

ADVANCING RESPONSIBLE AI WITH HUMAN-CENTERED EVALUATION

SUNNIE S. Y. KIM

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
COMPUTER SCIENCE
ADVISER: OLGA RUSSAKOVSKY

MAY 2025

© Copyright by Sunnie S. Y. Kim, 2025.

All rights reserved.

Abstract

As Artificial Intelligence (AI) technologies are increasingly transforming how we live, work, and communicate, AI evaluation must take a human-centered approach to realistically reflect real-world performance and impact. In this dissertation, I present research that advances human-centered evaluation, and subsequently responsible development and deployment of AI, by integrating knowledge and methods from AI and Human-Computer Interaction (HCI). First, using explainable AI as an example, I discuss the challenges and necessity of *human* (as opposed to automatic) evaluation and present human evaluation studies that identify new issues and opportunities in explainable AI research. Next, I bring *context* into the picture and present contextualized studies with real users that provide a holistic and nuanced understanding of users' explainability needs and trust in AI. Then, I illustrate the importance of evaluating human-AI *interaction* through studies that examine how users' trust, reliance, and other perceptions and behaviors are shaped by different features of AI technologies. In closing, I discuss the implications of these work, provide recommendations for human-centered evaluation, and describe future directions for responsible AI.

Acknowledgments

First and foremost, I would like to thank Olga Russakovsky for being an incredibly kind and supportive adviser. Thank you for creating an environment where I could do things I value, such as interdisciplinary research and DEI work, without having to worry about other things. I owe a great deal of my success and happiness to you.

I have been fortunate to have a number of additional truly amazing mentors. Andrés Monroy-Hernández, thank you for introducing me to HCI and training me to be a better researcher. I will always be grateful. Jenn Wortman Vaughan, the second half of this dissertation would not exist without your guidance. Thank you for everything. I aspire to be a researcher and mentor like you some day. Vera Liao, thank you for all the valuable research and career advice. I cannot stress how much I learned from you. Your future students are extremely lucky to have you as an adviser.

I am also grateful to Ruth Fong for being a big sister-like figure that I can always go to for advice, Parastoo Abtahi for being on my committee and providing thoughtful feedback, Elizabeth Anne Watkins for hand-holding me through my very first interview study, Tania Lombrozo for the immensely fun and intellectual discussions, and Mark Riedl for his advice and support for my career growth. I am grateful as well to John Lafferty, Jay Emerson, Gregory Shakhnarovich, and Woo-Kyoung Anh for their mentorship during my undergraduate and post-baccalaureate years. Thank you for believing in me and encouraging me to pursue a PhD.

Many of my research papers and presentations were supported by the NSF Graduate Research Fellowship, the Siebel Scholars Award, and travel awards from the Princeton School of Engineering and Applied Science, SIGCHI, FAccT, and the Women in Computer Vision Workshop. I also benefited from various mentoring programs such as the CHI Doctoral Consortium, the Georgia Tech Doctoral Consortium on Responsible Computing, AI, and Society, and the Rising Stars in EECS Workshop. I am very appreciative and hope to pay forward the help and guidance I received.

Now onto my friends! Thanks to members of the Princeton Visual AI Lab, the Princeton HCI Lab, and the MSR FATE & CSS internship cohort for the camaraderie. Special shout-outs to Vikram, Nicole, Sharon, Rohan, Indu, Allison, Amaya, and Amna who I have had the joy of working together. My fellow [redacted] club members Vivien and Angelina, thanks for your friendship and all the hilarious group chats. Grad school has been a thousand times more fun because of you two.

Thanks to my HCI friends outside of Princeton: Samantha for helping me navigate my first CHI, Wesley for introducing me to all of his friends, Zana for being my job market buddy, and countless others for welcoming me into the community. Thanks to friends who made Princeton feel more like home: Yeji for being my emotional support friend who I can talk to about anything and everything, Kun Woo for introducing me to friends and places in Princeton, and Esther, Junhwan, Seungwook, to name a few, for being my tennis, food and/or ranting buddies. Thanks also to my faraway friends Jhaeun, Jingyeong, Jaeyoon, So-Hyeon, Yeonwoo, and Jiyoona for always being there for me no matter where we are in the world.

I moved to the US alone at the age of 19 for college. This was only possible because my parents and grandparents were big supporters of my studies and wanted me to get the best education possible. I wouldn't be here today without them.

My final thanks goes to my partner Kyungjoo Noh who has been promoted from a friend of a friend to boyfriend to domestic partner to husband over the past eight years. In his dissertation acknowledgments [289], Kyungjoo thanked me: *"Doing research has sometimes been unavoidably stressful, but it has never been difficult thanks to her love and support. In fact, she has made doing a PhD so much fun that I am even willing to repeat it infinitely many times if it is with her. (Of course, I would not otherwise!)"* I can happily say the same (although my preference would still be doing a PhD once with him by my side). Thank you for all the love, laughter, and happiness. I am so excited about what lies ahead for us!

To my partner Kyungjoo Noh

Contents

Abstract	3
Acknowledgments	4
List of Tables	13
List of Figures	17
1 Introduction	1
2 Evaluating the usefulness of AI explanations	4
2.1 Introduction	5
2.2 Related work	8
2.2.1 Interpretability landscape in computer vision	8
2.2.2 Evaluating heatmaps	9
2.2.3 Evaluating interpretable-by-design models	10
2.2.4 Evaluating explanations with human studies	10
2.3 HIVE design principles	11
2.3.1 Falsifiable hypothesis testing	12
2.3.2 Cross-method comparison	14
2.3.3 Human-centered evaluation	16
2.3.4 Generalizability & Scalability	17
2.4 HIVE study design	17
2.5 HIVE study results	19

2.5.1	Experimental details	19
2.5.2	The issue of confirmation bias	20
2.5.3	Objective assessment of interpretability	21
2.5.4	A closer examination of prototype-based models	23
2.5.5	Subjective evaluation of interpretability	24
2.5.6	Interpretability-accuracy tradeoff	24
2.6	Follow-up study on concept-based explanations	25
2.6.1	Human study design	26
2.6.2	Key findings from the human studies	28
2.7	Conclusion	30
3	Understanding how explainability can support human-AI interaction	32
3.1	Introduction	33
3.2	Related work	36
3.2.1	From algorithm-centered to human-centered XAI	36
3.2.2	Understanding end-users' XAI needs	37
3.2.3	XAI's role in human-AI collaboration	38
3.2.4	XAI methods for computer vision	40
3.3	Study application: Merlin Bird ID	42
3.4	Methods	42
3.4.1	Participant recruitment and selection	42
3.4.2	Study instrument	44
3.4.3	Conducting and analyzing interviews	47
3.5	Results	49
3.5.1	XAI needs	49
3.5.2	XAI uses	53
3.5.3	XAI perceptions	58

3.6	Discussion	65
3.6.1	XAI as a medium for improving human-AI collaboration	65
3.6.2	XAI design for Merlin	66
3.6.3	Implications for future XAI research	67
3.7	Limitations and future work	69
3.8	Conclusion	69
4	Understanding end-users' trust in a real-world AI application	71
4.1	Introduction	72
4.2	Background and related work	75
4.2.1	Definitions and models of trust in AI	75
4.2.2	Empirical studies of trust in AI	77
4.3	Methods	78
4.3.1	Study application: Merlin Bird ID app	79
4.3.2	Participant recruitment and selection	79
4.3.3	Interview sessions and analysis	80
4.4	Results	82
4.4.1	Prelude: Is “trust” the right term?	82
4.4.2	Trustworthiness perception and trust attitude	84
4.4.3	Output acceptance	86
4.4.4	Adoption	89
4.4.5	Summary of the results	94
4.5	Discussion	94
4.5.1	Key findings and their implications	95
4.5.2	Adapting existing trust models to AI	98
4.5.3	Limitations and future work	98
4.5.4	Practical recommendations	99
4.6	Conclusion	100

5 Examining the impact of large language models' uncertainty expression on user reliance and trust	102
5.1 Introduction	103
5.2 Related work	107
5.2.1 Uncertainty expression	107
5.2.2 Uncertainty in LLMs	108
5.2.3 Measuring reliance and trust	110
5.3 Methods	112
5.3.1 Procedure and experimental conditions	112
5.3.2 Dependent variables	114
5.3.3 Hypothesis and analysis	115
5.3.4 Questions and AI responses used	117
5.3.5 Data collection and participants	119
5.4 Results: Confirmatory analysis	120
5.4.1 Agreement with AI	120
5.4.2 Confidence in answers	121
5.4.3 Source usage	121
5.4.4 Trust and perception of AI	123
5.4.5 Task performance	124
5.5 Results: Additional analyses	125
5.5.1 Effect of uncertainty expression on over- and underreliance . .	125
5.5.2 Participants' interpretations of AI's uncertainty expression . .	127
5.6 Discussion	128
5.7 Ethical considerations and positionality	131
6 Fostering appropriate reliance on large language models	134
6.1 Introduction	135
6.2 Related work	138

6.2.1	Appropriate reliance on AI	138
6.2.2	Explanations and inconsistencies	139
6.2.3	Sources	142
6.3	Study 1: Think-aloud study	143
6.3.1	Study 1 methods	144
6.3.2	Study 1 results	147
6.4	Study 2: Large-scale, pre-registered, controlled experiment	150
6.4.1	Study 2 methods	151
6.4.2	Study 2 results: Main analysis	159
6.4.3	Study 2 results: Additional analyses	166
6.5	Discussion	169
6.5.1	Implications of findings	169
6.5.2	Explanation of the answer vs. Explanation of how the LLM arrived at the Aaswer	172
6.5.3	Limitations	174
6.6	Conclusion	175
7	Conclusion	177
7.1	Key implications	177
7.2	Practical recommendations	179
7.3	Future research directions	181
A	Supplementary material for Chapter 2	183
A.1	Details on the evaluation tasks	183
A.2	Details on the evaluated interpretability methods	184
A.3	Details on the human studies	189
A.4	Additional results and analyses	189
A.5	Simple decision tree used for explaining ProtoTree	199

A.6 UI snapshots	201
B Supplementary material for Chapter 3	213
B.1 Interview protocol	213
B.2 Materials for studying XAI perceptions	218
B.3 Survey questions and results	221
C Supplementary material for Chapter 4	225
C.1 Domain and AI background levels	225
C.2 Interview protocol	226
D Supplementary material for Chapter 5	229
D.1 Participant demographics and background	229
D.2 Data collection and exclusion	231
D.3 Additional analyses	234
D.4 Full wording used in the experiment	237
D.5 Questions and AI responses used	252
E Supplementary material for Chapter 6	263
E.1 Additional analyses: Study 2 follow-up questions	263
E.2 Study 1 think-aloud protocol	266
E.3 Study 1 task questions used	269
E.4 Study 2 participant demographics and LLM background	270
E.5 Study 2 full wording of the experiment	271
E.6 Study 2 task questions and LLM responses used	276
Bibliography	288

List of Tables

- 2.1 Agreement task results.** For each study, we show mean accuracy, standard deviation of the participants’ performance, and mean confidence rating in parentheses. *Italics* denotes methods with accuracy not statistically significantly different from 50% random chance ($p > 0.05$); **bold** denotes the highest performing method in each group. In all studies, participants leaned towards believing that model predictions are correct when provided explanations, regardless of if they are actually correct. For example, for GradCAM on CUB, participants thought 72.4% of correct predictions were correct and $100 - 32.8 = 67.2\%$ of incorrect predictions were correct. These results reveal an issue of *confirmation bias*. See Section 2.5.2 for a discussion. 20

2.2 Distinction and output prediction task results. For each study, we report the mean accuracy and standard deviation of the participants' performance. <i>Italics</i> denotes methods that do not statistically significantly outperform 25% random chance ($p > 0.05$); bold denotes the highest performing method in each group. In the top half, we show the results of all four methods on CUB. In the bottom half, we show GradCAM and BagNet results on ImageNet, without vs. with ground-truth class labels. Overall, participants struggle to identify the correct prediction or the model output based on explanations. See Section 2.5.3 for a discussion.	22
3.1 Participants' domain (bird) and AI background. See Sec. 3.4.1 for a description of the background levels.	43
3.2 Summary of participants' feedback on four XAI approaches. See Sec. 3.5.3 for details.	60
4.1 Participants' domain (bird) and AI background. See the supp. mat. for a description of the categories.	82
4.2 Factors that influenced participants' trust in AI. See Section 4.4.5 for a discussion.	92
5.2 Within-condition analysis. For each of the two conditions with uncertainty, we compare DVs measured on instances with AI answers that are not uncertain vs. uncertain. We report the model-estimated means (and standard errors) from our confirmatory analysis. $>$ and $<$ note statistically significant differences with significance marked as * ($p < 0.05$) or ** ($p < 0.01$). \dagger The model does not fit properly due to large individual variance which we discuss in an exploratory analysis in Appendix C.1.	122

A.4 Interpretability-accuracy tradeoff results. We report the mean and standard deviation of the additional accuracy participants require for the baseline model, to use it over the model with explanations. For example in the GradCAM *agreement* study with CUB, participants require the baseline model to have +5.6% accuracy beyond the model that comes with GradCAM explanations and achieves 81.0% accuracy, in the low-risk setting. See Figure A.8 for a visualization of the results. 197

B.1 Summary of 19 participants’ responses to the XAI needs survey developed based on the XAI Question Bank [228]. For each question, we report the number of participants who selected “**Know** the answer to,” who selected “**Curious** to know (more),” and who selected neither (don’t know and **Not curious**). The numbers in each row don’t always add up to 19 because some participants selected both “**Know** the answer to” and “**Curious** to know (more)” for a question. 224

D.1 Demographics of our study sample (N=404) compared to the U.S. population. N/A indicates there is no data for the given category. * indicates multiple selection and self-description were possible. 231

List of Figures

2.1	Different forms of explanation. (Top left) Heatmap explanations (GradCAM [339], BagNet [52]) highlight decision-relevant image regions. (Bottom left) Prototype-based explanations (ProtoPNet [72], ProtoTree [285]) match image regions to prototypical parts learned during training. This schematic is much simpler than actual explanations. (Right) Actual ProtoPNet explanation example from the original paper. While existing evaluation methods typically apply to only one explanation form, HIVE evaluates and compares diverse interpretability methods.	6
2.2	Study user interfaces (UIs). We show simplified UIs for evaluating ProtoPNet [72] on the <i>agreement</i> task (top) and GradCAM [339] on the <i>distinction</i> task (bottom). Full UI snapshots are in the supp. mat. See Section 2.3 for description of the tasks.	12
2.3	Human study UI. We show a simplified version of the UI we developed for our human studies. In Part 1, we ask participants to guess the model’s prediction for a given image by recognizing concepts and using the provided explanations. In Part 2, we show participants explanations with different levels of simplicity and correctness, then ask which one they prefer the most.	27

3.1	Screenshots of Merlin [368], our study application. Merlin is an AI-based bird identification mobile phone app. Users upload photos on the <i>Photo ID</i> feature (top) or sounds on the <i>Sound ID</i> feature (bottom) to get a list of birds that best match the input. Users also share optional location and season data. The resulting bird list comes with example photos and sounds.	40
3.2	Identification examples and AI explanations shown to participants. We showed three <i>real</i> examples of Merlin Photo ID identifications, paired with <i>mock-up</i> explanations we designed for each. Each XAI approach was introduced with the following description. <i>Heatmap</i> : The AI model considers the red highlighted regions as evidence for [output, i.e., Evening Grosbeak. House Wren, Ruby-throated Hummingbird]. <i>Example</i> : The AI model considers the input photo most similar to the below photos of [output] the AI model has seen. <i>Concept</i> : The AI model considers the presence of the following concepts as positive/negative evidence for [output]. <i>Prototype</i> : The AI model considers the boxed regions of the input photo similar to parts of [output] photos the AI model has seen. See Sec. 3.4.2 for details.	47
4.1	Trust model of Mayer et al. [261]. Based on this model, we separate <i>trust</i> from <i>trustworthiness</i> perceptions that precede it, and two trust-related behaviors that proceed from it: (1) AI <i>adoption</i> , which corresponds to <i>risk taking in relationship</i> in the model, and (2) AI <i>output acceptance</i> , which corresponds to <i>outcomes</i> evaluation in the model. In this work, we describe both <i>general</i> trustworthiness perceptions and trust attitudes, and <i>instance-specific</i> trust-related behaviors. See Section 4.2.1 for further discussion. This figure is replicated from the work of Mayer et al. [261].	73

