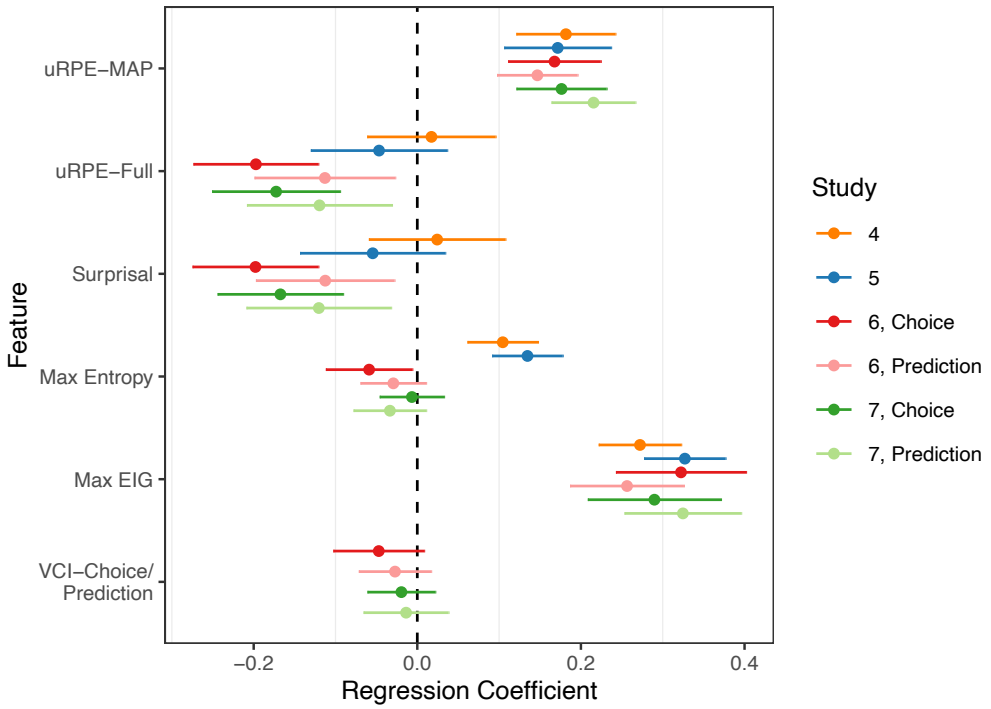


Figure D1 Association between alternative features and curiosity in individual regression models. Results for uRPE-MAP and Max EIG resemble results reported in Chapter 3 for uRPE and Global EIG, respectively. uRPE-Full, Surprisal, and Max Entropy are more variable across studies. VCI-Choice/Prediction is unrelated to curiosity.

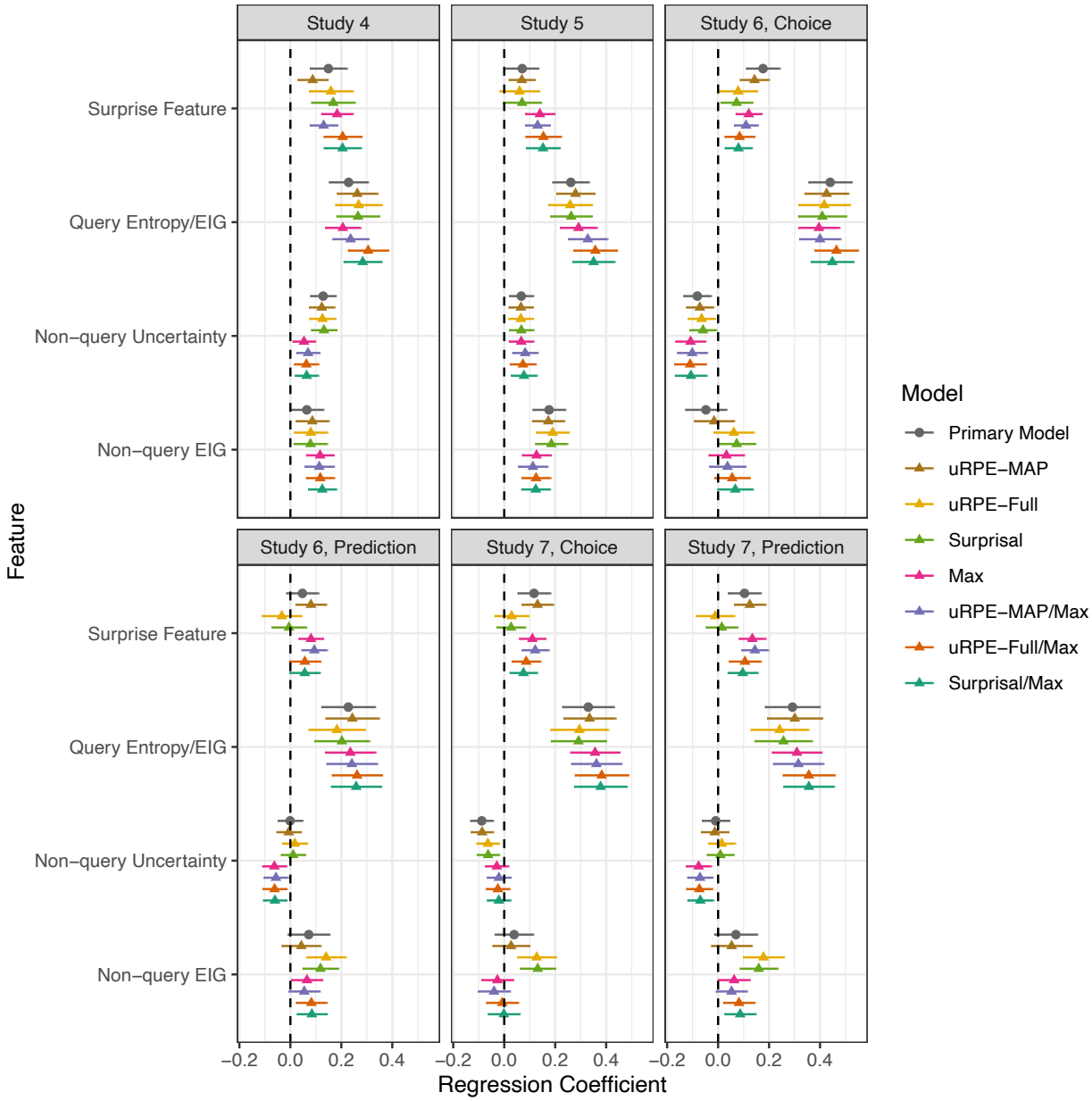


Figure D2 Results of multiple regression models with alternative epistemic features. "Primary Model" indicates the results reported in Chapter 3. All other models are indicated by the feature(s) that deviate from the primary model (e.g., uRPE-MAP includes the same features as the primary model, except uRPE-MAP is used as the surprise feature in place of uRPE). Results generally appear similar to those for the primary model, independent of which set of features are used.