4.2 Our study application Merlin [368]. Merlin is a popular AI-based bird identification mobile phone app. Users upload photos on the Photo ID feature (left) or audio recordings on the Sound ID feature (right), with optional location and season data, and get a list of bird(s) that best match the input. See Section 4.3.1 for more details about the app.	79
5.1 Screenshot of the information-seeking task (shown in the CONTROL, UNCERTAIN1ST, and UNCERTAINGENERAL conditions). Additional screenshots, including the variant shown in the NO-AI condition, can be found in the supp. mat.	106
5.2 Exploratory analysis of over- and underreliance (Section 5.5.1). We analyze Correct (%) separately for questions the system answered correctly vs. incorrectly. We show the model-estimated means and standard errors for each condition (Figure 5.2a) and for not uncertain vs. uncertain responses in the conditions with uncertainty (Figure 5.2b).	125
6.1 Overview of our studies. In Study 1, participants engaged in multi-turn interactions with ChatGPT to arrive at correct answers to objective questions. Based on a thematic analysis of think-aloud and behavioral data, we identified <i>explanations</i> , <i>inconsistencies</i> , and <i>sources</i> as three features of LLM responses likely to influence user reliance. These three features were then investigated in a controlled experiment (Study 2), with features operationalized as indicated in the schematic illustration. Similar to Study 1, participants solved question-answering tasks. However, this time, they had access to one LLM response whose features we experimentally manipulated.	143

6.2 Screenshots of Study 2’s experimental task. Here the LLM response provides an incorrect answer, includes sources, and includes an explanation (with inconsistencies). See Figure 6.3 for responses with a correct answer for the same task question.	151
6.3 Types of LLM responses used in Study 2. We vary three variables in the LLM responses: accuracy of the LLM’s answer to the question (correct/incorrect), presence of an explanation (absent/present), and presence of clickable sources (absent/present). In total there are 8 types of responses. Here we show 4 types of responses with a correct answer to the question: “Do more than two thirds of South America’s population live in Brazil?” See Figure 6.2 for a response with an incorrect answer.	153
6.4 Summary of participants’ accuracy in Study 2. We plot the raw data means and 95% confidence intervals for participants’ accuracy when provided with different types of LLM responses. When the LLM’s answer is correct, participants’ accuracy is highest when the LLM response includes an explanation and sources (Figure 6.4a left). When the LLM’s answer is incorrect, participants’ accuracy is highest when the LLM response includes sources but not an explanation (Figure 6.4a right). When the LLM response includes an explanation for an incorrect answer, participants’ accuracy is higher when the explanation is inconsistent (Figure 6.4b).	162

6.5 Study 2 results on inconsistencies. We plot the raw data means and 95% confidence intervals. Brackets indicate statistically significant differences between three types of incorrect LLM responses: No explanation, Consistent explanation, and Inconsistent explanation. Significance is marked as * ($p < .05$), ** ($p < .01$), and *** ($p < .001$). See Section 6.4.3 for details.	165
A.1 GradCAM explanations shown in the <i>distinction</i> task.	185
A.2 BagNet explanations shown in the <i>distinction</i> task.	187
A.3 ProtoPNet original and modified explanations. The original explanation (left) taken from the original paper [72] contains details such as activation maps, similarity scores, and class connection weights. In our version (right), we remove these to abstract away the complexities and have the participants focus on examining the similarity between prototypes and their matched image regions.	187
A.4 ProtoTree original explanation. We show the original explanation displayed in Fig. 9 of the original paper [285]. See Figure A.5 for our modified explanation.	188
A.5 ProtoTree modified explanation. See Figure A.4 for the original explanation.	188
A.6 BagNet <i>agreement</i> study input with example images. For the study version with example images, we additionally show three example images from the predicted class (highlighted in the blue box).	190
A.7 Participant vs. ProtoPNet prototype similarity rating. There exists a gap between ProtoPNet’s similarity scores and human judgments of similarity (Spearman’s $\rho = -0.25$, $p = 0.49$ for <i>distinction</i> ; $\rho = -0.52$, $p = 0.12$ for <i>agreement</i>).	194

A.8	Visualization of the interpretability-accuracy tradeoff results.	
	This plot shows that participants desire higher accuracies for the base-line model, especially in higher-risk settings. See Table A.4 for the full results.	198
A.9	A simple decision example. We use this model to introduce participants to decision trees before explaining the more complex ProtoTree. See Appendix A.5 for details.	200
A.10	1. Study introduction.	203
A.11	2. Demographics and background.	203
A.12	3. Method introduction. BagNet (top left), GradCAM (top right), ProtoPNet (bottom left), ProtoTree (bottom right).	204
A.13	4. Task preview and first subjective evaluation.	205
A.14	5. Task: GradCAM <i>distinction</i>.	206
A.15	5. Task: GradCAM <i>agreement</i>.	206
A.16	5. Task: BagNet <i>distinction</i>.	207
A.17	5. Task: BagNet <i>agreement</i>.	207
A.18	5. Task: ProtoPNet <i>distinction</i>.	208
A.19	5. Task: ProtoPNet <i>agreement</i>.	209
A.20	5. Task: ProtoTree <i>distinction</i>.	210
A.21	5. Task: ProtoTree <i>agreement</i>.	211
A.22	6. Second and third subjective evaluation.	212
A.23	7. Interpretability-accuracy tradeoff.	212

B.1 Snapshot of the survey. For each of the 10 question categories (e.g., <i>Data</i>), we ask participants to select all questions in the category they “know the answer to” in one block (as shown in the figure) and “curious to know (more)” in another block. See Chapter 3 for details about the survey and Table B.1 for the full survey questions and participants’ responses.	216
B.2 Heatmap-based explanations. (Left) Example explanation from Selvaraju and colleagues’ work [339]. (Right) Our mock-up.	219
B.3 Example-based explanations. (Left) Example explanation from Yeh, Kim and colleagues’ work [422]. (Right) Our mock-up.	220
B.4 Concept-based explanations. (Left) Example explanation from Ramaswamy and colleagues’ work [319]. (Right) Our mock-up.	220
B.5 Prototype-based explanations. (Left) Example explanation from Chen, Li, and colleagues’ work [72]. The full explanation consists of up to 10 rows. (Right) Our mock-up.	221
D.1 Task example page shown for participants in the Control, Uncertain1st, UncertainGeneral conditions.	239
D.2 Study stimuli for one of the tasks. The question is the same, but participants in different conditions were provided with different AI responses (e.g., with vs. without uncertainty) or not provided with an AI response.	245

Chapter 1

Introduction

Artificial Intelligence (AI) technologies are advancing at an incredible pace and transforming how we live, work, and communicate [10, 132, 238, 265, 281, 315, 427]. At the same time, AI technologies frequently and unexpectedly fail for various reasons. While technology failure is not new, the enormous complexity and unpredictability of AI technologies make it extremely difficult for users, and even developers, to know when and why AI technologies fail and how much to trust them [5, 105, 150, 154, 234, 257, 287, 345]. Nevertheless, AI technologies are increasingly deployed in the real world for their impressive capabilities, and this trend is widely predicted to grow [179, 343, 446]. We are in a pivotal moment where we must ensure AI technologies are developed and deployed responsibly in a way that supports users to have safe and successful interactions with these powerful, yet imperfect technologies.

At the heart of responsible AI is *human-centered evaluation* that foregrounds human needs and contexts in evaluating the performance and impact of AI [42, 49, 70, 227, 364]. As AI technologies are ultimately developed for human interaction and can have a profound impact on individuals and society at large, AI evaluation must take a human-centered approach to realistically reflect real-world performance and impact, and subsequently to reduce unexpected failures and negative impact of AI.

Current AI evaluation, however, is predominantly technology-centered. The most widely utilized methods are standardized benchmarks that automatically evaluate an AI model’s performance on some task (e.g., question answering). Other technology-centered methods include non-contextualized human ratings (e.g., having humans rate AI outputs on non-contextualized criteria such as fluency and coherence) and simulated evaluation (e.g., using AI models to simulate human judgments). Technology-centered evaluation methods are commonly used because they are relatively cost- and time-efficient, which make them *scalable*. However, they are less *realistic* due to their constrained and simple nature, and they have been criticized for their lack of validity and inability to capture nuanced meaning.

On the other end, there are human-centered evaluation methods, such as contextualized human ratings (e.g., describing the context of AI use and having humans rate AI outputs on context-specific criteria) and application-grounded studies with the highest context and human requirement realism [100, 227]. Human-centered evaluation methods tend to be more *realistic* because they are flexible and nuanced, allowing us to explore deeper questions like “how” and “why.” Thus, while they may be less *scalable*, when done well, they can better assess real-world performance and impact of AI technologies and guide us on where and in what direction to make technical innovation.

The goal of this thesis is to push AI evaluation to be more human-centered by both demonstrating the value of human-centered evaluation and making it more approachable to researchers and practitioners who may be less familiar with it.

To begin, in Chapter 2, I discuss the challenges and necessity of *human* (as opposed to automatic) evaluation using explainable AI as a case study. While explainable AI methods are developed to support human understanding and decision-making, they are typically evaluated with automatic metrics due to challenges associated with human evaluation. In this chapter, I demonstrate how we can address these challenges and present large-scale human evaluations [190, 319] that reveal critical issues in the

state-of-the-art methods, such as their tendency to increase users’ overreliance on incorrect AI outputs. I close the chapter with a discussion on how to responsibly develop and deploy explainable AI methods to minimize their unintended negative effects.

Next, I present *contextualized* studies with real AI users that contribute a holistic and nuanced understanding of users’ explainability needs and trust in AI. These studies were done with users of a real-world AI application (i.e., Merlin Bird ID [368]) with varying levels of AI and domain expertise. Chapter 3 provides insights on what explainability needs users have and how they perceive popular explainable AI approaches, highlighting new issues and opportunities in explainable AI research [192]. Chapter 4 elucidates when and how users trust (or not trust) AI and identifies human, AI, and context-related factors that can influence user trust in AI [191]. Overall, these contextualized studies offer guidance on what to evaluate and where to make technical innovation for effective and responsible AI development.

Then, I illustrate how studies of human-AI *interaction* can help foster appropriate user trust and reliance on AI, another key aspect of responsible AI. Chapter 5 presents a large-scale, pre-registered experiment that examines how user trust and reliance are influenced by natural language uncertainty expressions from large language models (LLMs), one of the most influential AI technologies in the current era [193]. Chapter 6 presents two studies: first, a think-aloud study that explores what features of LLM responses shape users’ reliance, and second, a large-scale, pre-registered experiment that evaluates the effects of the identified features — explanations, sources, and inconsistencies in LLM responses — on users’ reliance [194]. These studies of user perceptions and interactions with LLMs contribute actionable insights on how to adjust LLM response features to encourage appropriate user trust and reliance.

Finally, in Chapter 7, I discuss the implications and the lessons learned from the aforementioned work and describe future directions for advancing responsible AI.

Chapter 2

Evaluating the usefulness of AI explanations

As AI technology is increasingly applied to high-impact, high-risk domains, there have been a number of new methods aimed at making AI models more understandable to humans. However, there is a lack of systematic evaluation of proposed AI explanation methods, especially regarding their usefulness to human users. In this chapter, we introduce HIVE (Human Interpretability of Visual Explanations), a novel human evaluation framework that assesses the utility of AI explanations to human users in AI-assisted decision making scenarios, and enables falsifiable hypothesis testing, cross-method comparison, and human-centered evaluation of visual AI explanation methods. To the best of our knowledge, this is the first work of its kind. Using HIVE, we conduct IRB-approved human studies and evaluate a diverse range of methods: GradCAM, BagNet, ProtoPNet, and ProtoTree. Our results suggest that AI explanations engender human trust, even for incorrect predictions, yet are not distinct enough for users to distinguish between correct and incorrect predictions. We also demonstrate how HIVE can be extended to study other aspects of AI explanations (e.g., explanation complexity). In a follow-up study on concept-based

explanations, we find that existing AI explanations are too complex, highlighting the need to consider human capabilities and preferences in developing new methods. We open-source HIVE to enable future studies and encourage more human-centered approaches to AI interpretability (explainability) research. HIVE can be found at <https://princetonvisualai.github.io/HIVE>.

This chapter contains materials published in the ECCV 2022 paper, HIVE: Evaluating the Human Interpretability of Visual Explanations [190], in collaboration with Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky, and the CVPR 2023 paper, Overlooked Factors in Concept-based Explanations: Dataset Choice, Concept Learnability, and Human Capability [319], in collaboration with Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky.

2.1 Introduction

With the growing adoption of AI in high-impact, high-risk domains, there have been a surge of efforts aimed at making AI models more interpretable. Motivations for interpretability include allowing human users to trace through a model’s reasoning process (accountability, transparency), verify that the model is basing its predictions on the right reasons (fairness, ethics), and assess their level of confidence in the model (trustworthiness). The *interpretability* research field tackles these questions and is comprised of diverse works, including those that provide explanations of the behavior and inner workings of complex AI models [29, 30, 124, 125, 308, 339, 350, 428, 441], those that design inherently interpretable models [52, 64, 65, 72, 99, 103, 200, 285, 316], and those that seek to understand what is easy and difficult for these models [7, 397, 433] to make their behavior more interpretable.

Despite much methods development, there is a relative lack of standardized evaluation methods for proposed techniques. Existing evaluation methods for computer

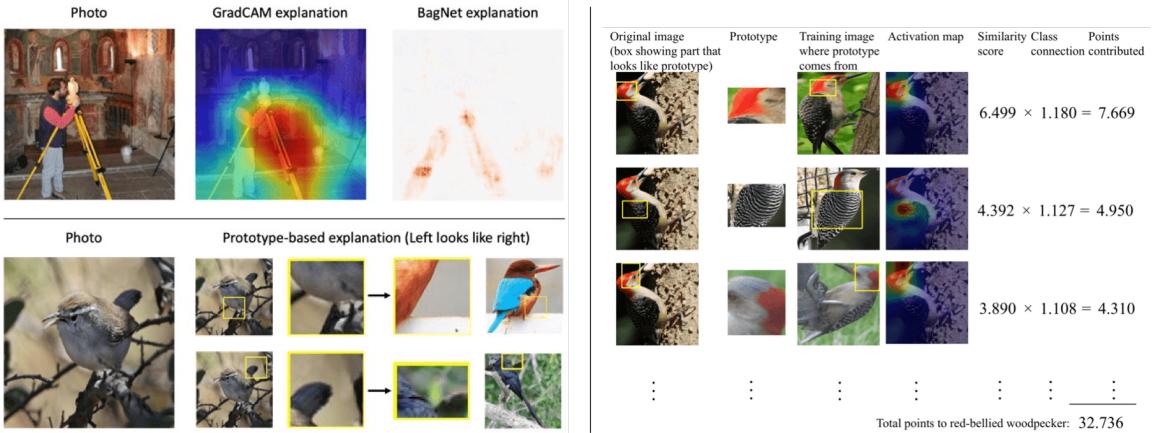


Figure 2.1: **Different forms of explanation.** (Top left) Heatmap explanations (GradCAM [339], BagNet [52]) highlight decision-relevant image regions. (Bottom left) Prototype-based explanations (ProtoPNet [72], ProtoTree [285]) match image regions to prototypical parts learned during training. This schematic is much simpler than actual explanations. (Right) Actual ProtoPNet explanation example from the original paper. While existing evaluation methods typically apply to only one explanation form, HIVE evaluates and compares diverse interpretability methods.

vision interpretability methods are focused on feature attribution heatmaps that highlight “important” image regions for a model’s prediction. Since we lack ground-truth knowledge about which regions are *actually* responsible for the prediction, different evaluation metrics use different proxy tasks for verifying these important regions (e.g., measuring the impact of deleting regions or the overlap between ground-truth objects and highlighted regions) [123, 157, 308, 311, 421, 432]. However, these automatic evaluation metrics are disconnected from downstream use cases of explanations; they don’t capture how useful end-users find heatmaps in their decision making. Further, these metrics don’t apply to other forms of explanations, such as prototype-based explanations produced by some of the recent interpretable-by-design models [72, 99, 285].

In part due to these challenges, the interpretability of a proposed method is often argued through a few exemplar explanations that highlight how a method is more interpretable than a baseline model. However, recent works suggest that some methods are not as interpretable as originally imagined and may engender over-trust

in automated systems [5, 105, 150, 154, 234, 257, 287, 345]. They caution against an over-reliance on intuition-based justifications and raise awareness for the need of falsifiable hypotheses [219] and proper evaluation in interpretability research.

Our contributions. As more diverse interpretability methods are being proposed, it is more important than ever to have a standardized and rigorous evaluation framework that allows for falsifiable hypothesis testing, cross-method comparison, and human-centered evaluation. To this end, we develop HIVE (Human Interpretability of Visual Explanations). HIVE evaluates diverse visual interpretability methods by evaluating all methods on a common task. We carefully design the tasks to reduce the effect of confirmation bias and human prior knowledge in interpretability evaluation, and assess the utility of explanations in AI-assisted decision making scenarios. HIVE also examines how well interpretable-by-design models’ reasoning process aligns with that of humans, and how human users tradeoff interpretability and accuracy.

To demonstrate the extensibility and applicability of HIVE, we conduct IRB-approved human studies with nearly 1000 participants and evaluate four existing methods that represent different streams of interpretability work (e.g., post-hoc explanations, interpretable-by-design models, heatmaps, and prototype-based explanations): GradCAM [339], BagNet [52], ProtoPNet [72], ProtoTree [285]. To the best of our knowledge, we are the first to compare interpretability methods with different explanation forms (see Figure 2.1) and the first to conduct human studies of the evaluated interpretable-by-design models [52, 72, 285].

We obtain a number of insights through our studies:

- When provided explanations, participants tend to believe that the model predictions are correct, revealing an issue of *confirmation bias*. For example, our participants found 60% of the explanations for *incorrect* model predictions convincing. Prior work has made similar observations for non-visual interpretability

methods [312]; we substantiate them for visual explanations and demonstrate a need for rigorous evaluation of proposed methods.

- When given multiple model predictions and explanations, participants struggle to distinguish between correct and incorrect predictions based on the explanations (e.g., achieving only 40% accuracy on a multiple-choice task with four options). This result suggests that interpretability methods need to be improved to be reliably useful for AI-assisted decision making.
- There exists a gap between the similarity judgments of humans and prototype-based models [72, 285] which can hurt the quality of their interpretability.
- Participants prefer to use a model with explanations over a baseline model without explanations. To switch their preference, they require the baseline model to have +6.2% to +10.9% higher accuracy.

As interpretability is fundamentally a human-centric concept, it needs to be evaluated in a human-centric way. We hope our work helps pave the way towards human evaluation becoming commonplace, by presenting and analyzing a human study design, demonstrating its effectiveness and informativeness for interpretability evaluation, and open-sourcing the code to enable future work.

2.2 Related work

2.2.1 Interpretability landscape in computer vision

Interpretability research can be described along several axes: first, whether a method is post-hoc or interpretable-by-design; second, whether it is global or local; and third, the form of an explanation (see [1, 27, 56, 74, 122, 135, 143, 323] for surveys). *Post-hoc explanations* focus on explaining predictions made by already-trained models, whereas

interpretable-by-design (IBD) models are intentionally designed to possess a more explicitly interpretable decision-making process [52, 64, 65, 72, 99, 103, 200, 285, 316]. Furthermore, explanations can either be *local explanations* of a single input-output example or *global explanations* of a network (or its component parts). Local, post-hoc methods include heatmap [125, 308, 339, 347, 350, 428, 441], counterfactual explanation [138, 383, 398], approximation [320], and sample importance [199, 422] methods. In contrast, global, post-hoc methods aim to understand global properties of CNNs, often by treating them as an object of scientific study [29, 30, 124, 189] or by generating class-level explanations [318, 442]. Because we focus on evaluating the utility of explanations in AI-assisted decision making, we do not evaluate global, post-hoc methods. *IBD* models can provide local and/or global explanations, depending on the model type. Lastly, explanations can take a variety of forms: two more popular ones we study are *heatmaps* highlighting important image regions and *prototypes* (i.e., image patches) from the training set that form interpretable decisions. In our work, we investigate four popular methods that span these types of interpretability work: GradCAM [339] (post-hoc, heatmap), BagNet [52] (IBD, heatmap), ProtoPNet [72] (IBD, prototypes), and ProtoTree [285] (IBD, prototypes). See Figure 2.1 for examples of their explanations.

2.2.2 Evaluating heatmaps

Heatmap methods are arguably the most-studied class of interpretability work. Several automatic evaluation metrics have been proposed [23, 123, 157, 308, 311, 421, 432], however, there is a lack of consensus on how to evaluate these methods. Further, the authors of [5, 6] and BAM [421] highlight how several methods fail basic “sanity checks” and call for more comprehensive metrics. Complementing these works, we use HIVE to study how useful heatmaps are to human users in AI-assisted decision

making scenarios and demonstrate insights that cannot be gained from automatic evaluation metrics.

2.2.3 Evaluating interpretable-by-design models

In contrast, there has been relatively little work on assessing interpretable-by-design models. Quantitative evaluations of these methods typically focus on demonstrating their competitive performance with a baseline CNN, while the quality of their interpretability is often demonstrated through qualitative examples. Recently, a few works revisited several methods’ interpretability claims. Hoffmann et al. [154] highlight that prototype similarity of ProtoPNet [72] does not correspond to semantic similarity and that this disconnect can be exploited. Margeloiu et al. [257] analyze concept bottleneck models [200] and demonstrate that learned concepts fail to correspond to real-world, semantic concepts. In this work, we conduct the first human study of three popular interpretable-by-design models [52, 72, 285] and quantify prior work’s [154, 285] anecdotal observation on the misalignment between prototype-based models [72, 285] and humans’ similarity judgment.

2.2.4 Evaluating explanations with human studies

Outside the computer vision field, human studies are commonly conducted for models trained on tabular datasets [210, 211, 216, 312, 437]; however, these do not scale to the complexity of modern vision models. Early human studies for visual explanations have been limited in scope: They typically ask participants which explanation they find more reasonable or which model they find more trustworthy based on explanations [169, 339]. Recently, more diverse human studies have been conducted [40, 47, 117, 287, 345, 347, 449].

Closest to our work are [117, 287, 345]. Shen and Huang [345] ask users to select incorrectly predicted labels with or without showing explanations; Nguyen et al. [287]

ask users to decide whether model predictions are correct based on explanations; Fel et al. [117] ask users to predict model outputs in a concurrent work. Regarding [287, 345], our *distinction* task also investigates how useful explanations are in distinguishing correct and incorrect predictions. However, different from these works, we ask users to select the correct prediction out of multiple predictions to reduce the effect of confirmation bias and don't show class labels to prevent users from relying their prior knowledge. Regarding [117], we also ask users to predict model outputs, but mainly as a supplement to our *distinction* task. Further, we ask users to identify the model output out of multiple predictions based on the explanations, whereas [117] first trains users to be a meta-predictor of the model by showing example model predictions and explanations, and then at test time asks users to predict the model output for a given image without showing any explanation. Most importantly, different from [117, 287, 345], we evaluate interpretability methods beyond heatmaps and conduct cross-method comparison. Our work is similar in spirit to work by Zhou et al. [445] on evaluating generative models with human perception. For general guidance on running human studies in computer vision, refer to work by Bylinskii et al. [63].

2.3 HIVE design principles

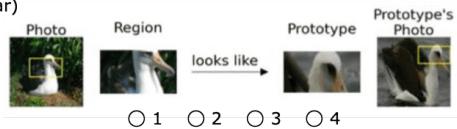
In this work, we focus on AI-assisted decision making scenarios, in particular those that involve an image classification model. For a given input image, a user is shown a model's prediction along with an associated explanation, and is asked to make a decision about whether the model's prediction is correct or more generally about whether to use the model. In such a scenario, explanations are provided with several goals in mind: help the user identify if the model is making an error, arrive at a more accurate prediction, understand the model's reasoning process, decide how much to trust the model, etc.

Agreement task: Rate the similarity of each row's prototype-region pair on a scale of 1-4.

(1: Not similar, 2: Somewhat not similar, 3: Somewhat similar, 4: Similar)

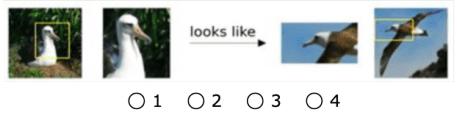


The model predicts **Species 2** for this photo. Shown below is the model's explanation for its prediction (all prototypes and their source photos are from **Species 2**).

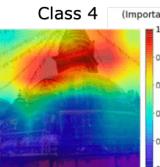
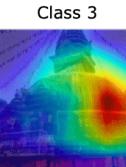
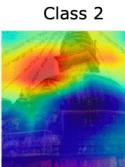
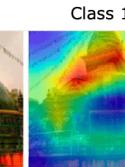


Q. What do you think about the model's prediction?

- Fairly confident that prediction is *correct*
- Somewhat confident that prediction is *correct*
- Somewhat confident that prediction is *incorrect*
- Fairly confident that prediction is *incorrect*



Distinction task: Select the class you think is correct based on the explanations.



Q. Which class do you think is correct?

- 1
- 2
- 3
- 4

Figure 2.2: **Study user interfaces (UIs).** We show simplified UIs for evaluating ProtoPNet [72] on the *agreement* task (top) and GradCAM [339] on the *distinction* task (bottom). Full UI snapshots are in the supp. mat. See Section 2.3 for description of the tasks.

To study whether and to what extent different visual interpretability methods are useful for AI-assisted decision making, we develop a novel human evaluation framework named HIVE (Human Interpretability of Visual Explanations). In particular, we design HIVE to allow for *falsifiable hypothesis testing* regarding the usefulness of explanations for identifying model errors, *cross-method comparison* between different explanation approaches, and *human-centered evaluation* for understanding the practical effectiveness of interpretability.

2.3.1 Falsifiable hypothesis testing

We join a growing body of work that cautions against intuition-based justification and subjective self-reported ratings in interpretability evaluation [5, 206, 219, 332] and calls for objective assessment with behavior indicators [212, 312, 425, 437]. To this end, we design two evaluation tasks, the *agreement* and *distinction* tasks, that enable *falsifiable hypothesis testing* about the evaluated interpretability method.

In the *agreement* task, we present participants with one prediction-explanation pair at a time and ask how confident they are in the model’s prediction based on the explanation. We evaluate methods on this task in part because it is closer to existing interpretability evaluation schemes that consider a model’s top-1 prediction and its explanation [339], and also because it allows us to quantify the degree to which participants believe in model predictions based on explanations.

The *agreement* task measures the amount of *confirmation bias* that arises for a given interpretability method. However, it doesn’t measure the utility of explanations in distinguishing correct and incorrect predictions, a crucial functionality of explanations in AI-assisted decision making. Hence, we design and use the *distinction* task as our main evaluation task. Here we simultaneously show four predictions and their associated explanations for a given input image and ask users to identify the correct prediction based on the provided explanations. The *distinction* task also mitigates the effect of confirmation bias in interpretability evaluation, as participants now have to reason about multiple explanations at once. See Figure 2.3 for the evaluation task UIs.

One concern with this setup is ensuring that participants use the provided explanations rather than their knowledge to complete the task. We take two measures to remove the effect of *human prior knowledge* in our evaluations. First, we evaluate all interpretability methods in the context of fine-grained bird species classification [391], which is a challenging task for non-bird experts. Second, as a more general measure, we omit the semantic class labels of the predictions. This measure is particularly important when evaluating interpretability methods in easier contexts, e.g., coarse-grained object classification with ImageNet [324], because the task becomes too easy otherwise (i.e., participants can select the correct prediction based on the class labels instead of using the explanations). Note that ground-truth class labels are also omitted to simulate a realistic decision making scenario where users do not have access to the ground truth.

2.3.2 Cross-method comparison

Existing evaluation methods typically apply to only one explanation form (e.g., heatmaps are compared against each other). In contrast, HIVE enables *cross-method comparison* between different explanation forms by focusing on downstream uses of explanations and evaluating all methods on a common task.

However, there remains a number of practical roadblocks. First, different methods may have been developed for different scenarios (e.g., fine-grained vs. coarse-grained classification), requiring us to carefully analyze the effect of the particular setting during evaluation. Second, different methods may be more or less digestible to the users. While this is an inherent part of what we are trying to evaluate, we also want to ensure that the evaluation task is doable by study participants with limited machine learning background, given most human studies in the field are run through Amazon Mechanical Turk. Hence, we actualize a specific evaluation setup for each interpretability method by creating an individual evaluation UI that respects the method’s characteristics (e.g., its explanation form, dataset used for model training). We briefly describe the four methods we evaluate in this work (see Figure 2.1 for example explanations) and their evaluation setups. When making any adaptations, we tried to present each method in as favorable of a way as possible. More details are in the supp. mat.

GradCAM [339]. GradCAM is a post-hoc method that produces a heatmap that highlights important regions in an input image that contribute to a model’s prediction. We evaluate GradCAM on ImageNet [324], which it was originally developed for, as well as on CUB [391], for which we train a standard CNN model to use as the underlying model for generating GradCAM heatmaps.

BagNet [52]. In contrast, BagNet is an interpretable-by-design model that collects evidence for a class from small regions of an image. For each class, BagNet creates a heatmap where higher values (i.e., darker red in our visualizations) imply stronger

evidence for the class. BagNet then sums the values in each heatmap and predicts the class with the highest sum. We evaluate BagNet on ImageNet, for which it was originally designed, as well as on CUB, for which we train a new BagNet model using the authors’ code.

ProtoPNet [72]. The next two methods reason with *prototypes*, which are small image patches from the training set that these models deem as representative for a certain class. At test time, ProtoPNet compares a given image to the set of prototypes it learned during training and finds regions in the image that are the most similar to each prototype. It computes a similarity score between each prototype-region pair, then predicts the class with the highest weighted sum of the similarity scores. The ProtoPNet model for CUB learns 10 prototypes for each of the 200 bird species (2,000 total) and produces one of the most complex explanations. Its explanation for a single prediction consists of 10 prototypes and their source images, heatmaps that convey the similarity between matched image regions and prototypes, continuous and unnormalized similarity scores, and weights multiplied to the scores (see Figure 2.1 right). In our evaluation, we abstract away most technical details based on our pilot studies, and focus on showing the most crucial component of ProtoPNet’s reasoning process: the prototype-image region matches. We also ask participants to rate the similarity of each match (see Figure 2.3 top) to assess how well the model’s similarity judgment aligns with that of humans. See the supp. mat. for the task and explanation modification details.

ProtoTree [285]. Finally, the ProtoTree model learns a tree structure along with the prototypes. Each node in the tree contains a prototype from a training image. At each node, the model compares a given test image to the node’s prototype and produces a similarity score. If the score is above some threshold, the model judges that the prototype is present in the image and absent if not. The model then proceeds to the next node and repeats this process until it reaches a leaf node, which corresponds to

a class. The ProtoTree model for CUB trained by the authors has 511 decision nodes and up to 10 decision steps, and our pilot studies revealed that is too overwhelming for participants. Thus in our evaluation, we significantly simplify the decision process. Participants are shown the model’s decisions until the penultimate decision node, and then are asked to make decisions for only the final two nodes of the tree by judging whether the prototype in each node is present or absent in the image. This leads the participants to select one of the four (2^2) classes as the final prediction. One additional challenge is that participants may not be familiar with decision trees and thus may have trouble following the explanation. To help understanding, we introduce a simple decision tree model with two levels, walk through an example, and present two warm up exercises so that participants can get familiar with decision trees before encountering ProtoTree. See the supp. mat. for more information.

2.3.3 Human-centered evaluation

HIVE complements existing algorithmic evaluation methods by bringing humans back into the picture and taking a *human-centered* approach to interpretability evaluation. The design of HIVE, particularly the inclusion/exclusion of class labels in Sec. 2.3.1 and careful actualization of the evaluation setup in Sec. 2.3.2, is focused on making this evaluation tractable for the participants and as fair as possible with respect to different interpretability methods. We also went through multiple iterations of UI design to present visual explanations in digestible bits so as to not overwhelm participants with their complexity. Despite the challenges, there is a very important payoff from human studies. We are able to evaluate different interpretability methods through participants’ 1) ability to *distinguish* between correct and incorrect predictions based on the provided explanations, simulating a more realistic AI-assisted decision-making setting, and 2) level of *alignment* with the model’s intermediate reasoning process in

the case of prototype-based, interpretable-by-design models. We also gain a number of valuable insights that can only be obtained through human studies.

2.3.4 Generalizability & Scalability

In closing we discuss two common concerns about human studies: generalizability and scalability. We have shown HIVE’s *generalizability* by using it to evaluate a variety of methods (post-hoc explanations, interpretable-by-design models, heatmaps, prototype-based explanations) in two different settings (coarse-grained object recognition with ImageNet, fine-grained bird recognition with CUB). Further, a recent work by Ramaswamy et al. [319] uses HIVE to set up new human studies, for evaluating example-based explanations and finding the ideal complexity of concept-based explanations, demonstrating that HIVE can be easily generalized to new methods and tasks. Regarding *scalability*, human study costs are not exorbitant contrary to popular belief and can be budgeted for like we budget for compute. For example, our Grad-CAM distinction study cost \$70 with 50 participants compensated at \$12/hr. The real obstacles are typically the time, effort, and expertise required for study design and UI development; with HIVE open-sourced, these costs are substantially mitigated.

2.4 HIVE study design

In this section, we describe our IRB-approved study design. See the supp. mat. and <https://princetonvisualai.github.io/HIVE> for UI snapshots and code.

Introduction. For each participant, we first introduce the study and receive their informed consent. We also request optional demographic data regarding gender identity, race and ethnicity, and ask about the participant’s experience with machine learning; however, no personally identifiable information was collected. Next we explain the evaluated interpretability method in simple terms by avoiding technical jargon

(i.e., replacing terms like “image” and “training set” to “photo” and “previously-seen photos”). We then show a preview of the evaluation task and provide example explanations for one correct and one incorrect prediction made by the model to give the participant appropriate references. The participant can access the method description at any time during the task.

Objective evaluation tasks. Next we evaluate the interpretability method on a behavioral task (*distinction* or *agreement*) introduced in Section 2.3.1 and Figure 2.3. Detailed task descriptions are available in the supp. mat.

Subjective evaluation questions. While the core of HIVE is in the objective evaluation tasks, we also ask subjective evaluation questions to make the most out of the human studies. Specifically, we ask the participant to self-rate their level of understanding of the evaluated method before and after completing the task, to investigate if the participant’s self-rated level of understanding undergoes any changes during the task. After the task completion, we disclose the participant’s performance on the task and ask the question one last time.