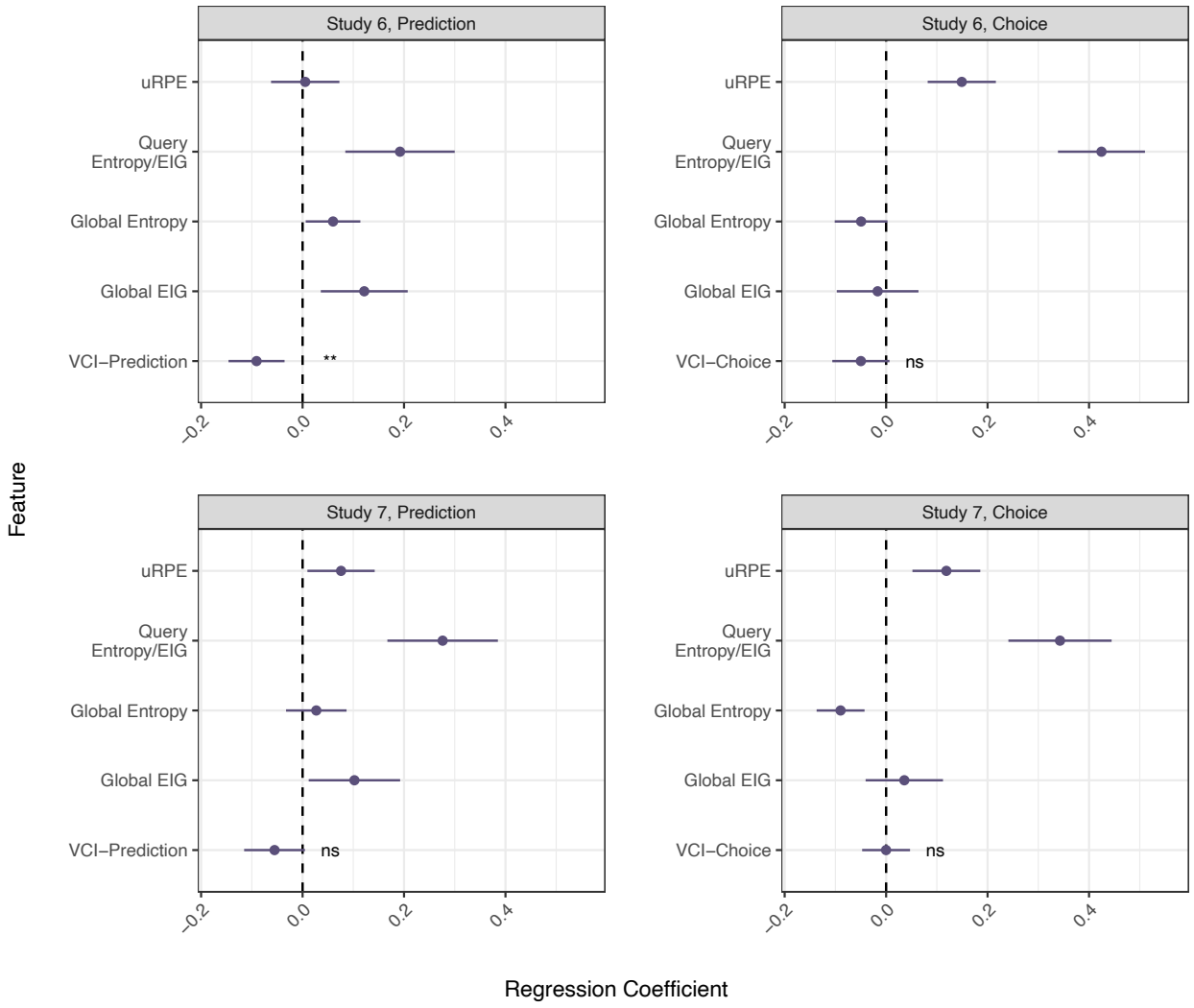


Figure D3 Added variance explained by alternative instrumental features. Each model (one per panel) included all four epistemic features (uRPE, query entropy/EIG, global entropy, and global EIG) and one instrumental feature (VCI-prediction/choice). The significance of each instrumental feature is indicated. *** $p < .001$; ** $p < .01$; * $p < .05$; ns = not significant. VCI-prediction explained unique variance in curiosity in the prediction condition of Study 6, such that participants were more curious on trials where learning the true value of all three mines would be *less* valuable for the prediction task.

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

Aguiar, N. R., Stoess, C. J., & Taylor, M. (2012). The development of children's ability to fill the gaps in their knowledge by consulting experts. *Child Development*, *83*(4), 1368–1381. https://doi.org/10.1111/j.1467-8624.2012.01782.x

Ahn, W.-K., Marsh, J. K., Luhmann, C. C., & Lee, K. (2002). Effect of theory-based feature correlations on typicality judgments. *Memory & Cognition*, *30*(1), 107–118. https://doi.org/10.3758/BF03195270

Ahn, W.-K., Novick, L. R., & Kim, N. S. (2003). Understanding behavior makes it more normal. *Psychonomic Bulletin & Review*, *10*(3), 746–752. https://doi.org/10.3758/BF03196541

Aleven, V. A. W. M. M., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, *26*(2), 147–179. https://doi.org/10.1207/s15516709cog2602_1

Anderson, C. A., Krull, D. S., & Weiner, B. (1996). Explanations: Processes and consequences. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 271–296). Guilford Press.