Interpretability-accuracy tradeoff questions. While interpretability methods offer useful insights into a model’s decision, some explanations come at the cost of lower model accuracy. Hence in the final part of the study, we investigate the *interpretability-accuracy tradeoff* participants are willing to make when comparing an interpretable method against a baseline model that doesn’t come with any explanation. In high-risk scenarios a user may prefer to maximize model performance over interpretability. However, another user may prefer to prioritize interpretability in such settings so that there would be mechanisms for examining the model’s predictions. To gain insight into the tradeoff users are willing to make, we present three scenarios: low-risk (e.g., bird species recognition for scientific or educational purposes), medium-risk (e.g., object recognition for automatic grocery checkout), and high-risk (e.g., scene understanding for autonomous driving). For each scenario, we

then ask the participant to input the minimum accuracy of the baseline model that would convince them to use it over the model with explanations and also describe the reason for their choices.

2.5 HIVE study results

2.5.1 Experimental details

Datasets & Models. We evaluate all interpretability methods on classification tasks and use images from the CUB [391] test set and the ImageNet [324] validation set to generate model predictions and explanations. On CUB, we evaluate all four methods: GradCAM [339], BagNet [52], ProtoPNet [72], ProtoTree [285]. On ImageNet, we evaluate GradCAM and BagNet. See the supp. mat. for details.

Human studies. For each study, i.e., an evaluation of one interpretability method on one task (*distinction* or *agreement*), we recruited 50 participants through Amazon Mechanical Turk (AMT). In total, we conducted 19 studies with 950 participants; see the supp. mat. for the full list. The self-reported machine learning experience of the participants was 2.5 ± 1.0 , between “2: have heard about...” and “3: know the basics...” The mean study duration was 6.9 minutes for GradCAM, 6.6 for BagNet, 13.6 for ProtoPNet, and 10.4 for ProtoTree. Participants were compensated based on the state-level minimum wage of \$12/hr.

Statistical analysis. For each study, we report the mean task accuracy and standard deviation of the participants’ performance which captures the variability between individual participants’ performance. We also compare the study result to random chance and compute the p -value from a 1-sample t -test.¹ When comparing results

¹We compare our results to chance performance instead of a baseline without explanations because we omit semantic class labels to remove the effect of human prior knowledge (see Section 2.3.1); so such a baseline would contain no relevant information.

Table 2.1: **Agreement task results.** For each study, we show mean accuracy, standard deviation of the participants’ performance, and mean confidence rating in parentheses. *Italics* denotes methods with accuracy not statistically significantly different from 50% random chance ($p > 0.05$); **bold** denotes the highest performing method in each group. In all studies, participants leaned towards believing that model predictions are correct when provided explanations, regardless of if they are actually correct. For example, for GradCAM on CUB, participants thought 72.4% of correct predictions were correct and $100 - 32.8 = 67.2\%$ of incorrect predictions were correct. These results reveal an issue of *confirmation bias*. See Section 2.5.2 for a discussion.

CUB	GradCAM [339]	BagNet [52]	ProtoPNet [72]	ProtoTree [285]
Correct	$72.4\% \pm 21.5$ (2.9)	$75.6\% \pm 23.4$ (3.0)	$73.2\% \pm 24.9$ (3.0)	$66.0\% \pm 33.8$ (2.8)
Incorrect	$32.8\% \pm 24.3$ (2.8)	$42.4\% \pm 28.7$ (2.7)	$46.4\% \pm 35.9$ (2.4)	$37.2\% \pm 34.4$ (2.7)
ImageNet	GradCAM [339]	BagNet [52]	-	-
Correct	$70.8\% \pm 26.6$ (2.9)	$66.0\% \pm 27.2$ (2.8)	-	-
Incorrect	$44.8\% \pm 31.6$ (2.7)	$35.6\% \pm 26.9$ (2.7)	-	-

between two groups, we compute the p -value from a 2-sample t -test. Results are deemed statistically significant under $p < 0.05$ conditions.

2.5.2 The issue of confirmation bias

Let us first examine how the four methods perform on the *agreement* task, where we present participants with one prediction-explanation pair at a time and ask how confident they are in the model’s prediction. Results are summarized in Table 2.1. On CUB, participants found 72.4% of correct predictions convincing for GradCAM, 75.6% for BagNet, 73.2% for ProtoPNet, and 66.0% ProtoTree. However, they also thought 67.2% of incorrect predictions were correct for GradCAM, 57.6% for BagNet, 53.6% for ProtoPNet, and 62.8% for ProtoTree. Similarly on ImageNet, participants found 70.8% of correct predictions convincing for GradCAM and 66.0% for BagNet, yet also believed in 55.2% and 64.4% of incorrect predictions, respectively. These results reveal an issue of *confirmation bias*: When given explanations, participants tend to believe model predictions are correct, even if they are wrong. Still, the confidence ratings are overall higher for correct predictions than incorrect predictions, suggesting

there is some difference between their explanations. More results and discussion are in the supp. mat.

2.5.3 Objective assessment of interpretability

Next we discuss findings from our main evaluation task, the *distinction* task, where we ask participants to select the correct prediction out of four options based on the provided explanations. Results are summarized in Table 2.2.

Participants perform better on correctly predicted samples. On correctly predicted samples from CUB, the mean task accuracies are 71.2% on GradCAM, 45.6% on BagNet, 54.5% on ProtoPNet and 33.8% on ProtoTree, all above the 25% chance baseline. That is, participants can identify which of the four explanations correspond to the ground-truth class correctly predicted by the model. On incorrect predictions, however, the accuracies drop from 71.2% to 26.4% for GradCAM and from 45.6% to 32.0% for BagNet, and we observe a similar trend in the ImageNet studies. These results suggest that explanations for correct predictions may be more coherent and convincing than those for incorrect predictions. Even so, all accuracies are far from 100%, indicating that the evaluated methods are not yet reliably useful for AI-assisted decision making.

Participants struggle to identify the model’s prediction. For GradCAM and BagNet, we ask participants to select the class they think the model predicts (*output prediction*) in addition to the class they think is correct (*distinction*). For BagNet, this is a straightforward task where participants just need to identify the most activated (most red, least blue) heatmap among the four options, as BagNet by design predicts the class with the most activated heatmap. However, accuracy is not very high, only marginally above the *distinction* task accuracy. This result suggests that BagNet heatmaps for the top-4 (or top-3 plus ground-truth) classes look similar to the human eye, and may not be suitable for assisting humans with tasks that involve

Table 2.2: **Distinction and output prediction task results.** For each study, we report the mean accuracy and standard deviation of the participants’ performance. *Italics* denotes methods that do not statistically significantly outperform 25% random chance ($p > 0.05$); **bold** denotes the highest performing method in each group. In the top half, we show the results of all four methods on CUB. In the bottom half, we show GradCAM and BagNet results on ImageNet, without vs. with ground-truth class labels. Overall, participants struggle to identify the correct prediction or the model output based on explanations. See Section 2.5.3 for a discussion.

CUB		GradCAM [339]	BagNet [52]	ProtoPNet [72]	ProtoTree [285]
Distinction	Correct	71.2% \pm 33.3	45.6% \pm 28.0	54.5% \pm 30.3	33.8% \pm 15.9
	Incorrect	<i>26.4% \pm 19.8</i>	32.0% \pm 20.8	-	-
Output prediction	Correct	69.2% \pm 32.3	50.4% \pm 32.8	-	-
	Incorrect	53.6% \pm 27.0	<i>30.0% \pm 24.1</i>	-	-
ImageNet		GradCAM [339]	with labels	BagNet [52]	with labels
Distinction	Correct	51.2% \pm 24.7	49.2% \pm 30.8	38.4% \pm 28.0	34.8% \pm 27.7
	Incorrect	30.0% \pm 22.4	<i>27.2% \pm 20.3</i>	<i>26.0% \pm 18.4</i>	<i>27.2% \pm 18.7</i>
Output prediction	Correct	48.0% \pm 28.3	48.0% \pm 35.6	46.8% \pm 29.0	42.8% \pm 27.4
	Incorrect	35.6% \pm 24.1	33.2% \pm 25.2	34.0% \pm 24.1	32.8% \pm 25.5

distinguishing one class from another. For GradCAM, participants also struggle on this task but to a lesser degree.

Showing ground-truth labels hurts performance. For GradCAM and BagNet, we also investigate the effect of showing ground-truth class labels for the presented images. We have not been showing them to simulate a realistic decision making scenario where users don’t have access to the ground truth. However, since the task may be ambiguous for datasets like ImageNet whose images may contain several objects, we run a second version of the ImageNet studies showing ground-truth class labels on the same set of images and compare results. Somewhat surprisingly, we find that accuracy decreases, albeit by a small amount, with class labels. One possible explanation is that class labels implicitly bias participants to value heatmaps with better localization properties, which could be a suboptimal signal for the *distinction* and *output prediction* tasks.

Automatic evaluation metrics correlate poorly with human study results. We also analyze GradCAM results using three automatic metrics that evaluate the localization quality of post-hoc attribution maps: pointing game [432], energy-based

pointing game [396], and intersection-over-union [441]. In the *agreement* studies, we find near-zero correlation between participants’ confidence in the model prediction and localization quality of heatmaps. In the *distinction* studies, we also do not see meaningful relationships between the participants’ choices and these automatic metrics. These observations are consistent with the findings of [117, 287], i.e., automatic metrics poorly correlate with human performance in post-hoc attribution heatmap evaluation. See the supp. mat. for details.

2.5.4 A closer examination of prototype-based models

We are the first to conduct human studies of ProtoPNet and ProtoTree which produce some of the most complex visual explanations. As such, we take a closer look at their results to better understand how human users perceive them.

A gap exists between similarity ratings of ProtoPNet & ProtoTree and those of humans. We quantify prior work’s [154, 285] anecdotal observation that there exists a gap between model and human similarity judgment. For ProtoTree, the Pearson correlation coefficient between the participants’ similarity ratings and the model similarity scores is 0.06, suggesting little to no relationship. For ProtoPNet, whose similarity scores are not normalized across images, we compute the Spearman’s rank correlation coefficient ($\rho = -0.25, p = 0.49$ for *distinction* and $\rho = -0.52, p = 0.12$ for *agreement*). There is no significant negative correlation between the two, indicating a gap in similarity judgment that may hurt the models’ interpretability. See the supp. mat. for more discussion.

Participants perform relatively poorly on ProtoTree, but they understand how a decision tree works. Since the previously described ProtoTree *agreement* study does not take into account the model’s inherent tree structure, we run another version of the study where, instead of asking participants to rate each prototype’s similarity, we ask them to select the first step they disagree with in the model’s ex-

planation. The result of this study ($52.8\% \pm 19.9\%$) is similar to that of the original study ($53.6\% \pm 15.2\%$); in both cases, we cannot conclude that participants outperform 50% random chance ($p = 0.33$, $p = 0.10$). To ensure participants understand how decision trees work, we provided a simple decision tree example and subsequent questions asking participants if the decision tree example makes a correct or incorrect prediction. Participants achieved 86.5% performance on this task, implying that the low task accuracy for ProtoTree is not due to a lack of comprehension of decision trees. See the supp. mat. for details.

2.5.5 Subjective evaluation of interpretability

To complement the objective evaluation tasks, we asked participants to self-rate their level of method understanding three times. The average ratings are 3.7 ± 0.9 after the method explanation, 3.8 ± 0.9 after the task, and 3.5 ± 1.0 after seeing their task performance, which all lie between the fair (3) and good (4) ratings. Interestingly, the rating tends to *decrease* after participants see their task performance ($p < 0.05$). Several participants indicated that their performance was lower than what they expected, whereas no one suggested the opposite, suggesting that participants might have been disappointed in their task performance, which in turn led them to lower their self-rated level of method understanding.

2.5.6 Interpretability-accuracy tradeoff

In the final part of our studies, we asked participants for the minimum accuracy of a baseline model they would require to use it over the evaluated interpretable model with explanations for its predictions. Across all studies, participants require the baseline model to have a higher accuracy than the model that comes with explanations, and by a greater margin for higher-risk settings. On average, participants require the baseline model to have +6.2% higher accuracy for low-risk, +8.2% for medium-risk,

and +10.9% for high-risk settings. See the supp. mat. for the full results and the participants’ reasons for their choices.

2.6 Follow-up study on concept-based explanations

Building on the study design and user interface (UI) of HIVE [190], we designed and conducted a follow-up study [319] to examine how understandable and useful concept-based explanations are to humans.

Concept-based explanation methods aim to explain a deep neural network model’s components and predictions using a pre-defined set of semantic concepts. These methods evaluate a trained model on a new, “probe” dataset and correlate the model’s outputs with concepts labeled in that dataset. Existing concept-based explanations use a large number of concepts: NetDissect [29] and Net2Vec [124] use all 1197 concepts labelled within the Broden [29] dataset; IBD [442] uses Broden object and art concepts with at least 10 examples (660 concepts); and Concept Bottleneck [200] uses all concepts that are predominantly present for at least 10 classes from CUB [391] (112 concepts). However, can people actually reason with these many concepts?

In this study, we examine this important yet overlooked aspect of concept-based explanations: *explanation complexity* and how it relates to human capability and preference. Specifically, we investigate: (1) How well do people recognize concepts in images? (2) How do the (concept recognition) task performance and time change as the number of concepts vary? (3) How well do people predict the model output for a new image using explanations? (4) How do people trade off simplicity and correctness of concept-based explanations? To answer these questions, we design and conduct a human study. We describe the study design in Section 2.6.1 and report findings in Section 2.6.2.

2.6.1 Human study design

We build on the study design and user interface (UI) of HIVE [190], and design a two-part study to understand how understandable and useful concept-based explanations are to human users with potentially limited knowledge about machine learning . To the best of our knowledge, we are the first to investigate such properties of concept-based explanations for computer vision models.²

Part 1: Recognize concepts and predict the model output. First, we present participants with an image and a set of concepts and ask them to identify whether each concept is present or absent in the image. We also show explanations for 4 classes whose scores are calculated real-time based on the concepts selected. As a final question, we ask participants to select the class they think the model predicts for the given image. See Figure 2.3 (*left*) for the study UI.

To ensure that the task is doable and is only affected by explanation complexity (number of concepts used) and not the complexity of the model and its original prediction task (e.g., 365 scenes classification), we generate explanations for only 4 classes and ask participants to identify which of the 4 classes corresponds to the model’s prediction. We only show images where the model output matches the explanation output (i.e., the model predicts the class with the highest explanation score, calculated with ground-truth concept labels), since our goal is to understand how people reason with concept-based explanations with varying complexity.

Part 2: Choose the ideal tradeoff between simplicity and correctness. Next, we ask participants to reason about two properties of concept-based explanations: *simplicity*, i.e., the number of concepts used in a given set of explanations, and *correctness*, i.e., the percentage of model predictions correctly explained by explanations, which is the percentage of times the model output class has the highest explanation

²We note that there are works examining complexity of explanations for other types of models, for example, Lage et al. [211] investigate complexity of explanations over decision sets, Bolubasi et al. [43] investigate this for concept-based explanations for language models.

Part 1: Recognize concepts and predict the model output

Concepts

- wall
- floor
- windowpane
- table
- plant
- chair
- carpet
- lamp
- bed
- sofa
- cushion
- vase
- armchair
- sconce
- coffee table
- fireplace

Explanation for Scene W

$$= \textbf{1.88}$$

- = + 1.88 x 1 (bed)
- 0.95 x 0 (chair)
- 0.60 x 0 (sofa)
- 0.28 x 0 (armchair)
- 0.04 x 0 (table)
- 0.03 x 0 (sconce)
- + 0.00

Explanation for Scene X

$$= \textbf{-2.74}$$

- = - 3.20 x 1 (bed)
- + 1.47 x 0 (chair)
- 1.38 x 0 (sofa)
- 0.80 x 1 (cushion)
- 0.39 x 0 (coffee table)
- 0.14 x 0 (armchair)
- 0.14 x 1 (lamp)
- + 1.40

Explanation for Scene Y

$$= \textbf{1.03}$$

- = + 1.36 x 1 (bed)
- 1.02 x 0 (windowpane)
- 0.92 x 1 (wall)
- 0.31 x 0 (plant)
- 0.24 x 1 (carpet)
- + 0.19 x 0 (sconce)
- 0.18 x 1 (floor)
- 0.15 x 1 (cushion)
- 0.11 x 0 (vase)
- + 1.16

Explanation for Scene Z

$$= \textbf{-0.54}$$

- = + 2.00 x 0 (sofa)
- 1.73 x 1 (bed)
- 0.88 x 0 (table)
- + 0.68 x 0 (coffee table)
- 0.52 x 0 (chair)
- 0.38 x 1 (wall)
- + 0.30 x 0 (armchair)
- + 0.20 x 0 (fireplace)
- + 0.17 x 1 (cushion)
- + 1.40

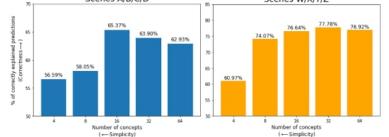
Q. Which scene class do you think the model predicts?