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

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology*, *95*(4), 774–783. https://doi.org/10.1037/0022-0663.95.4.774

Baranes, A. F., Oudeyer, P.-Y., & Gottlieb, J. (2014). The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration. *Frontiers in Neuroscience*, *8*, 317. https://doi.org/10.3389/fnins.2014.00317

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

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

Baum, L. A., Danovitch, J. H., & Keil, F. C. (2008). Children's sensitivity to circular explanations. *Journal of Experimental Child Psychology*, *100*(2), 146–155. https://doi.org/10.1016/j.jecp.2007.10.007

Bechlivanidis, C., Lagnado, D. A., Zemla, J. C., & Sloman, S. (2017). Concreteness and abstraction in everyday explanation. *Psychonomic Bulletin & Review*, *24*(5), 1451–1464. https://doi.org/10.3758/s13423-017-1299-3

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

Berlyne, D. E. (1950). Novelty and curiosity as determinants of exploratory behavior. *British Journal of Psychology. General Section*, *41*(1–2), 68–80. https://doi.org/10.1111/j.2044-8295.1950.tb00262.x

Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology*, *45*, 180–191. https://doi.org/10.1111/j.2044-8295.1954.tb01243.x

Berlyne, D. E. (1960). *Conflict, arousal, and curiosity.* McGraw-Hill. https://doi.org/10.1037/11164-000

Berlyne, D. E. (1962). Uncertainty and epistemic curiosity. *British Journal of Psychology*, *53*(1), 27–34. https://doi.org/10.1111/j.2044-8295.1962.tb00811.x

Berlyne, D. E. (1966). Curiosity and exploration. *Science*, *153*(3731), 25–33. https://doi.org/10.1126/science.153.3731.25

Berlyne, D. E. (1978). Curiosity and learning. *Motivation and Emotion*, *2*(2), 97–175. https://doi.org/10.1007/BF00993037

Berlyne, D. E., & Frommer, F. D. (1966). Some determinants of the incidence and content of children's questions. *Child Development*, *37*(1), 177–189. https://doi.org/10.2307/1126438

Bernacki, M. L., & Walkington, C. (2018). The role of situational interest in personalized learning. *Journal of Educational Psychology*, *110*(6), 864–881. https://doi.org/10.1037/edu0000250

Blanchard, T. C., Hayden, B. Y., & Bromberg-Martin, E. S. (2015). Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. *Neuron*, *85*(3), 602–614. https://doi.org/10.1016/j.neuron.2014.12.050

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

Blanco, N. J., & Sloutsky, V. (2019). *Systematic exploration and uncertainty dominate young children's choices*. PsyArXiv. https://doi.org/10.31234/osf.io/72sfx

Blanco, N. J., & Sloutsky, V. M. (2020). Attentional mechanisms drive systematic exploration in young children. *Cognition*, *202*, 104327. https://doi.org/10.1016/j.cognition.2020.104327

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

Bonawitz, E. B., Ullman, T. D., Bridgers, S., Gopnik, A., & Tenenbaum, J. B. (2019). Sticking to the evidence? A behavioral and computational case study of micro-theory change in the domain of magnetism. *Cognitive Science*, *43*(8), e12765. https://doi.org/10.1111/cogs.12765

Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*(4), 215–234. https://doi.org/10.1016/j.cogpsych.2011.12.002

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731. https://doi.org/10.1037/xlm0000061

Brod, G., & Breitwieser, J. (2019). Lighting the wick in the candle of learning: Generating a prediction stimulates curiosity. *Npj Science of Learning*, *4*, 17. https://doi.org/10.1038/s41539-019-0056-y

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

Brydevall, M., Bennett, D., Murawski, C., & Bode, S. (2018). The neural encoding of information prediction errors during non-instrumental information seeking. *Scientific Reports*, *8*, 6134. https://doi.org/10.1038/s41598-018-24566-x

Busch, J. T. A., & Legare, C. H. (2019). Using data to solve problems: Children reason flexibly in response to different kinds of evidence. *Journal of Experimental Child Psychology*, *183*, 172–188. https://doi.org/10.1016/j.jecp.2019.01.007

Callanan, M. A., Legare, C. H., Sobel, D. M., Jaeger, G. J., Letourneau, S., McHugh, S. R., Willard, A., Brinkman, A., Finiasz, Z., Rubio, E., Barnett, A., Gose, R., Martin, J. L., Meisner, R., & Watson, J. (2020). Exploration, explanation, and parent–child interaction in museums. *Monographs of the Society for Research in Child Development*, *85*(1). https://doi.org/10.1111/mono.12412

Callaway, F., Gul, S., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2018). Learning to select computations. *Uncertainty in Artificial Intelligence: Proceedings of the Thirty-Fourth Conference*. https://arxiv.org/pdf/1711.06892.pdf

Callaway, F., Rangel, A., & Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLOS Computational Biology*, *17*(3), e1008863. https://doi.org/10.1371/journal.pcbi.1008863

Carey, S. (1985). *Conceptual change in childhood*. MIT press.

Carvalho, P. F., Vales, C., Fausey, C. M., & Smith, L. B. (2018). Novel names extend for how long preschool children sample visual information. *Journal of Experimental Child Psychology*, *168*, 1–18. https://doi.org/10.1016/j.jecp.2017.12.002

Charpentier, C. J., Bromberg-Martin, E. S., & Sharot, T. (2018). Valuation of knowledge and ignorance in mesolimbic reward circuitry. *Proceedings of the National Academy of Sciences*, *115*(31), E7255–E7264. https://doi.org/10.1073/pnas.1800547115

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

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. https://doi.org/10.1207/s15516709cog1302_1

Chi, M. T., De Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477. https://doi.org/10.1016/0364-0213(94)90016-7

Chouinard, M. M. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, *72*(1), 1–129. https://doi.org/10.1111/j.1540-5834.2007.00412.x.

Cimpian, A., & Steinberg, O. D. (2014). The inherence heuristic across development: Systematic differences between children's and adults' explanations for everyday facts. *Cognitive Psychology*, *75*, 130–154. https://doi.org/10.1016/j.cogpsych.2014.09.001

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

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133. https://doi.org/10.1016/j.cogpsych.2015.02.004

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–349. https://doi.org/10.1016/j.cognition.2011.03.003

Corriveau, K. H., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science*, *12*(3), 426–437. https://doi.org/10.1111/j.1467-7687.2008.00792.x

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

Corriveau, K. H., Meints, K., & Harris, P. L. (2009). Early tracking of informant accuracy and inaccuracy. *British Journal of Developmental Psychology*, *27*(2), 331–342. https://doi.org/10.1348/026151008X310229

Craik, K. J. W. (1943). *The nature of explanation*. Cambridge University Press.

Croissant, Y. (2020). Estimation of Random Utility Models in R: The mlogit Package. *Journal of Statistical Software*, *95*(1), 1–41. https://doi.org/10.18637/jss.v095.i11

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

Danovitch, J. H., & Mills, C. M. (2018). Understanding when and how explanation promotes exploration. In M. M. Saylor & P. A. Ganea (Eds.), *Active learning from infancy to childhood: Social motivation, cognition, and linguistic mechanisms* (pp. 95–112). Springer International Publishing. https://doi.org/10.1007/978-3-319-77182-3_6

Daw, N. D., O’Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI)*, 761–768.