Scene W Scene X Scene Y Scene Z

Part 2: Simplicity-Correctness tradeoff

Simplicity refers to the number of concepts used in a given set of explanations. **Correctness** refers to the percentage of times the explanations correctly explain the model prediction.

You can choose the level of simplicity and correctness of concept-based explanations.



Q. Which would you prefer?

- Explanations that use 4 concepts
- Explanations that use 8 concepts
- Explanations that use 16 concepts
- Explanations that use 32 concepts
- Explanations that use 64 concepts

Figure 2.3: **Human study UI.** We show a simplified version of the UI we developed for our human studies. In Part 1, we ask participants to guess the model’s prediction for a given image by recognizing concepts and using the provided explanations. In Part 2, we show participants explanations with different levels of simplicity and correctness, then ask which one they prefer the most.

score. See Figure 2.3 (*right*) for the study UI. We convey the notion of a simplicity-correctness tradeoff through bar plots that show the correctness of explanations of varying simplicity/complexity (4, 8, 16, 32, 64 concepts). We then ask participants to choose the explanation they prefer the most and provide a short justification for their choice.

Full study design and experimental details. In summary, our study consists of the following steps. For each participant, we introduce the study, receive informed consent for participation in the study, and collect information about their demographic (optional) and machine learning experience. We then introduce concept-based explanations in simple terms, and show a preview of the concept recognition and model output prediction task in Part 1. The participant then completes the task for 10 images. In Part 2, the participant indicates their preference for explanation complexity, given simplicity and correctness information. There are no foreseeable risks in participation in the study, and our study design was approved by our institution’s IRB.

Using this study design, we investigate explanations that take the form of a linear combination of concepts (e.g., Baseline, IBD [442], Concept Bottleneck [200]).

Explanations are generated using the *Baseline* method, which is a logistic regression model trained to predict the model’s output using concepts. Note that we are evaluating the form of explanation (linear combination of concepts) rather than a specific explanation method. The choice of the method does not impact the task.

Specifically, we compare four types of explanations: concept-based explanations that use (1) 8 concepts, (2) 16 concepts, (3) 32 concepts, and (4) example-based explanations that consist of 10 example images for which the model predicts a certain class. We include (4) as a method that doesn’t use concepts. In Jeyakumar et al. [169], this type of explanation is shown to be preferred over saliency-type explanations for image classification; here, we compare this to concept-based explanations.

For a fair comparison, all four are evaluated on the same set of images. In short, we conduct a between-group study with 125 participants recruited through Amazon Mechanical Turk. Participants were compensated based on the state-level minimum wage of \$12/hr. In total, $\sim \$800$ was spent on running human studies. See the supp. mat. for more details.

2.6.2 Key findings from the human studies

When presented with more concepts, participants spend more time but are worse at recognizing concepts. The median time participants spend on each image is 17.4 sec. for 8 concept-, 27.5 sec. for 16 concept-, and 46.2 sec. for 32 concept-explanations. This is expected, since participants are asked to make a judgment for each and every concept. When given example-based explanations with no such task, participants spend only 11.6 seconds on each image. Interestingly, the concept recognition performance, reported in terms of mean recall (i.e., the percentage of concepts in the image that are recognized) and standard deviation, decreases from $71.7\% \pm 27.7\%$ (8 concepts) to $61.0\% \pm 28.5\%$ (16 concepts) to $56.8\% \pm 24.9\%$ (32

concepts). While these numbers are far from perfect recall (100%), participants are better at judging whether concepts are present when shown fewer number of concepts.

Concept-based explanations offer little to no advantage in model output prediction over example-based explanations. Indeed, we see that the participants' errors in concept recognition result in an incorrect class having the highest explanation score. When predicting the model output as the class with the highest explanation score, calculated based on the participants' concept selections, the mean accuracy and standard deviation of model output prediction are $64.8\% \pm 23.9\%$ (8 concepts), $63.2\% \pm 26.9\%$ (16 concepts), $63.6\% \pm 22.2\%$ (32 concepts). These are barely higher than $60.0\% \pm 30.2\%$ of example-based explanations, which are simpler and require less time to complete the task.

The majority of participants prefer explanations with 8, 16, or 32 concepts. When given options of explanations that use 4, 8, 16, 32, or 64 concepts, 82% of participants prefer explanations with 8, 16, or 32 concepts (28%, 33%, 21% respectively). Only 6% prefer those with 64 concepts, suggesting that existing explanations that use hundreds or even thousands of concepts do not cater to human preferences. In the written responses, many favored having fewer concepts (e.g., “the lesser, the better”) and expressed concerns against having too many (e.g., “I think 32 is a lot, but 16 is an adequate enough number that it could still predict well...”). In making the tradeoff, some valued correctness above all else (e.g., “Out of all the options, 32 is the most correct”), while others reasoned about marginal benefits (e.g., “I would prefer explanations that use 16 concepts because it seems that the difference in percentage of correctness is much closer and less, than other levels of concepts”). Overall, we find that participants actively reason about both simplicity and correctness of explanations.

2.7 Conclusion

In this chapter, we introduced HIVE, a novel human evaluation framework for evaluating diverse visual explanation methods, and used it to evaluate four existing methods: GradCAM, BagNet, ProtoPNet, and ProtoTree.

There are a few limitations of this work: First, we use a relatively small sample size of 50 participants for each study due to our desire to evaluate four methods, some under multiple conditions. Second, while HIVE takes a step towards use case driven evaluation, our evaluation setup is still far from real-world uses of interpretability methods. An ideal evaluation would be contextually situated and conducted with domain experts and/or end-users of a real-world application (e.g., how would bird experts choose to use one method over another when given multiple interpretability methods for a bird species recognition model).

Nonetheless, we believe our work will facilitate more user studies — as demonstrated by the follow-up study on concept-based explanations [319] — and encourage human-centered AI explanation research [107, 109, 110, 225]. Our human evaluation reveals several key insights about the field. In particular, we find that participants generally believe model predictions are correct when given explanations for them. Humans are naturally susceptible to confirmation bias; thus, interpretable explanations will likely engender trust from humans, even if they are incorrect. Our findings underscore the need for evaluation methods that fairly and rigorously assess the usefulness and effect of explanations. We hope our work helps shift the field’s objective from focusing on method development to also prioritizing the development of high-quality evaluation methods.

Acknowledgments

This material is based upon work partially supported by the National Science Foundation (NSF) under Grant No. 1763642. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF. We also acknowledge support from the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award (OR), Princeton SEAS Project X Fund (RF, OR), Open Philanthropy (RF, OR), and Princeton SEAS and ECE Senior Thesis Funding (NM). We thank the authors of [29, 52, 72, 148, 154, 200, 285, 339, 442] for open-sourcing their code and/or trained models. We also thank the AMT workers who participated in our studies, anonymous reviewers who provided thoughtful feedback, and Princeton Visual AI Lab members (especially Dora Zhao, Kaiyu Yang, and Angelina Wang) who tested our user interface and provided helpful suggestions.

Chapter 3

Understanding how explainability can support human-AI interaction

Despite the proliferation of explainable AI (XAI) methods, little is understood about end-users' explainability needs and behaviors around AI explanations. To address this gap and contribute to understanding how explainability can support human-AI interaction, we conducted a mixed-methods study with 20 end-users of a real-world AI application, the Merlin bird identification app, and inquired about their XAI needs, uses, and perceptions. We found that participants desire practically useful information that can improve their collaboration with the AI, more so than technical system details. Relatedly, participants intended to use AI explanations for various purposes beyond understanding the AI's outputs: calibrating trust, improving their task skills, changing their behavior to supply better inputs to the AI, and giving constructive feedback to developers. Finally, among existing XAI approaches, participants preferred part-based explanations that resemble human reasoning and explanations. We discuss the implications of our findings and provide recommendations for future XAI design.

This chapter contains materials published in the CHI 2023 paper, “Help Me Help the AI”: Understanding How Explainability Can Support Human-AI Interaction [192],

in collaboration with Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández.

3.1 Introduction

Artificial Intelligence (AI) systems are ubiquitous: from unlocking our phones with face identification, to reducing traffic accidents with autonomous cars, to assisting radiologists with medical image analysis. Being able to better *understand* these AI systems is becoming increasingly important—although what exactly that means is different in different settings: a smartphone user may want to *understand* how best to position their face to quickly unlock their phone, a researcher may want to *understand* what particular design decisions led to an autonomous car accident, and a radiologist may want to *understand* where the medical decision support tool is looking in suggesting a particular diagnosis.

Over the past years, numerous explainable AI (XAI) methods have been developed to provide transparency into these AI systems and make them more understandable to people (see [1, 4, 27, 93, 122, 135, 142, 143] for surveys). However, arguably these are being developed without embracing the full spectrum of end-user needs. Particularly for computer vision AI systems (such as the ones described above), with millions of model parameters processing thousands of low-level image pixels, translating model outputs into understandable insights is so challenging that proposed XAI methods are frequently limited by what XAI researchers *can do* rather than what AI end-users *might need*.

In this work, we connect XAI development with end-users and study a real-world context in which XAI methods might be deployed. Concretely, we set out to answer three research questions:

- **RQ1:** What are end-users' XAI **needs** in real-world AI applications?

- **RQ2:** How do end-users **intend to use** AI explanations¹?
- **RQ3:** How are existing XAI approaches **perceived** by end-users?

In scoping our study, we focus on Merlin, an AI-based mobile phone application that uses computer vision to identify birds in user-uploaded photos and audio recordings. We chose Merlin because it is a widely-used application that allows us to connect with a diverse set of active end-users. Concretely, we conducted a mixed-methods study with 20 Merlin users who span the range from low-to-high AI background (representing both consumers and creators of AI systems) and low-to-high domain background (representing both users who know *less* and *more* about birding than the AI system).

With each participant, we conducted an hour-long interview, which included a survey and an interactive feedback session, to understand their XAI needs, uses, and perceptions. Our study bridges the gap between XAI research done in the HCI and AI communities by directly connecting end-users of a real-world AI application with the XAI methods literature. We do so by mocking up four XAI approaches that could be potentially implemented into Merlin, i.e., heatmap, example, concept, and prototype-based explanations of the AI’s outputs. The mock-up explanations enabled us to get concrete and detailed data about how participants intended to use AI explanations, as well as how they perceived each approach, in an actual AI use context.

Through our study, we found:

- Participants’ XAI needs varied depending on their domain/AI background and interest level. While participants were generally *curious* about AI system details, those with high-AI background or notably high interest in birds had higher XAI needs. However, participants unanimously expressed a need for practically

¹In this paper, we use the term “AI explanations” to refer to explanations produced by XAI methods to explain specific AI system outputs.

useful information that can improve their *collaboration* with the AI, suggesting an important area of focus for future XAI development (RQ1, Sec. 3.5.1).

- Participants intended to use AI explanations for various purposes beyond understanding the AI’s outputs: determining when to trust the AI, learning to perform the task better on their own without needing to consult the AI, changing their behavior to supply better inputs to the AI, and giving constructive feedback to the developers to improve the AI. This highlights the broad range of XAI needs that should be considered in XAI development (RQ2, Sec. 3.5.2).
- Among existing XAI approaches, participants preferred part-based explanations, i.e., concept [318, 442] and prototype [72, 285] based explanations. Participants found them similar to human reasoning and explanations, and the most useful for the aforementioned purposes. This suggests that to the extent possible, the XAI community should pay particular attention to these methods, despite the challenges with their development and evaluation (RQ3, Sec. 3.5.3).

Following our findings, we discuss XAI’s potential as a medium for enhancing human-AI collaboration, and conclude with a set of recommendations for future XAI design. However, as with any case study, our findings and recommendations may have limited generalizability. This is an intentional trade-off made to gain an in-depth understanding of end-users’ XAI needs, uses, and perceptions in a real-world context, in line with growing calls for human-centered XAI research [107, 109, 110, 225, 229]. We are hopeful that our study design and insights will aid future XAI research in other contexts.

3.2 Related work

3.2.1 From algorithm-centered to human-centered XAI

With the growing adoption of AI, there has been a surge of interest in explainable AI (XAI) research that aims to make AI systems more understandable to people. XAI is one of the fastest growing fields with hundreds of new papers published each year. See [1, 2, 4, 27, 93, 100, 122, 135, 142, 143, 278, 323, 330, 356] for in-depth surveys, and the following for examples of XAI research done in different disciplines: AI [125, 188, 200, 306, 352], HCI [153, 355, 395, 434], social and cognitive science [54, 88, 269, 271, 358, 366], and philosophy [31, 181, 275]. XAI is also increasingly being researched and applied in various domains, including but not limited to healthcare [15, 252, 258, 310, 351, 354, 420, 435], autonomous driving [20, 267, 294], energy and power systems [253], and climate science [255].

Much of the field’s efforts originally focused on the *algorithms*, i.e., on providing explanations of AI systems’ inner workings and outputs, rather than the *people* or the *context* where these systems are deployed. Recently, there has been a growing recognition that XAI methods cannot be developed “in a vacuum” without an understanding of people’s needs in specific contexts [107, 109, 110, 225, 228, 229]. In response, researchers have proposed *conceptual frameworks* to characterize XAI needs based on people’s roles [218, 314, 371], expertise [276], or more fine-grained axes of knowledge and objectives [361]. Others interviewed *industry practitioners* who work on AI products to identify their common XAI needs [38, 156, 228].

We join this relatively new line of research, called “human-centered XAI” [107, 109, 110, 225, 228, 229], and foreground the people who use AI systems and their needs, goals, and contexts in *understanding how explainability can support human-AI interaction*. In doing so, we build on the aforementioned frameworks to study end-users’ explainability needs. Concretely, we developed a survey based on Liao and

colleagues' XAI Question Bank [228] to collect concrete data on which aspects of AI end-users want to know about.

3.2.2 Understanding end-users' XAI needs

Although human-centered XAI is an actively growing area of research, much of the work still focuses on developers rather than end-users of AI systems [38, 156, 228]. This gap is unsurprising, since XAI methods have been primarily developed for and used by developers to inspect AI systems [38, 271]. But it is critical because end-users may have different explainability needs that XAI methods should but don't yet support.

Recently, some researchers began looking at end-users' XAI needs in context of specific applications [66, 67, 372]. Tonekaboni and colleagues [372] placed clinicians in *hypothetical scenarios* where AI models are used for health risk assessment, and found that clinicians wanted to know what features the model uses so they can understand and rationalize the model's outputs. In a lab setting, Cai and colleagues [67] studied clinicians' needs in their interaction with a *prototype* AI model that can assist with cancer diagnoses, and found that clinicians desired overall information about the model (e.g., capabilities and limitations, design objective) in addition to explanations of the model's individual outputs. In another lab setting, Cai and colleagues [66] examined what needs pathologists have when using a *prototype* AI model for retrieving similar medical images. They also studied how pathologists use their proposed refinement tools, finding that pathologists often re-purposed them to test and understand the underlying search algorithm and to disambiguate AI errors from their own errors.

These studies delivered rich insights. However, they studied *hypothetical* or *prototype* AI applications. Hence, an important question remains, which we tackle in this work: *What are end-users' XAI needs in real-world AI applications? (RQ1)*. Elish and Watkins [114] recently provided insights into this question through an in-situ

study of a deployed, real-world AI system. Specifically, they documented the types of inquiries physicians asked of nurses tasked with monitoring Sepsis Watch [340], an AI system designed to predict patients’ risk of sepsis development. However, they did not study how XAI methods could answer the physicians’ inquiries. In this paper, we take a step further and contribute to understanding how XAI methods can satisfy (or not satisfy) end-users’ needs by studying: *How do end-users intend to use AI explanations? (RQ2)* and *How are existing XAI approaches perceived by end-users? (RQ3)*.

Our work extends prior work in three more ways. First, while all aforementioned work [66, 67, 114, 372] studies AI applications that make or support high-stakes medical decisions, we focus on an ordinary application that a diverse set of people use in everyday life. Second, while prior work does not differentiate their participants, we study group differences with respect to domain and AI background levels. We are inspired by recent findings of Ehsan and colleagues [111] on how people’s perceptions of AI explanations differed based on their AI background. Third, we connect to the XAI methods literature directly, by mocking-up AI explanations in the studied application. These in-situ mock-up explanations allowed us to gather detailed data on how end-users perceive and intend to use AI explanations in their actual use of the AI.

3.2.3 XAI’s role in human-AI collaboration

Our work also connects to the literature of human-AI collaboration [17, 19, 67, 203, 214, 394], sometimes called human-AI teaming [25, 26, 298] or human-AI partnership [286], that studies how people work together with AI to achieve shared goals. In this work, we didn’t set out to study human-AI collaboration. Our use of this term emerged from our findings: while studying participants’ XAI needs, uses, and perceptions, we found that participants described a process for which the language of “collaboration” proved the best fit. Participants described a two-way exchange, where they help Merlin succeed in bird identification and obtain more accurate re-

sults in return, and expressed a strong desire to improve their collaboration with AI explanations and other information. Hence, we give a brief overview of the human-AI collaboration literature and describe how our work connects to existing work.