Dennis, S. A., Goodson, B. M., & Pearson, C. (2018). *MTurk Workers' Use of Low-Cost "Virtual Private Servers" to Circumvent Screening Methods: A Research Note* (SSRN Scholarly Paper ID 3233954). Social Science Research Network. https://papers.ssrn.com/abstract=3233954

Dorfman, H. M., Bhui, R., Hughes, B. L., & Gershman, S. J. (2019). Causal inference about good and bad outcomes. *Psychological Science*, *30*(4), 516–525. https://doi.org/10.1177/0956797619828724

Douven, I. (2013). Inference to the best explanation, Dutch books, and inaccuracy minimisation. *The Philosophical Quarterly*, *63*(252), 428–444. https://doi.org/10.1111/1467-9213.12032

Douven, I. (2020). The ecological rationality of explanatory reasoning. *Studies in History and Philosophy of Science Part A*, *79*, 1–14. https://doi.org/10.1016/j.shpsa.2019.06.004

Douven, I., & Mirabile, P. (2018). Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(11), 1792–1813. https://doi.org/10.1037/xlm0000545

Douven, I., & Schupbach, J. N. (2015a). Probabilistic alternatives to Bayesianism: The case of explanationism. *Frontiers in Psychology*, *6*, 459. https://doi.org/10.3389/fpsyg.2015.00459

Douven, I., & Schupbach, J. N. (2015b). The role of explanatory considerations in updating. *Cognition*, *142*, 299–311. https://doi.org/10.1016/j.cognition.2015.04.017

Douven, I., & Wenmackers, S. (2017). Inference to the best explanation versus Bayes's rule in a social setting. *The British Journal for the Philosophy of Science*, *68*(2), 535–570. https://doi.org/10.1093/bjps/axv025

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). If it's important, then I am curious: A value intervention to induce curiosity. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 282–288). Cognitive Science Society.

Dubey, R., Mehta, H., & Lombrozo, T. (2021). Curiosity is contagious: A social influence intervention to induce curiosity. *Cognitive Science*, *45*(2), e12937. https://doi.org/10.1111/cogs.12937

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

Fandakova, Y., & Gruber, M. J. (2021). States of curiosity and interest enhance memory differently in adolescents and in children. *Developmental Science*, *24*(1), e13005. https://doi.org/10.1111/desc.13005

Fastrich, G. M., Kerr, T., Castel, A. D., & Murayama, K. (2018). The role of interest in memory for trivia questions: An investigation with a large-scale database. *Motivation Science*, *4*(3), 227–250. https://doi.org/10.1037/mot0000087

Fitneva, S. A., Lam, N. H., & Dunfield, K. A. (2013). The development of children's information gathering: To look or to ask? *Developmental Psychology*, *49*(3), 533–542. https://doi.org/10.1037/a0031326

FitzGibbon, L., Moll, H., Carboni, J., Lee, R., & Dehghani, M. (2019). Counterfactual curiosity in preschool children. *Journal of Experimental Child Psychology*, *183*, 146–157. https://doi.org/10.1016/j.jecp.2018.11.022

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science*, *49*(2), 78–88. http://dx.doi.org/10.1037/cbs0000069

Foster, M. I., & Keane, M. T. (2015). Why some surprises are more surprising than others: Surprise as a metacognitive sense of explanatory difficulty. *Cognitive Psychology*, *81*, 74–116. https://doi.org/10.1016/j.cogpsych.2015.08.004

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

Galli, G., Sirota, M., Gruber, M. J., Ivanof, B. E., Ganesh, J., Materassi, M., Thorpe, A., Loaiza, V., Cappelletti, M., & Craik, F. I. M. (2018). Learning facts during aging: The benefits of curiosity. *Experimental Aging Research*, *44*(4), 311–328. https://doi.org/10.1080/0361073X.2018.1477355

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.

Gopnik, A. (2020). Childhood as a solution to explore-exploit tensions. *Philosophical Transactions of the Royal Society B*, *375*, 20190502. https://doi.org/10.1098/rstb.2019.0502

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. MIT Press.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085. https://doi.org/10.1037/a0028044

Gottlieb, J., Cohanpour, M., Li, Y., Singletary, N., & Zabeh, E. (2020). Curiosity, information demand and attentional priority. *Current Opinion in Behavioral Sciences*, *35*, 83–91. https://doi.org/10.1016/j.cobeha.2020.07.016

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*(12), 758–770. https://doi.org/10.1038/s41583-018-0078-0

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, 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

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716. https://doi.org/10.1037/a0017201

Grimm, S. R. (2008). Explanatory inquiry and the need for explanation. *The British Journal for the Philosophy of Science*, *59*(3), 481–497. https://doi.org/10.1093/bjps/axn021

Grimm, S. R. (2010). The goal of explanation. *Studies in History and Philosophy of Science Part A*, *41*(4), 337–344. https://doi.org/10.1016/j.shpsa.2010.10.006

Gruber, M. J., Gelman, B. D., & Ranganath, C. (2014). States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit. *Neuron*, *84*(2), 486–496. https://doi.org/10.1016/j.neuron.2014.08.060

Gruber, M. J., & Ranganath, C. (2019). How curiosity enhances hippocampus-dependent memory: The Prediction, Appraisal, Curiosity, and Exploration (PACE) framework. *Trends in Cognitive Sciences*, *23*(12), 1014–1025. https://doi.org/10.1016/j.tics.2019.10.003

Gweon, H., & Schulz, L. (2011). 16-month-olds rationally infer causes of failed actions. *Science*, *332*(6037), 1524. https://doi.org/10.1126/science.1204493

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066–9071. https://doi.org/10.1073/pnas.1003095107

Gwynne, N. Z., & Lombrozo, T. (2010). The cultural transmission of explanations: Evidence that teleological explanations are preferentially remembered. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1301–1306). Cognitive Science Society.

Harman, G. H. (1965). The inference to the best explanation. *The Philosophical Review*, *74*(1), 88–95. https://doi.org/10.2307/2183532

Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Harvard University Press.

Harris, P. L., & Koenig, M. A. (2006). Trust in testimony: How children learn about science and religion. *Child Development*, *77*(3), 505–524. https://doi.org/10.1111/j.1467-8624.2006.00886.x

Hay, N., Russell, S., Tolpin, D., & Shimony, S. E. (2012). Selecting computations: Theory and applications. In N. de Freitas & K. Murphy (Eds.), *Uncertainty in artificial intelligence: Proceedings of the twenty-eighth conference*. AUAI Press.

Heider, F. (1958). *The psychology of interpersonal relations*. John Wiley & Sons. https://doi.org/10.1037/10628-000

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135–175. https://doi.org/10.1086/286983

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

Hickling, A. K., & Wellman, H. M. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology*, *37*(5), 668–683. https://doi.org/10.1037/0012-1649.37.5.668

Higginson, A. D., McNamara, J. M., & Houston, A. I. (2016). Fatness and fitness: Exposing the logic of evolutionary explanations for obesity. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1822), 20152443. https://doi.org/10.1098/rspb.2015.2443

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46–54. https://doi.org/10.1016/j.tics.2014.10.004

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

Howard, R. A. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, *2*(1), 22–26. https://doi.org/10.1109/TSSC.1966.300074

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

Huang, F. L., & Cornell, D. G. (2016). Using multilevel factor analysis with clustered data: Investigating the factor structure of the Positive Values Scale. *Journal of Psychoeducational Assessment*, *34*(1), 3–14. https://doi.org/10.1177/0734282915570278

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306. https://doi.org/10.1016/j.visres.2008.09.007

Jepma, M., Verdonschot, R. G., van Steenbergen, H., Rombouts, S. A. R., & Nieuwenhuis, S. (2012). Neural mechanisms underlying the induction and relief of perceptual curiosity. *Frontiers in Behavioral Neuroscience*, *6*, 5. https://doi.org/10.3389/fnbeh.2012.00005

Jirout, J., & Klahr, D. (2012). Children's scientific curiosity: In search of an operational definition of an elusive concept. *Developmental Review*, *32*(2), 125–160. https://doi.org/10.1016/j.dr.2012.04.002

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*, 101222. https://doi.org/10.1016/j.cogpsych.2019.05.004

Johnston, A. M., Johnson, S. G. B., Koven, M. L., & Keil, F. C. (2017). Little Bayesians or little Einsteins? Probability and explanatory virtue in children's inferences. *Developmental Science*, *20*(6), e12483. https://doi.org/10.1111/desc.12483

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

Johnston, A. M., Sheskin, M., & Keil, F. C. (2019). Learning the relevance of relevance and the trouble with truth: Evaluating explanatory relevance across childhood. *Journal of Cognition and Development*, *20*(4), 555–572. https://doi.org/10.1080/15248372.2019.1631167

Jones, A., Swaboda, N., & Ruggeri, A. (2020). Developmental changes in question-asking. In L. P. Butler, S. Ronfard, & K. H. Corriveau (Eds.), *The Questioning Child: Insights from Psychology and Education* (pp. 118–143). Cambridge University Press.

Kachergis, G., Rhodes, M., & Gureckis, T. (2017). Desirable difficulties during the development of active inquiry skills. *Cognition*, *166*, 407–417. https://doi.org/10.1016/j.cognition.2017.05.021

Kahan, D. M., Landrum, A., Carpenter, K., Helft, L., & Hall Jamieson, K. (2017). Science curiosity and political information processing. *Political Psychology*, *38*(S1), 179–199.

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science*, *20*(8), 963–973. https://doi.org/10.1111/j.1467-9280.2009.02402.x

Karmiloff-Smith, A., & Inhelder, B. (1974). If you want to get ahead, get a theory. *Cognition*, *3*(3), 195–212. https://doi.org/10.1016/0010-0277(74)90008-0

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7*(8), 368–373. https://doi.org/10.1016/S1364-6613(03)00158-X

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*(1), 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

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

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107–128. https://doi.org/10.1037/h0034225

Kerrod, R., Madgwick, W., Reed, S., Collins, F., & Brooks, P. (2006). *1000 questions & answers factfile*. Kingfisher.

Khemlani, S. S., & Johnson-Laird, P. N. (2011). The need to explain. *The Quarterly Journal of Experimental Psychology*, *64*(11), 2276–2288. https://doi.org/10.1080/17470218.2011.592593

Khemlani, S. S., & Johnson-Laird, P. N. (2012). Hidden conflicts: Explanations make inconsistencies harder to detect. *Acta Psychologica*, *139*(3), 486–491. https://doi.org/10.1016/j.actpsy.2012.01.010

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

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449–460. https://doi.org/10.1016/j.neuron.2015.09.010

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE*, *7*(5), e36399. https://doi.org/10.1371/journal.pone.0036399

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.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211–228. https://doi.org/10.1037/0033-295X.94.2.211

Knox, W. B., Otto, A. R., Stone, P., & Love, B. (2012). The Nature of Belief-Directed Exploratory Choice in Human Decision-Making. *Frontiers in Psychology*, *2*, 398. https://doi.org/10.3389/fpsyg.2011.00398

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

Kon, E., & Lombrozo, T. (in prep). *Why explainers take exceptions to exceptions*.

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

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

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

Lastovicka, J. L., & Thamodaran, K. (1991). Common factor score estimates in multiple regression problems. *Journal of Marketing Research*, 105–112. https://doi.org/10.2307/3172730

Lau, J. K. L., Ozono, H., Kuratomi, K., Komiya, A., & Murayama, K. (2020). Shared striatal activity in decisions to satisfy curiosity and hunger at the risk of electric shocks. *Nature Human Behaviour*, *4*, 531–543. https://doi.org/10.1038/s41562-020-0848-3

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, *12*(2), 164–174. https://doi.org/10.1016/j.cogsys.2010.07.007

Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Development*, *83*(1), 173–185. https://doi.org/10.1111/j.1467-8624.2011.01691.x

Legare, C. H., Evans, E. M., Rosengren, K. S., & Harris, P. L. (2012). The coexistence of natural and supernatural explanations across cultures and development. *Child Development*, *83*(3), 779–793. https://doi.org/10.1111/j.1467-8624.2012.01743.x

Legare, C. H., & Gelman, S. A. (2008). Bewitchment, biology, or both: The co-existence of natural and supernatural explanatory frameworks across development. *Cognitive Science*, *32*(4), 607–642. https://doi.org/10.1080/03640210802066766

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, *126*, 198–212. https://doi.org/10.1016/j.jecp.2014.03.001

Legare, C. H., Mills, C. M., Souza, A. L., Plummer, L. E., & Yasskin, R. (2013). The use of questions as problem-solving strategies during early childhood. *Journal of Experimental Child Psychology*, *114*(1), 63–76. https://doi.org/10.1016/j.jecp.2012.07.002

Legare, C. H., Wen, N. J., Herrmann, P. A., & Whitehouse, H. (2015). Imitative flexibility and the development of cultural learning. *Cognition*, *142*, 351–361. https://doi.org/10.1016/j.cognition.2015.05.020

Ligneul, R., Mermillod, M., & Morisseau, T. (2018). From relief to surprise: Dual control of epistemic curiosity in the human brain. *NeuroImage*, *181*, 490–500. https://doi.org/10.1016/j.neuroimage.2018.07.038

Lim, J. B., & Oppenheimer, D. M. (2020). Explanatory preferences for complexity matching. *PLoS ONE*, *15*(4), e0230929. https://doi.org/10.1371/journal.pone.0230929

Lin, L.-M., Moore, D., & Zabrucky, K. M. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology*, *22*(2), 111–128. https://doi.org/10.1080/02702710119125

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

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005. https://doi.org/10.1214/aoms/1177728069

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). Routledge.