Prior work has studied how people collaborate with different types of AI systems (e.g., robots [117, 205, 287, 288, 293, 389], virtual agents [19, 83, 106, 291], embedded systems [12, 67, 117, 119, 170, 190, 212, 214, 286, 288, 376]) in different task contexts (e.g., content generation [221, 248, 438], medical diagnosis [67, 119, 376], content moderation [170, 214], deception detection [212, 286], cooperative games [19], and fine-grained visual recognition [117, 190, 287, 288]). Among these, our work is most closely related to [117, 190, 212, 287, 288] that studied XAI’s role in AI-assisted decision making, where AI makes a recommendation and a human makes the final decision. In this work, we explored what role AI explanations could play in Merlin where for each bird identification, end-users make the final call based on the app’s output and their knowledge of birds and the app.

However, different from our work, [117, 190, 212, 287, 288] focused on measuring the usefulness of specific XAI methods in AI-assisted decision making through lab experiments. These experiments typically consisted of simple tasks (e.g., binary choice) and were conducted with participants recruited from Amazon Mechanical Turk. Further, because they were lab experiments, it was well-defined in advance how participants should use AI explanations in their collaboration with AI (e.g., look at the provided explanation and judge whether or not to accept the AI’s output). On the other hand, our qualitative descriptive study allowed us to find that participants intended to use AI explanations for various purposes, highlighting a broad range of XAI needs and uses that should be considered in XAI development.

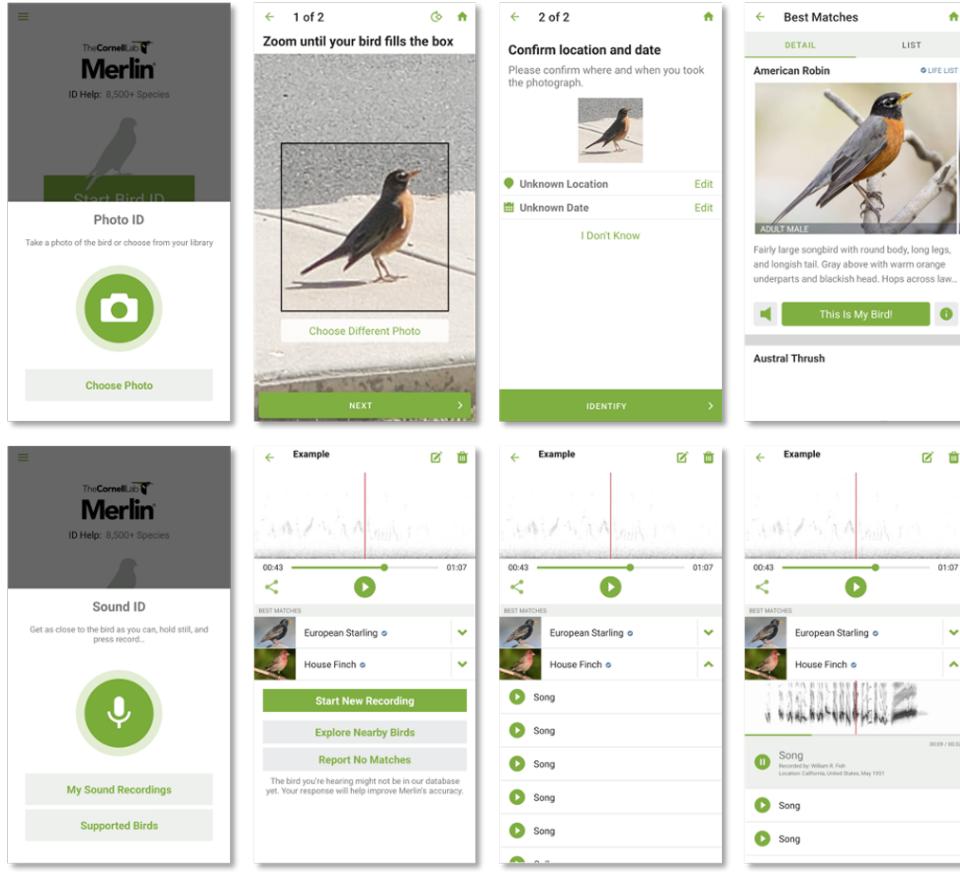


Figure 3.1: **Screenshots of Merlin [368], our study application.** Merlin is an AI-based bird identification mobile phone app. Users upload photos on the *Photo ID* feature (top) or sounds on the *Sound ID* feature (bottom) to get a list of birds that best match the input. Users also share optional location and season data. The resulting bird list comes with example photos and sounds.

3.2.4 XAI methods for computer vision

Finally, we review the XAI methods literature to provide background on how we mocked up AI explanations for Merlin. We focus on methods developed for computer vision AI models because Merlin uses computer vision to identify birds in user-input photos and audio recordings. See [1, 27, 56, 122, 135, 143, 323] for overviews.

XAI methods can be categorized along several axes: first, whether a method is post-hoc or interpretable-by-design; second, whether it provides a global or local explanation; and third, by the explanation form. To begin, the majority of existing XAI methods are *post-hoc* methods that explain certain aspects of already-trained

models [29, 30, 124, 125, 189, 199, 308, 318, 320, 339, 347, 350, 422, 428, 441, 442]. Recently, more *interpretable-by-design* methods are being proposed; these are typically new types of computer vision models with an explicitly-interpretable reasoning process [52, 64, 65, 72, 99, 103, 200, 285, 316]. Second, XAI methods provide either a *local* explanation of a model’s individual output or a *global* explanation of a model and its behavior. Local, post-hoc methods include feature attribution [125, 308, 339, 347, 350, 428, 441], approximation [320], and sample importance [199, 422] methods. Global, post-hoc methods include methods that generate class-level explanations [318, 442] and summaries of what a model has learned [29, 30, 124, 189]. Interpretable-by-design models can provide local and/or global explanations, depending on the model type. Lastly, explanations come in a variety of forms. Representative ones are *heatmaps* [52, 123, 308, 339, 347, 350, 396, 428, 441], *examples* [199, 422], *concepts* [200, 318, 442], and *prototypes* [72, 99, 285, 288]. To the best of our knowledge, these cover the range of XAI methods for computer vision.

Since we are not affiliated with the Merlin development team and do not have access to its AI models, it was not possible to produce *actual* explanations of how Merlin identifies birds. Hence, we created *mock-up* explanations. For comprehensiveness, we mocked up all four aforementioned explanation forms. We know they all are plausible XAI approaches for Merlin because they have been demonstrated on bird image classification models in prior work (e.g., heatmaps in [190, 305, 396], examples in [288], concepts in [200, 318, 319], prototypes in [72, 99, 285, 288]). See Fig. 3.2 and Sec. 3.4.2 for the mock-ups and their descriptions, and the supp. mat. for details about how we created the mock-ups.

3.3 Study application: Merlin Bird ID

As described in the related work section, we looked for a research setting that involves real-world AI use by end-users with a diverse domain and AI knowledge base, and that people use in ordinary, everyday life scenarios. Furthermore, we looked for a domain with significant AI and XAI research. We found Merlin [368] fit what we were looking for. Merlin is a mobile phone app with over a million downloads that end-users, with diverse birding and AI knowledge, use for bird identification as they go out and about outdoors. Most birding apps are digital field guides that don't use AI (e.g., Audubon Bird Guide [284], iBird Pro Guide [273], eBird [367], Birdadvisor 360° [309]). Merlin is unique in that it uses computer vision AI models to identify birds in user-input photos and audio recordings.

Merlin provided a grounded context with real end-users whose experience we can augment with mock-ups of AI explanations. Furthermore, a large proportion of XAI methods for computer vision have been developed and evaluated on bird image classification [72, 99, 103, 138, 200, 285, 288, 305, 316, 383] using the Caltech-UCSD Birds (CUB) dataset [391]. Hence, the feedback we collect on the mock-up explanations for Merlin can provide concrete and immediate insights to XAI researchers.

3.4 Methods

In this section, we describe our study methods, all of which were reviewed and approved by our Institutional Review Board prior to conducting the study.

3.4.1 Participant recruitment and selection

We recruited participants who are end-users of Merlin's Photo ID and/or Sound ID, the app's AI-based bird identification features, with considerations for diversity in participants' domain and AI background. Concretely, we created a screening survey

Table 3.1: **Participants’ domain (bird) and AI background.** See Sec. 3.4.1 for a description of the background levels.

	Low-AI	Medium-AI	High-AI
Low-domain	P7, P12, P16	P8, P14	P11, P13
Medium-domain	P2, P20	P1, P4, P10	P6
High-domain	P5, P17	P3, P9, P15	P18, P19

with questions about the respondent’s domain background, AI background, and app usage pattern (e.g., regularly used app features, frequency of app use). We posted the survey on a variety of channels: Birding International Discord, AI for Conservation Slack, various Slack workspaces within our institution, and Twitter. On Twitter, in addition to posting the survey, we reached out to accounts with tweets about Merlin via @mentions and Direct Messages.

Based on the screening survey responses, we selectively enrolled participants to maximize the diversity of domain and AI background of the study sample. See the table for a summary of participants’ background. The subgroups were defined based on participants’ survey responses and interview answers. We refer to individual participants by identifier P#.

- *Low-domain*: From “don’t know anything about birds” (P11, P12) to “recently started birding” (P7, P8, P13, P14, P16). Participants who selected the latter option typically have been birding for a few months or more than a year but in an on-and-off way, and were able to identify some local birds.
- *Medium-domain*: Have been birding for a few years and/or can identify most local birds (P1, P2, P4, P6, P10, P20).
- *High-domain*: Have been birding for more than a few years and/or do bird-related work (e.g., ornithologist) (P3, P5, P9, P15, P17, P18, P19).
- *Low-AI*: From “don’t know anything about AI” (P16, P17) to “have heard about a few AI concepts or applications” (P2, P5, P7, P12, P20). Participants

in this group either didn't know that Merlin uses AI (P12, P16) or knew but weren't familiar with the technical aspects of AI (P2, P5, P7, P17, P20).

- *Medium-AI*: From “know the basics of AI and can hold a short conversation about it” (P1, P3, P8, P9, P14) to “have taken a course in AI or have experience working with an AI system” (P4, P10, P15). Participants in this group had a general idea of how Merlin’s AI might work, e.g., it is neural network based and has learned to identify birds based on large amounts of labeled examples.
- *High-AI*: Use, study, or work with AI in day-to-day life (P6, P11, P13, P18, P19). Participants in this group were extremely familiar with AI in general and had detailed ideas of how Merlin’s AI might work at the level of specific data processing techniques, model architectures, and training algorithms.

Note that our referral here and elsewhere to “high-AI background” participants describes their expertise with AI in general, not necessarily with Merlin’s AI. All participants were active Merlin users who could provide vivid anecdotes of when the app worked well and not. Regarding frequency of use, 11 participants used it several times a week, 8 used it once a week, and one used it once a month.

3.4.2 Study instrument

Our interviews were structured in three parts and included a short survey and an interactive feedback session. The complete study instrument is attached in the supp. mat.

Context First, we asked the participant a series of questions aimed at learning the context of their app use. These include questions about their background; when, where, why, and how they use the app; stakes in their use; and their knowledge and perception of AI.

XAI needs Next, we inquired about the participant’s explainability needs through open-ended questions and a survey we developed from the XAI Question Bank [228]. The survey lists 10 categories of questions that people might have about an AI system. Nine categories (*Data, Output, Performance, How, Why, Why not, What if, How to be that, How to still be this*) are from [228], and we added a new *Transparency* category on expert and social transparency [108]. The survey asks the participant to select questions they “know the answer to” and/or are “curious to know (more).” We directed the participant to interpret “know the answer to” as “have a good idea of the answer” for questions whose exact answers are not available to end-users (e.g., What is the size of the data?).

XAI uses and perceptions Finally, we assessed the participant’s perception of existing XAI approaches. Using screen sharing during the Zoom interview, we showed three examples of Merlin Photo ID identifications: the first is a correct identification; the second is a misidentification, one that people—even experienced birders—would make; and the third is a misidentification, but one that people wouldn’t make. Using these as running examples, we introduced four XAI approaches one at a time and in random order (see the next paragraph for more information). For each, we asked the participant what they like and dislike about the approach, what they think can be improved, whether they want to see it in the app, and how much it helps them understand the AI’s reasoning and output. See Fig. 3.2 for the identification examples and AI explanations shown to participants.

To get detailed answers from participants about XAI uses and perceptions, we created mock-ups of representative XAI approaches that could potentially be embedded into Merlin. These included:

- *Heatmap*-based explanations that highlight regions in the input image that are important for the AI model’s output. They represent feature attribution methods that visualize results via heatmaps (also known as saliency maps) [52, 123, 308, 339, 347, 350, 428, 441] and include popular techniques like Grad-CAM [339].
- *Example*-based explanations that show examples in the training data that are important for the AI model’s output. They include methods that use influence functions [199] and representer points [422] to identify important positive/negative training examples for a particular output.
- *Concept*-based explanations that explain the AI model’s output with text-based concepts. They include concept bottleneck models [200], as well as methods like IBD [442] and ELUDE [318] that generate class-level explanations as a linear combination of concepts.
- *Prototype*-based explanations that explain the AI model’s output with visual prototypical parts. They represent methods such as ProtoPNet [72], ProtoTree [285], and their recent variations [99, 288].

As described in Sec. 3.2.4, these cover the range of XAI methods for computer vision. In the supp. mat., we describe in detail how we created the mock-up explanations and show examples of explanations beyond bird image classification. We emphasize that the three identification examples are *real* app outputs that we obtained by uploading photos to Merlin Photo ID. However, the shown AI explanations are *mock-ups* that we designed; they are not representative of how Merlin Photo ID actually identifies birds. We made this point clear to participants during the interviews. We also communicated that we were not affiliated with Merlin’s AI development team.

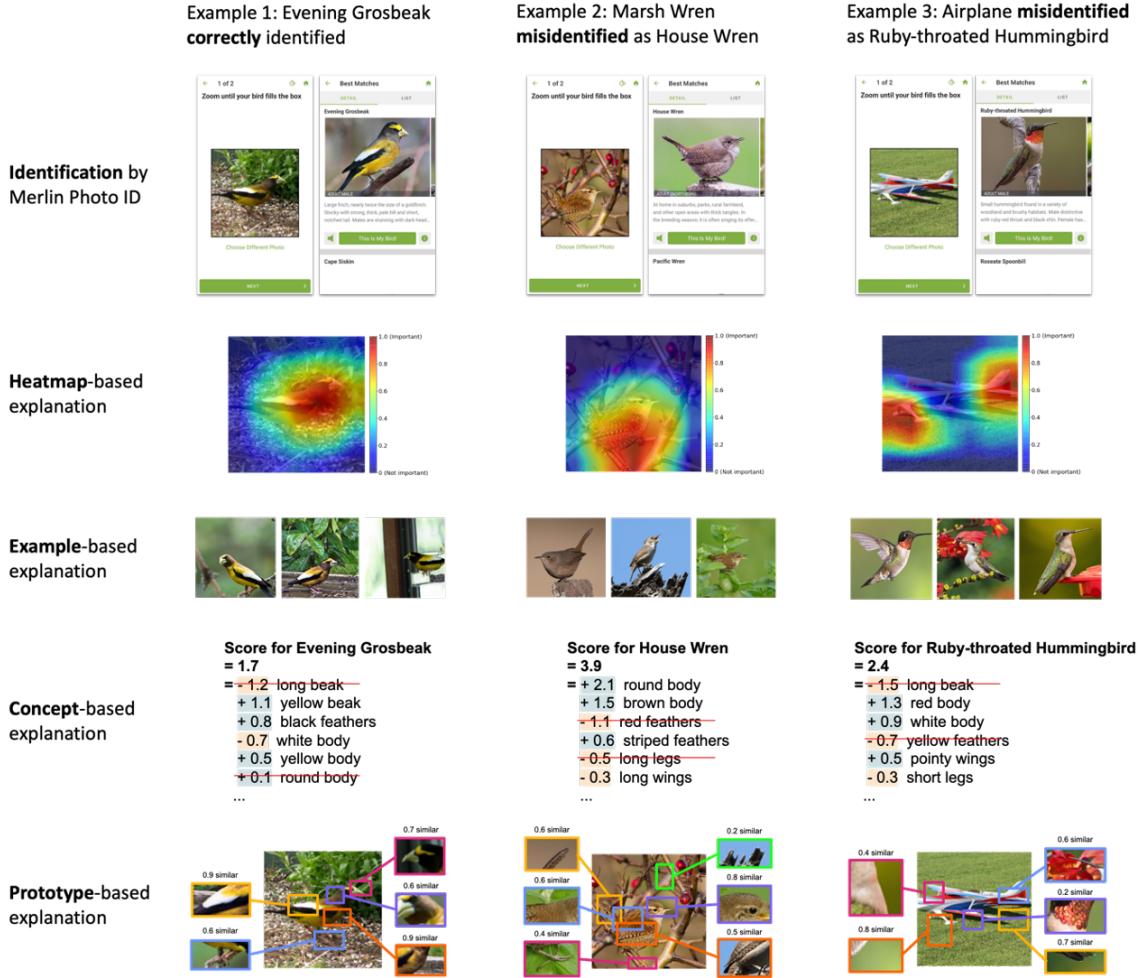


Figure 3.2: Identification examples and AI explanations shown to participants. We showed three *real* examples of Merlin Photo ID identifications, paired with *mock-up* explanations we designed for each. Each XAI approach was introduced with the following description. *Heatmap*: The AI model considers the red highlighted regions as evidence for [output, i.e., Evening Grosbeak, House Wren, Ruby-throated Hummingbird]. *Example*: The AI model considers the input photo most similar to the below photos of [output] the AI model has seen. *Concept*: The AI model considers the presence of the following concepts as positive/negative evidence for [output]. *Prototype*: The AI model considers the boxed regions of the input photo similar to parts of [output] photos the AI model has seen. See Sec. 3.4.2 for details.

3.4.3 Conducting and analyzing interviews

We interviewed 20 participants, each over a Zoom video call, from July to August 2022. The interviews lasted one hour on average. Participants were compensated with their choice of a 30 USD gift card or a donation to a bird conservation organi-

zation made on their behalf. We transcribed the interviews and then analyzed the transcripts. First, two authors read through five transcripts to develop an initial shared codebook, primarily using descriptive coding, i.e., describing participants' expressions of their beliefs, perceptions, and actions. Descriptions were at the semantic level within the surface meanings of the data (e.g., *desire to learn from the AI to improve their task skills*). Next, during multiple group meetings, all authors iterated on and refined the codebook, by shifting from descriptions of participants' beliefs to identification of shared latent, conceptual themes [326]. *Desire for improved human-AI collaboration* is an example of a latent, conceptual theme we interpreted based on participants' expressed needs for information that would help them understand the AI's capabilities and limitations, identify sources of errors, and supply better inputs, in order to more effectively interact with the AI and achieve better outcomes. After we collectively agreed that our conceptual themes were exhaustive, we then identified and eliminated themes which were redundant or overlapping. Once we had a final revised codebook, one author then used this codebook to re-code all of the data. Example codes include: *desire to know the AI's confidence in its output* (XAI needs), *learn how to take better pictures/audio recordings for the AI* (XAI uses), and *heatmap-based explanations are too coarse* (XAI perceptions). We deliberately did not calculate inter-rater reliability (IRR) as part of our analytic process. McDonald and colleagues [262] argue that such a calculative process is a poor fit for the interpretive paradigm from which qualitative research has developed. Our codebook is derived from our collective and agreed-upon interpretations of our participants' responses to our questions, and so a mathematical post-hoc comparison of individuals' coding selections would bring little rigor to the process. Instead, we focused on bringing rigor to our analysis through the discussions and selections involved in the codebook development.

3.5 Results

We present our results in three parts. We begin by describing participants’ explainability needs (RQ1, Sec. 3.5.1). We then describe how they intended to use AI explanations (RQ2, Sec. 3.5.2). Finally, we describe how they perceived the four representative XAI approaches we mocked-up for Merlin (RQ3, Sec. 3.5.3).

3.5.1 XAI needs

Based on open-ended questions and the survey we developed from the XAI Question Bank [228], we found that while participants were generally curious about AI system details, only those with high-AI background and/or high-domain interest were willing to actively seek out this information. However, participants unanimously expressed a need for information that can improve their collaboration with the AI system.

Participants were generally curious about AI system details, but curiosity levels differed based on AI background and domain interest

As most other AI applications, Merlin does not provide much information about its underlying technology. Hence, when we asked participants what they knew about the app’s AI, all replied that they didn’t know much about system details, although those with high-AI background (P6, P11, P13, P18, P19) had detailed guesses about the app’s data, model architectures, and training algorithms.

So what did participants want to know? According to the survey results, participants wanted to know everything about the app’s AI. For all questions in the survey, most if not all participants responded they “know (or have a good idea of) the answer” and/or are “curious to know (more).” That is, participants were curious about overall system details (questions in the *Data*, *Output*, *Performance*, *How*, *Transparency* categories), as well as how the AI reasons and makes judgments on

specific inputs (questions in the *Why*, *Why not*, *What if*, *How to be that*, *How to still be this* categories). We report the full survey results in the supp. mat.

But how curious are they, really? When we tempered self-reported levels of curiosity with interview questions about the effort participants were willing to invest to satisfy that curiosity, the picture changed. “*I wouldn’t go tremendously out of my way to find the answer to these questions*” (P12) was a sentiment shared by many participants (P1, P5, P6, P7, P9, P10, P12, P13, P16, P20). For instance, P5 said: “*If there’s an opportunity that arises, I’d love to ask about it [...] but I don’t think I would be contacting people at Cornell [app developers]*.” Other participants were open to searching around a bit (P9, P10), listening to talks or podcasts (P12), or reading some documentation if easily available (P1, P6, P7, P13, P16, P20), but didn’t want to take the initiative to seek out more information about the AI system, as described by the questions in the survey.

Exceptions were some participants with *high-AI background* (P11, P18, P19) or notably *high interest in birds* (P1, P4, P8). P11, P18, and P19, likely because they develop AI systems in their work, were very curious about the app’s AI and were willing to go to the extent of reaching out the app developers (P11, P18) or playing with the data themselves (P19). For example, P19 said: “*I’d love to talk to one of the engineers and pick their brain [...] or get some data and play with it myself*.” P1, P4, P8 have medium-AI background but their exceptionally high interest in birds seemed to fuel their curiosity about the app’s AI. They were particularly curious about how the AI tackles difficult identifications such as mockingbirds that mimic other birds or birds that are difficult for experienced human birders to identify (e.g., “little brown birds”).

In contrast, participants with *low-to-medium AI background* (P7, P8, P9, P10, P12, P16) had lower explainability needs. For instance, P7, P8, and P10 had little-to-no interest about the how the AI reasons and makes judgments on specific inputs. P8 said questions in the *Why*, *Why not*, *What if*, *How to be that*, *How to still be this*

categories were not what they would ever think about on their own. P7 expressed more bluntly that they prefer to keep the AI as a black box: “*No, I don’t want to ruin the mystique.*” P9, P12, and P16, on the other hand, became more curious during the interview, however, their responses suggest that they were not very curious about the AI in their natural use environment prior to the interview.

In short, all participants were interested in learning more about the AI, but only those with high-AI background and/or high-domain interest were willing to expend effort to gain more information about the AI’s system details.

Participants desired information that can improve collaboration with AI

Participants’ expressed needs for explanation shifted, however, when our interview questions moved away from gauging their curiosity about AI system details, and towards querying their use of the app. While participants’ needs for system details differed based on background and interest, they unanimously expressed a need for practically useful information that could improve their collaboration with the AI system.

To begin, participants wanted a general understanding of the AI’s capabilities and limitations (P1, P4, P5, P16, P19, P20). P1 described a number of ways this understanding would help their use of the app: “*It would definitely first help me understand more about when certain identifications may be more or less reliable. But also it will help me supply better inputs to the app to try and get the best quality identification results that I can*” (P1). Participants had already tried to gain this understanding by pushing the AI to its limits (P4, P5, P16, P19, P20). Some had tried to fool the AI with non-bird sounds (e.g., sounds of other animals, bird impersonations) to understand when it works and when it breaks (P4, P5, P16, P19). Others had conducted more rigorous experimentation by altering their input (e.g., clip the audio recording, remove location information) and observing changes in the AI’s output to understand what factors influence the AI’s output and how (P4, P20).

Another frequently expressed need was for a display of the AI’s confidence (P1, P2, P3, P4, P6, P13, P18, P20). Participants wanted this information to better determine when to trust the AI’s output. Concretely, P2 demanded for percentage-based confidence scores: “*If it doesn’t give a percentage [...] I just don’t have a gauge of how correct it is*” (P2). P7 requested the AI to qualify its output by saying “it may not be the exact match” or give a general answer (e.g., “we don’t know the exact species but this bird is in the Wren family”).

Lastly, participants wanted the AI to give more *detailed outputs* (P2, P10, P11, P12). They demanded information that would help them verify the AI’s output. For instance, P10 wanted the AI to “*highlight the time period of the [sound] clip that it calls a certain species*” because it is hard to know which sound corresponds to which bird when multiple birds are singing at once. Going a step further, P2, P11, and P12 wanted the AI to specify the type of bird sound it heard. Currently, the verification process is arduous because each bird species has a number of songs and calls, as well more specific sounds such as juvenile calls, flock calls, and alarm calls. They said the suggested features will make the verification process easier and provide more information about how the AI has made its identification, with which they can more readily check the AI’s output and determine whether to trust it.

In sum, when we queried participants about their actual, real-world use of the app, they expressed a desire for information which could improve their use of the app, particularly in deciding whether or not to trust the AI’s outputs. Intriguingly, they expressed these desires before we showed them our mock-ups of what AI explanations for the app might look like. This suggests that these XAI needs were not prompted solely by seeing AI explanations.

3.5.2 XAI uses

Next, when we showed AI explanations to participants, they were excited to use them for various purposes beyond understanding the AI’s outputs: for determining when to trust the AI, which is a well-known use and commonly-stated motivation for XAI [118, 270, 333, 425, 437], but also for learning to perform the task better on their own, changing their behavior to supply better inputs to the AI, and giving feedback to the developers to improve the AI, which are less discussed uses in existing literature.

Participants intended to use explanations to determine when to trust AI

Many participants said they would use explanations to determine when to believe the app’s identification result (P1, P4, P8, P11, P13, P18, P20). The need underlying this use is consistent with the aforementioned need for information that helps them decide when to trust the AI. While looking at different explanation mock-ups, participants gave examples of when their trust would increase and decrease. For instance, participants said they would feel more confident in the AI’s output when heatmap-based explanations show that the AI is “*looking at the right things*” (P8) and when example-based explanations show example photos that look similar to their input photo. Conversely, they said they would feel more skeptical when heatmap-based explanations suggest that an “*artifact was important*” (P8), when concept-based explanations have errors in their concept recognition (e.g., says there is a long beak when there is not) (P18), and when prototype-based explanations match photo regions and prototypes that don’t look similar to them (P4). These findings confirm existing literature [118, 270, 333, 425, 437] and suggest that trust calibration will be an important use of XAI.

Participants desired to learn via explanations to better achieve the task on their own

Intriguingly, a greater number of participants said that they intend to use explanations to improve their task skills (P1, P2, P4, P6, P7, P8, P9, P10, P11, P13, P15, P17, P19, P20). Participants viewed the AI as a teacher and were keen to learn the features it looks at via explanations, so they can look for these features in the future when they are birding on their own. Participants were aware that the features the AI looks at may be different from what expert human birders look at. But they weren't very concerned about the potential differences. One participant even said it would be interesting if the AI finds new ways of identifying birds and explanations can “call attention towards things that people did not really think of before” (P1). Still, participants preferred that explanation *forms* be similar to those of human birders. We elaborate on this point further in Sec. 3.5.3.

Overall, participants were excited about how explanations could make birding more accessible for themselves and others who lack access to expert resources (e.g., mentoring from human birders):

“It [the explanation] is kind of training or giving me more information and I’m kind of learning these things [what features to look at]. Whereas before, birders or ornithologists are learning this from mentors or teachers in the field. But those opportunities are limited based on social relations, privilege, how closely you are connected to birding groups and stuff. And so it will be much more openly accessible if that kind of more comparative identification knowledge was accessible through just an app.” –

P1

Even participants with high-domain background, whose main goal for using the app was not to obtain such knowledge, appreciated the educational value of explanations and said explanations would help them learn faster (P16).

These findings are closely related to recent works by Goyal and colleagues [138] and Pazzani and colleagues [305]. They demonstrated that AI explanations help non-bird-experts (graduate students in machine learning [138] and undergraduate students in psychology, cognitive science, or linguistics courses [305]) learn to distinguish birds. While their experiments employed relatively easy tasks, i.e., assigning bird images to one of two species options, they showed the potential of *learning from AI via AI explanations*. While [138, 305] did not establish that this is a need that people have, our work provides empirical evidence for it, suggesting *learning from AI* as another important use case for XAI.

We postulate this use case stemmed from Merlin’s status as an expert AI system. Many AI applications are deployed to automate tasks that are easy for people (e.g., face verification, customer service chatbot) in settings where it is costly or implausible to have humans in the loop. In contrast, Merlin possesses expertise that most people don’t have and need to invest time and effort to gain. This expertise is likely the source of Merlin explanations’ educational value. In other types of AI applications, end-users may not intend to learn from AI via explanations.

Participants viewed explanations as an opportunity to be better AI-collaborators

Participants also saw explanations as an opportunity for action. They looked for feedback on their own behavior that would in turn enable them to help the AI better achieve the task (P1, P7, P9, P20). P20 said explanations, by providing insights into how the AI got an identification wrong, can help them figure out the answer to: “*What would I have to do to change this photo to make it [AI] understand it better?*”

Participants sought out opportunities to improve their own collaborative skills when working with the AI to achieve a task, because at the end of the day they want to achieve best possible outcomes:

“You’re still trying to look for the right bird. So if you can adjust human behavior to get the right answer out of the robot [AI], then that’s helpful.”

– P20

Because of this need, participants were critical towards XAI approaches they thought didn’t provide actionable feedback. For instance, P9 questioned the utility of heatmap and example-based explanations: *“How is it helpful to the user in the future? Besides just being cool and interesting? How does it change the user’s use of the app? Does it make you take a different photo?”* They critiqued that these approaches don’t help them help the AI be more correct.

We view use this intended use of AI explanations as an extension of participants’ current efforts to help out the AI. When describing their use of the app, participants mentioned several different ways they help the AI perform better. Some were smaller adjustments on the spot, such as facing the microphone closer to the bird and getting a sufficiently long recording for Sound ID (P9). Others were more involved, such as the efforts P1 described as part of their “*general workflow*” for using Photo ID:

“I basically don’t use images that are either too blurry or do not feature the bird in an unobstructed manner. I know from my personal experience using it that Merlin works a lot better if it has a more silhouetted side profile shot of the bird. [...] So I try to feed Merlin photos taken from similar angles, also in acceptable lighting conditions. I might have to boost the contrast or the brightness of a picture artificially to feed it into Merlin to get better results. If there’s no real contrast, then it’s much harder to get credible results.” – P1

In short, participants viewed the AI as a collaborator. They have already found ways to better work with it, and they intended to use AI explanations to further improve their collaboration. To this end, they wanted explanations to provide actionable feedback on their own behavior so that they can supply better inputs to the AI.

Participants saw explanations as a medium to give feedback to developers and improve AI

Finally, participants with high-AI background intended to use explanations as a medium to give feedback to developers and contribute to improving the AI (P13, P18, P19). These participants mentioned that explanations, by providing more information to end-users about how the AI produced its output, enable end-users to give more detailed feedback. This feedback can then help developers improve the AI system. P13 illustrated this process using prototype-based explanations as an example:

“The fact that it [AI] identifies parts of the tree, that’s a great opportunity to [to have end-users] tap that region and say ‘not a part of the bird’ so that you can get the users helping you to do some curation and labeling on the images, which someone could review or whatever. You can make much higher quality models by getting this sort of the labeling right.” – P13

P18 suggested a similar feedback process for example-based explanations. They said when end-users disagree with the provided examples of similar looking birds, they can correct them by saying “*no, I think it actually looks more like bird number three*” and help developers align the AI’s notion of perceptual similarity with that of humans, and improve the AI.

Lastly, P19 described XAI’s potential for creating a positive feedback loop that helps both end-users and the AI system:

“So there’s a feedback loop here, right? Because if that [teaching people to better identify birds] is your goal, and you’re successful in doing that, then you’re able to rely on people to verify their data, contribute solid data, and that data can help inform Merlin, which makes Merlin better, which makes it do its job better. [...] I think no matter what, it [providing explanations] is kind of beneficial.” – P19

P13 and P18 shared this view and said they would be excited to help developers improve the app by providing feedback via explanations. P18, in particular, expressed a strong desire to contribute. They had already been signing up for beta versions of the app, and the first answer they gave to the question “What would you like to know more about Merlin?” was: *“How I can contribute more”* (P18).

In short, participants with high-AI background desired to use explanations to help improve the AI, so that they can achieve better outcomes with it in the future. We interpret this as another example of participants viewing the AI as a collaborator whom they work together with.

3.5.3 XAI perceptions

In this last results section, we describe how participants perceived the four XAI approaches we mocked up: Heatmap (Sec. 3.5.3), Example (Sec. 3.5.3), Concept (Sec. 3.5.3), and Prototype (Sec. 3.5.3). We also summarize concerns expressed toward AI explanations (Sec. 3.5.3), and explore how existing XAI approaches might satisfy end-users’ explainability needs and goals identified in the previous sections.