Liquin, E. G., Callaway, F., & Lombrozo, T. (2020). Quantifying curiosity: A formal approach to dissociating causes of curiosity. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 309–315). Cognitive Science Society.

Liquin, E. G., & Gopnik, A. (2020). *Children are more exploratory and learn more than adults in an approach-avoid task*. PsyArXiv. https://doi.org/10.31234/osf.io/6watc

Liquin, E. G., & Lombrozo, T. (2020). A functional approach to explanation-seeking curiosity. *Cognitive Psychology*, *119*, 101276. https://doi.org/10.1016/j.cogpsych.2020.101276

Liquin, E. G., & Lombrozo, T. (2018). Determinants and consequences of the need for explanation. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 696–701). Cognitive Science Society.

Litman, J. A. (2008). Interest and deprivation factors of epistemic curiosity. *Personality and Individual Differences*, *44*(7), 1585–1595. https://doi.org/10.1016/j.paid.2008.01.014

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

Loewenstein, G., Adler, D., Behrens, D., & Gillis, J. (1992). *Why Pandora opened the box: Curiosity as a desire for missing information* [Unpublished Manuscript].

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. (2009). Explanation and categorization: How "why?" informs "what?" *Cognition*, *110*(2), 248–253. https://doi.org/10.1016/j.cognition.2008.10.007

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass*, *6*(8), 539–551. https://doi.org/10.1111/j.1747-9991.2011.00413.x

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

Lombrozo, T., Bonawitz, E. B., & Scalise, N. R. (2018). Young children's learning and generalization of teleological and mechanistic explanations. *Journal of Cognition and Development*, *19*(2), 220–232. https://doi.org/10.1080/15248372.2018.1427099

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204. https://doi.org/10.1016/j.cognition.2004.12.009

Lombrozo, T., & Gwynne, N. Z. (2014). Explanation and inference: Mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, *8*, 700. https://doi.org/10.3389/fnhum.2014.00700

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

Lyons, K. E., & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child Development*, *82*(6), 1778–1787. https://doi.org/10.1111/j.1467-8624.2011.01649.x

Maguire, P., Moser, P., Maguire, R., & Keane, M. T. (2018). Seeing patterns in randomness: A computational model of surprise. *Topics in Cognitive Science*, *11*, 103–118. https://doi.org/10.1111/tops.12345

Maguire, R., Maguire, P., & Keane, M. T. (2011). Making sense of surprise: An investigation of the factors influencing surprise judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 176–186. https://doi.org/10.1037/a0021609

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

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT Press.

Markant, D. B., & Gureckis, T. (2014a). A preference for the unpredictable over the informative during self-directed learning. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (Issue 36, pp. 958–963). Cognitive Science Society.

Markant, D. B., & Gureckis, T. (2012). Does the utility of information influence sampling behavior? In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (Issue 34, pp. 719–724). Cognitive Science Society.

Markant, D. B., & Gureckis, T. M. (2014b). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94–122. https://doi.org/10.1037/a0032108

Markant, D. B., Settles, B., & Gureckis, T. M. (2015). Self-directed learning favors local, rather than global, uncertainty. *Cognitive Science*, *40*(1), 100–120. https://doi.org/10.1111/cogs.12220

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

Matheson, J. E. (1968). The economic value of analysis and computation. *IEEE Transactions on Systems Science and Cybernetics*, *4*(3), 325–332. https://doi.org/10.1109/TSSC.1968.300126

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*, *23*(2), 209–237. https://doi.org/10.1080/02699930802204677

McCormack, T., Bramley, N., Frosch, C., Patrick, F., & Lagnado, D. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology*, *141*, 1–22. https://doi.org/10.1016/j.jecp.2015.06.017

McGillivray, S., Murayama, K., & Castel, A. D. (2015). Thirst for knowledge: The effects of curiosity and interest on memory in younger and older adults. *Psychology and Aging*, *30*(4), 835–841. https://doi.org/10.1037/a0039801

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*(2), 119–148.

Meder, B., Nelson, J. D., Jones, M., & Ruggeri, A. (2019). Stepwise versus globally optimal search in children and adults. *Cognition*, *191*, 103965. https://doi.org/10.1016/j.cognition.2019.05.002

Meder, B., Wu, C. M., Schulz, E., & Ruggeri, A. (2021). Development of directed and random exploration in children. *Developmental Science*, e13095. https://doi.org/10.1111/desc.13095

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, *31*(5), 838–850. https://doi.org/10.1037/0012-1649.31.5.838

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., Danovitch, J. H., Rowles, S. P., & Campbell, I. L. (2017). Children's success at detecting circular explanations and their interest in future learning. *Psychonomic Bulletin & Review*, *24*(5), 1465–1477. https://doi.org/10.3758/s13423-016-1195-2

Mills, C. M., Legare, C. H., Bills, M., & Mejias, C. (2010). Preschoolers use questions as a tool to acquire knowledge from different sources. *Journal of Cognition and Development*, *11*(4), 533–560. https://doi.org/10.1080/15248372.2010.516419

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), e12706. https://doi.org/10.1111/cogs.12706

Mullaney, K. M., Carpenter, S. K., Grotenhuis, C., & Burianek, S. (2014). Waiting for feedback helps if you want to know the answer: The role of curiosity in the delay-of-feedback benefit. *Memory & Cognition*, *42*(8), 1273–1284. https://doi.org/10.3758/s13421-014-0441-y

Munnich, E. L., Foster, M. I., & Keane, M. T. (2019). Editors’ introduction and review: An appraisal of surprise: Tracing the threads that stitch it together. *Topics in Cognitive Science*, *11*(1), 37–49. https://doi.org/10.1111/tops.12402

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

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 904–919. https://doi.org/10.1037/0278-7393.20.4.904

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316. https://doi.org/10.1037/0033-295X.92.3.289

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77. https://doi.org/10.1016/j.cogpsych.2016.01.001

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120–134. https://doi.org/10.1037/a0021110

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979–999. https://doi.org/10.1037/0033-295X.112.4.979

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*(1), 74–80. https://doi.org/10.1016/j.cognition.2013.09.007

Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience Matters: Information Acquisition Optimizes Probability Gain. *Psychological Science*, *21*(7), 960–969. https://doi.org/10.1177/0956797610372637

Nguyen, C. T. (2021). The seductions of clarity. *Royal Institute of Philosophy Supplement*, *89*, 227–255. https://doi.org/10.1017/S1358246121000035

Noordewier, M. K., & Dijk, E. van. (2017). Curiosity and time: From not knowing to almost knowing. *Cognition and Emotion*, *31*(3), 411–421. https://doi.org/10.1080/02699931.2015.1122577

Nussenbaum, K., Cohen, A. O., Davis, Z. J., Halpern, D. J., Gureckis, T. M., & Hartley, C. A. (2020). Causal information-seeking strategies change across childhood and adolescence. *Cognitive Science*, *44*(9), e12888. https://doi.org/10.1111/cogs.12888

Nyrup, R. (2015). How explanatory reasoning justifies pursuit: A Peircean view of IBE. *Philosophy of Science*, *82*(5), 749–760. https://doi.org/10.1086/683262

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631. https://doi.org/10.1037/0033-295X.101.4.608

Oudeyer, P.-Y. (2018). Computational theories of curiosity-driven learning. In G. Gordon (Ed.), *The New Science of Curiosity* (pp. 43–72). Nova Science Publishers.

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

Patalano, A. L., Chin-Parker, S., & Ross, B. H. (2006). The importance of being coherent: Category coherence, cross-classification, and reasoning. *Journal of Memory and Language*, *54*(3), 407–424. https://doi.org/10.1016/j.jml.2005.10.005

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. *Journal of Personality and Social Psychology*, *62*(2), 189–206. https://doi.org/10.1037/0022-3514.62.2.189

Perez, J., & Feigenson, L. (2020). *Violations of expectation trigger infants to search for explanations* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/eahjd

Poli, F., Serino, G., Mars, R. B., & Hunnius, S. (2020). Infants tailor their attention to maximize learning. *Science Advances*, *6*(39), eabb5053. https://doi.org/10.1126/sciadv.abb5053

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

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, *130*(3), 323–360. https://doi.org/10.1037/0096-3445.130.3.323

Rehder, B., & Hastie, R. (2004). Category coherence and category-based property induction. *Cognition*, *91*(2), 113–153. https://doi.org/10.1016/S0010-0277(03)00167-7

Reisenzein, R. (2000). Exploring the strength of association between the components of emotion syndromes: The case of surprise. *Cognition & Emotion*, *14*(1), 1–38. https://doi.org/10.1080/026999300378978

Rittle-Johnson, B., Loehr, A. M., & Durkin, K. (2017). Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM*, *49*(4), 599–611. https://doi.org/10.1007/S11858-017-0834-Z

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

Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*(1), 69–89. https://doi.org/10.1007/s42113-018-0005-5

Rouhani, N., & Niv, Y. (2021). Signed and unsigned reward prediction errors dynamically enhance learning and memory. *ELife*, *10*, e61077. https://doi.org/10.7554/eLife.61077

Rouhani, N., Norman, K. A., & Niv, Y. (2018). Dissociable effects of surprising rewards on learning and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(9), 1430–1443. https://doi.org/10.1037/xlm0000518

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

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*, 203–216. https://doi.org/10.1016/j.cognition.2015.07.004

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2016). Sources of developmental change in the efficiency of information search. *Developmental Psychology*, *52*(12), 2159–2173. https://doi.org/10.1037/dev0000240

Ruggeri, A., Sim, Z. L., & Xu, F. (2017). "Why is Toma late to school again?" Preschoolers identify the most informative questions. *Developmental Psychology*, *53*(9), 1620–1632. https://doi.org/10.1037/dev0000340

Ruggeri, A., Xu, F., & Lombrozo, T. (2019). Effects of explanation on children's question asking. *Cognition*, *191*, 103966. https://doi.org/10.1016/j.cognition.2019.05.003

Salmon, W. C. (1989). *Four decades of scientific explanation*. University of Minnesota Press. https://doi.org/10.2307/j.ctt5vkdm7

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

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (2019). Searching for rewards like a child means less generalization and more directed exploration. *Psychological Science*, *30*(11), 1561–1572. https://doi.org/10.1177/0956797619863663

Schulz, L. E. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*(7), 382–389. https://doi.org/10.1016/j.tics.2012.06.004

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*(4), 1045–1050. https://doi.org/10.1037/0012-1649.43.4.1045

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, *40*(2), 162–176. https://doi.org/10.1037/0012-1649.40.2.162

Schwitzgebel, E. (1999). Children's theories and the drive to explain. *Science & Education*, *8*(5), 457–488. https://doi.org/10.1023/A:1008633709447

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shtulman, A., & Lombrozo, T. (2016). Bundles of contradiction: A coexistence view of conceptual change. In D. Barner & A. S. Baron (Eds.), *Core Knowledge and Conceptual Change* (pp. 53–71). Oxford University Press.

Sim, Z. L., & Xu, F. (2017). Infants preferentially approach and explore the unexpected. *British Journal of Developmental Psychology*, *35*(4), 596–608. https://doi.org/10.1111/bjdp.12198

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. H. Spieler & E. Schumacher (Eds.), *New Methods in Cognitive Psychology*. Psychology Press.

Smock, C. D., & Holt, B. G. (1962). Children's reactions to novelty: An experimental study of "curiosity motivation." *Child Development*, *33*(3), 631–642. https://doi.org/10.2307/1126663

Somerville, L. H., Sasse, S. F., Garrad, M. C., Drysdale, A. T., Abi Akar, N., Insel, C., & Wilson, R. C. (2017). Charting the expansion of strategic exploratory behavior during adolescence. *Journal of Experimental Psychology: General*, *146*(2), 155–164. https://doi.org/10.1037/xge0000250

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and Exploration in a Restless Bandit Problem. *Topics in Cognitive Science*, *7*(2), 351–367. https://doi.org/10.1111/tops.12145

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4444–4451.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*(4), 605–632. https://doi.org/10.1037/0033-295X.99.4.605

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94. https://doi.org/10.1126/science.aaa3799

Stahl, A. E., & Feigenson, L. (2017). Expectancy violations promote learning in young children. *Cognition*, *163*, 1–14. https://doi.org/10.1016/j.cognition.2017.02.008

Stare, C. J., Gruber, M. J., Nadel, L., Ranganath, C., & Gómez, R. L. (2018). Curiosity-driven memory enhancement persists over time but does not benefit from post-learning sleep. *Cognitive Neuroscience*, *9*(3–4), 100–115. https://doi.org/10.1080/17588928.2018.1513399

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179. https://doi.org/10.1016/j.jmp.2008.11.002

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489. https://doi.org/10.1016/S0364-0213(03)00010-7

Stojić, H., Orquin, J. L., Dayan, P., Dolan, R. J., & Speekenbrink, M. (2020). Uncertainty in learning, choice, and visual fixation. *Proceedings of the National Academy of Sciences*, *117*(6), 3291–3300. https://doi.org/10.1073/pnas.1911348117

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

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, *44*(3), 510–515. https://doi.org/10.1016/j.shpsa.2012.12.005

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

Subbotsky, E. (2010). Curiosity and exploratory behaviour towards possible and impossible events in children and adults. *British Journal of Psychology*, *101*(3), 481–501. https://doi.org/10.1348/000712609X470590

Sumner, E., Li, A. X., Perfors, A., Hayes, B., Navarro, D., & Sarnecka, B. W. (2019). *The Exploration Advantage: Children's instinct to explore allows them to find information that adults miss.* PsyArXiv. https://doi.org/10.31234/osf.io/h437v

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). MIT Press.

Ten, A., Kaushik, P., Oudeyer, P.-Y., & Gottlieb, J. (2020). *Humans monitor learning progress in curiosity-driven exploration*. PsyArXiv. https://doi.org/10.31234/osf.io/7dbr6

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

Thomson, G. H. (1951). *The factorial analysis of human ability*. London University Press.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

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*, 100802. https://doi.org/10.1016/j.cogdev.2019.100802

van Lieshout, L. L., Vandenbroucke, A. R., Müller, N. C., Cools, R., & de Lange, F. P. (2018). Induction and relief of curiosity elicit parietal and frontal activity. *Journal of Neuroscience*, *38*(10), 2579–2588. https://doi.org/10.1523/JNEUROSCI.2816-17.2018

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable Causal Relationships are Better Causal Relationships. *Cognitive Science*, *42*(4), 1265–1296. https://doi.org/10.1111/cogs.12605

Vasilyeva, N., Wilkenfeld, D., & Lombrozo, T. (2017). Contextual utility affects the perceived quality of explanations. *Psychonomic Bulletin & Review*, *24*, 1436–1450. https://doi.org/10.3758/s13423-017-1275-y

Vicario, M. D., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559. https://doi.org/10.1073/pnas.1517441113

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.

Vogl, E., Pekrun, R., Murayama, K., & Loderer, K. (2020). Surprised–curious–confused: Epistemic emotions and knowledge exploration. *Emotion*, *20*(4), 625–641. https://doi.org/10.1037/emo0000578

Vogl, E., Pekrun, R., Murayama, K., Loderer, K., & Schubert, S. (2019). Surprise, curiosity, and confusion promote knowledge exploration: Evidence for robust effects of epistemic emotions. *Frontiers in Psychology*, *10*, 2474. https://doi.org/10.3389/fpsyg.2019.02474

Von Stumm, S., Hell, B., & Chamorro-Premuzic, T. (2011). The hungry mind: Intellectual curiosity is the third pillar of academic performance. *Perspectives on Psychological Science*, *6*(6), 574–588. https://doi.org/10.1177/1745691611421204

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

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

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, *82*(1), 27–58. https://doi.org/10.1016/S0010-0277(01)00141-X

Walker, C. M., Bonawitz, E. B., & Lombrozo, T. (2017). Effects of explaining on children's preference for simpler hypotheses. *Psychonomic Bulletin & Review*, *24*(5), 1538–1547. https://doi.org/10.3758/s13423-016-1144-0

Walker, C. M., & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition*, *167*, 266–281. https://doi.org/10.1016/j.cognition.2016.11.007

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*(2), 343–357. https://doi.org/10.1016/j.cognition.2014.07.008

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. *Child Development*, *88*(1), 229–246. https://doi.org/10.1111/cdev.12590

Wang, M. Z., & Hayden, B. Y. (2019). Monkeys are curious about counterfactual outcomes. *Cognition*, *189*, 1–10. https://doi.org/10.1016/j.cognition.2019.03.009

Weiner, B. (1985a). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548–573. https://doi.org/10.1037/0033-295X.92.4.548

Weiner, B. (1985b). "Spontaneous" causal thinking. *Psychological Bulletin*, *97*(1), 74–84. http://dx.doi.org/10.1037/0033-2909.97.1.74

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.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*(1), 337–375. https://doi.org/10.1146/annurev.ps.43.020192.002005

Wiggin, K. L., Reimann, M., & Jain, S. P. (2019). Curiosity tempts indulgence. *Journal of Consumer Research*, *45*(6), 1194–1212. https://doi.org/10.1093/jcr/ucy055

Wilkenfeld, D. A. (2019). Understanding as compression. *Philosophical Studies*, *176*, 2807–2831. https://doi.org/10.1007/s11098-018-1152-1

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. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*(5), 776–806. https://doi.org/10.1111/j.1551-6709.2010.01113.x

Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology*, *66*(1), 55–84. https://doi.org/10.1016/j.cogpsych.2012.09.002

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

Wittmann, B. C., Daw, N. D., Seymour, B., & Dolan, R. J. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron*, *58*(6), 967–973. https://doi.org/10.1016/j.neuron.2008.04.027

Wong, P. T., & Weiner, B. (1981). When people ask" why" questions, and the heuristics of attributional search. *Journal of Personality and Social Psychology*, *40*(4), 650. https://doi.org/10.1037/0022-3514.40.4.650

Wong, W., & Yudell, Z. (2015). A normative account of the need for explanation. *Synthese*, *192*(9), 2863–2885. https://doi.org/10.1007/s11229-015-0690-8

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, *100*(1), 221–228. https://doi.org/10.1093/biomet/ass048

Woodward, J. F. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924. https://doi.org/10.1038/s41562-018-0467-4

Wu, Y., & Gweon, H. (2020). *Preschool-aged children jointly consider others' emotional expressions and prior knowledge to decide when to explore*. PsyArXiv. https://doi.org/10.31234/osf.io/ckh6j

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012–5015. https://doi.org/10.1073/pnas.0704450105

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