Heatmap-based explanations: Most mixed opinions

We received the most mixed reviews for heatmap-based explanations. Participants who liked heatmaps described them as “*fun*” (P15), “*aesthetically pleasing*” (P3), and intuitive—“*it’s very easy, it hits you right away*” (P9). Some participants were

positive because they often use heatmaps in their work and find them helpful for representing information (P12, P19). Conversely, a few participants expressed a strong dislike (P14, P16), e.g., “*I hate those things [...] They are simply not intuitive*” (P14). P20 didn’t like heatmaps as an explanation form because “*heatmaps feel like they should be related to weather*,” revealing individual differences in perception.

Regarding utility, some said heatmaps help them understand how the AI had made a mistake (P7, P9, P13). For instance, P19 said they see how the AI made a mistake for the Marsh Wren photo because the heatmap (in Fig. 3.2) did not highlight areas that are important for distinguishing different species of Wrens (e.g., Marsh Wren has a white eyebrow that House Wren doesn’t). However, many participants criticized that the shown heatmaps were too coarse and uninformative (P1, P2, P3, P4, P6, P10, P11, P16, P17, P19). “*It’s just highlighting the bird*” was a common remark. Participants said heatmaps would be more helpful if they highlight a few salient features of the bird, just like how human birders focus on a few field markers when identifying birds.

Finally, some participants thought heatmap-based explanations were inherently limited by its form. P9, P11, and P17 said heatmaps were unsatisfying because they don’t answer the “why” question. Regarding heatmaps’ highlighted regions, P17 asked: “*Yes it’s important, but why was it important?*” Other participants were dissatisfied because heatmaps lacked actionable information (P9, P11). They said knowing which parts of the photo were important to the AI does not help them change their behavior to help the AI be more correct in future uses.

Example-based explanations: Intuitive but uninformative

There was a consensus among participants that example-based explanations are “*really easily understandable*.” However, opinions diverged regarding their utility. Some found them helpful for determining when to trust the AI (P4, P5, P17, P20) since

Table 3.2: **Summary of participants’ feedback on four XAI approaches.** See Sec. 3.5.3 for details.

XAI	Praises	Complaints
<i>Heatmap</i>	Intuitive, pleasing Helpful for spotting AI’s mistakes	Unintuitive, confusing Uninformative, too coarse Doesn’t explain why certain parts are important Doesn’t give actionable information
<i>Example</i>	Intuitive Helpful for verifying AI’s outputs Allows end-users to do their own moderation	Uninformative, impression-based Doesn’t add much to current examples Doesn’t give actionable information
<i>Concept</i>	Part-based form Resembles human reasoning and explanations Helpful for verifying AI’s outputs Helpful for learning bird identification Final scores and coefficients are helpful	Current concepts are too generic Meaning of coefficients is unclear Numbers are overwhelming
<i>Prototype</i>	Part-based form Resembles human reasoning and explanations Intuitive, visual Helpful for verifying AI’s outputs Helpful for learning bird identification	Cluttered, difficult to see on small screens Some prototypes are ambiguous and uninteresting

they themselves can compare their input photo to the examples in the explanations. P4 noted that example-based explanations feel “*much more collaborative*” since they allow end-users to do their own moderation of the provided information. P19, on the other hand, were concerned that they would “*open the door for user error*.” Especially for difficult identifications where there are only subtle differences between candidate birds, P19 said example-based explanations wouldn’t help non-bird-expert end-users arrive at a more accurate identification.

Many participants described example-based explanations as rather uninformative (P1, P4, P6, P8, P10, P11, P12, P18). Some thought they didn’t add much information to example photos that are already shown in the app with the identification result (P1, P6, P10, P11). They understood the difference between the two, that example-based explanations convey what the AI considers similar to the input photo, while the currently provided example photos are part of a fixed bird description and independent of the input. Still, they thought the explanations were not very useful. Some even preferred the current example photos because they are high-quality and well-curated (P1, P6).

Another frequent criticism against example-based explanations was that they are too general and impression-based (P4, P8, P10, P12, P18). Participants were frustrated that they don't communicate what features the AI was using to make its identifications, e.g., P8 said "*This kind of tells you nothing.*" Due to this lack of specificity, many mentioned that example-based explanations were not helpful for their various intended uses, ranging from understanding the AI's reasoning to supplying better inputs to the AI to improving their own bird identification skills.

Concept-based explanations: Well-liked overall but overwhelming to some

Participants were largely positive towards concept-based explanations. Most praises were about their part-based form. They liked that the AI's output was broken down into chunks that human birders reason with, i.e., concepts (P3, P4, P11). "*This is what a person looks for basically when they're identifying a bird,*" remarked P3. Relatedly, participants liked that concept-based explanations resemble the way bird identifications are taught and shared between birders (P3, P8, P17). P17 said, "*before all this technology, this is exactly how you would basically learn to ID a bird.*" For these reasons, participants mentioned that concept-based explanations seem helpful for learning to identify birds on their own.

Participants also mentioned other use cases where concept-based explanations can help. For instance, P11 said they would allow people to check the AI's output more thoroughly because people can agree or disagree with the explanation at the level of individual concepts. As an example, they said they would not believe the AI's output if the explanation says there are red feathers in the photo when there are not. Participants also liked that the shown explanations provided a final score for the output because they display the AI's confidence in the identification (P1, P5, P17). P5 said such scores would be particularly helpful when they are comparing similar-looking candidate birds.

Nonetheless, participants mentioned a few areas of improvement. Several participants pointed out that the concepts in the shown explanations (e.g., long beak, black feathers, white body) were too general (P1, P4, P5, P10). They suggested adopting birders' language and describing birds with more specific terms such as "*underbelly, chest, rump, wing, wingbars, neck, head, cap*" (P4). Participants also recommended making the numbers in the explanations as easily understandable as possible (P6, P9, P12, P13, P15, P16, P18). P6 pointed out that the current concept coefficients are confusing: "*I have no idea what any of the numbers mean? Like is 1.7 good?*" Specifying what are good and bad numbers and constraining the coefficients' range may mitigate some of the confusions. Even with these changes, however, concept-based explanations may not be everyone's cup of tea. Some participants shared that they find the explanation form inherently overwhelming and less attractive (P5, P13, P16, P20). P16 shared: "*I sort of tune out with numbers after a while.*" P20 also expressed their preferences for more visual explanations: "*I'm such a visual person that stuff like this would go right over my head and make no sense for the most part.*"

Prototype-based explanations: Most preferred

Many participants picked prototype-based explanations as their favorite (P2, P3, P4, P6, P7, P9, P10, P12, P13, P15, P16, P17, P19, P20). The part-based form was clearly preferred, for similar reasons mentioned for concept-based explanations. P15 and P20 said prototype-based explanations are analogous to how they think about birds, and P1 that they are analogous to how birders teach each other. Between prototypes and concepts, participants tended to prefer prototypes for their visual nature and information content: prototype-based explanations locate and draw a box around relevant bird parts in the user-input photo, whereas concept-based explanations only list the bird parts. P4 summarized the advantages: "*It makes a very clear match between the photo that you're looking at and a larger base of what this bird should*

look like. It also skips over the whole language issue and is incredibly visual which I really appreciate." Participants also noted that prototype-based explanations can help many uses, e.g., learning how to identify new birds (P2, P8, P13, P15, P19, P20), understanding how the AI is working (P11, P13, P15, P16, P20), spotting the AI's mistakes (P4, P13), and changing their own behavior to supply better inputs to the AI (P20).

Complaints against prototype-based explanations were mostly minor. Some participants described the current version as "*cluttered*" and "*difficult to see*" (P1, P4, P5, P6, P11) and made UI design recommendations, e.g., having one prototype-photo region match pop up at a time (P11). Participants also mentioned that some prototypes were ambiguous (P2, P11, P18). For instance, P11 said they had to "*examine the prototype and the example to figure out what the concept was that they corresponded to.*" As a solution, P2 suggested providing a textual description of the prototype. Another complaint was that some prototypes (e.g., feet) were uninteresting (P1, P13, P18). "*Very few bird species are differentiated based on their feet,*" remarked P1. For solving this problem, participants suggested curating prototypes with domain experts and end-users so that the explanation focuses on salient and interesting features, those that human birders would use to identify birds.

Finally, several participants suggested combining prototype-based explanations with other approaches (P2, P4, P11, P12, P16, P18, P19). Concretely, P2 suggested combining it with heatmap-based, P2, P12, P16 and P18 with concept-based, and P4 and P11 with example-based explanations. P19 didn't specify an approach. Regarding the combination, some suggestions were general (e.g., show both types of explanations) while others were more specific (e.g., add concept labels to prototypes). P12 and P18 particularly advocated for using information from multiple sources (e.g., photo, sound, location) for both improving the AI's performance and explaining its results to end-users.

Concerns about AI explanations

Participants were overall excited to see AI explanations in Merlin, however, some expressed concerns regarding the faithfulness and potential negative effects of explanations. In particular, participants who were familiar with XAI questioned how faithfully the shown approaches would explain the app’s identification process, if they were to be implemented in the app (P6, P10). For example, P6 said example-based explanations feel like “*cheating interpretability*” unless the AI actually makes identifications using clustering or other techniques that group similar photos together. Regarding concept-based explanations, P6 and P10 asked if they imply that the AI system is *interpretable-by-design* and actually reasons in two steps (first concept recognition, then bird identification), or if they were *post-hoc* explanations produced by a separate “explainer” system. These questions highlight the importance and challenges of communicating what AI explanations are actually showing. In some cases, explanations of AI explanations (“meta-explanations”) may be more complex than the AI explanations themselves.

Another concern was that explanations might lead to mistrust or overtrust in AI systems. P20 said a convincing explanation for a misidentification would be “*detrimental*” to end-users who are trying to learn bird identification on their own, because they might more readily believe in the misidentification and accumulate wrong knowledge. Similarly, P19 said explanations might encourage end-users to “*double down on the incorrect identification*,” and even create a negative feedback loop if the AI system relies on end-users to input or verify data. These concerns are consistent with findings from recent research [190, 312] that people tend to believe in AI outputs when given explanations for them, and raise caution against negative effects explanations might have on end-users, irrespective of XAI designers’ intent.

3.6 Discussion

3.6.1 XAI as a medium for improving human-AI collaboration

The most surprising finding from our study was the degree to which Merlin end-users wanted to improve their *collaboration* with the AI system through AI explanations. Participants desired information upon which they can act. In particular, they wanted AI explanations to *help them help the AI*, e.g., by supplying better inputs to the AI and providing constructive feedback to developers. We found this an intriguing repurposing of explanations, which are typically developed to help people understand AI’s inner workings and outputs. Collaboration is distinct from *usability*. Usability is already often discussed in the XAI literature, where XAI is presented as a means to provide meaningful information about how an AI output is reached, so that users can “make decisions more quickly, or to increase decision quality” [218]. However, our participants desired information that not only improves their decisions based on the AI’s outputs (usability), but also empowers them to help the AI be more accurate in the first place and achieve better outcomes together (collaboration).

For designing XAI that supports human-AI collaboration, research in the accessibility field can be instructive. Bigham and colleagues’ work on the VizWiz system [41] combined computer vision with human labor to support blind end-users in solving daily visual problems. The VizWiz system allowed end-users to upload a photo of their environment and ask remote workers visual questions (e.g., where is the tomato soup can?). In their study, the authors found that input photos presented challenges to the VizWiz system because the blind end-users provided photos which were “often blurred, tilted, varied in scale, and improperly framed .. and susceptible to problems related to lighting” [41]. To overcome these challenges, the authors developed a subsystem that uses computer vision to provide end-users with prompts around lighting,

centering, and distance-to-object. These prompts helped end-users take photos that help remote workers better answer their questions. Like how VizWiz’s subsystem helped end-users collaborate with VizWiz and remote workers, we argue AI explanations can and should serve as a medium for improving end-users’ collaboration with AI systems. In addition to providing technical transparency, AI explanations should move towards providing *actionable feedback* to and from end-users, empowering end-users to have more rich and meaningful interactions with AI systems.

3.6.2 XAI design for Merlin

So what should Merlin’s AI explanations look like? In this section, we propose a design based on participants’ perceptions of the four XAI approaches (Sec. 3.5.3 and Tab. 3.2). We hope this serves as a helpful example of how end-users’ feedback can inform and guide the design of AI explanations.

Participants preferred specific, part-based explanations that resemble human reasoning and explanations. Participants repeatedly demanded that explanations highlight a few specific features that the AI uses to make its identifications. XAI approaches that best satisfied this need were concept and prototype-based explanations that break down the AI’s output with human-digestible units of information, i.e., concepts and prototypes. Participants also appreciated how similar concept and prototype-based explanations were to the way human birders identify birds and explain their identification to others. Heatmap and example-based explanations were generally less preferred: many participants noted that they were too coarse and impression-based to be useful.

At the same time, participants wanted explanations to be easily understandable. “*I don’t want to spend extra time trying to understand what I’m looking at*” (P5) was a common sentiment. However, what counts as easily understandable differed between participants. For instance, participants with low-AI background found concept-based

explanations overwhelming due to the numbers and calculations. Contrarily, participants with high-AI background found all explanations intuitive, with some wanting even more numbers and other technical details. These findings underline the importance of our research agenda, as end-users’ XAI needs, perceptions, and variations therein cannot be anticipated by studying only AI experts, which have made up a primary study population in XAI research.

Based on these observations, if we were to design AI explanations for Merlin, we would combine prototype and concept-based. For simplicity, we would start with a visualization of the user’s input photo overlaid with boxes indicating regions matched to prototypes. We would then give the user an option to tap on each box to get more information. The pop-up would show the prototype matched to the boxed region, accompanied with a short description, i.e., concept name. As some participants mentioned, prototypes may be ambiguous. For instance, a prototype visualizing a part of a bird’s wing may not be informative if end-users don’t know what is special about that part. A text description (e.g., white stripes across the folded wing) would help remove the ambiguity. Further, we would curate the set of prototypes for each bird with end-users and domain experts. We would exclude prototypes that are uninteresting (e.g., feet) or redundant, and limit the total number of prototypes so as to not overwhelm users. Finally, we would optionally provide similarity scores, coefficients, and final class scores for end-users who are interested in diving into the details.

3.6.3 Implications for future XAI research

Our findings have broader implications for XAI research beyond designing explanations for Merlin. Below we discuss directions we believe AI explanations should be improved to better serve the needs of AI system end-users.

1. *Explanations should be designed with end-users.* Participants often exposed blind spots in existing XAI approaches, revealing a creator-consumer gap in XAI

research [111]. For example, they pointed out that the concepts used in concept-based explanations were disconnected from birders' language (Sec. 3.5.3). The shown concepts (e.g., white body, long wings) were too generic compared to birders' field mark terms (e.g., wingbar, supercilium). Participants also proposed solutions, e.g., develop the bank of concepts with end-users, and offered to contribute their experience and domain expertise. This example highlights the need for end-users' *participation* in the explanation design process and calls for more participatory approaches [282] to XAI research.

2. *Explanations should answer “why” not just “what.”* Several of our participants were unsatisfied with existing XAI approaches that only explain “what” features the AI system was using to produce its output, e.g., heatmap explanations that highlight “what” image regions were important but don't explain “why” those regions were important (Sec. 3.5.3). They expressed a desire for explanations that answer “why” question so that they can gain a deeper understanding of the AI's reasoning process. Explaining causal relationships in computer vision AI models is an open problem, however, it is promising that more researchers are tackling causal questions in XAI research [279].
3. *Explanations should use multiple forms and modalities if warranted.* There is no reason for explanations to be limited to one form or modality. Participants often suggested combining two or more XAI approaches to produce more informative explanations (Sec. 3.5.3). They also questioned why Merlin's identification features (i.e., Photo ID, Sound ID) and our explanation mock-ups were not multimodal, when human birders combine evidence from as many sources as possible (e.g., photo, sound, location) for more accurate bird identification. Expanding the design space of explanations will lead to XAI methods that better satisfy end-users' needs.

4. *Explanations should be rigorously evaluated.* Explanations sometimes have (unintended) negative effects. Recent works have revealed that explanations can engender over-trust in AI or give misleading understandings [5, 111, 154, 190, 234, 257, 287, 345]. Our participants were also concerned about the faithfulness and potential negative effects of explanations (Sec. 3.5.3). To preemptively, not reactively, address these issues, it is crucial to rigorously evaluate XAI methods on both their method goals and use case goals throughout the development process.

3.7 Limitations and future work

First, as most of our interview questions and study materials pertain to the Merlin app, our findings may not generalize to other contexts. This is an intentional trade-off made in favor of deepening our understanding of end-users' XAI needs, uses, and perceptions in a specific context. However, our study design can aid future research on other types of AI applications and the disparate domains and contexts into which they are integrated. Another limitation is that we had relatively few participants in some of the background subgroups; in future research we aim to have more participants from these subgroups. Finally, we did not have access to other stakeholder groups, such as the developers and deployers of the Merlin app. However, different stakeholders in AI systems might have different needs for AI explanations [131, 218]. We plan to pursue additional research with a more comparative design and study if and how needs differ across stakeholder groups.

3.8 Conclusion

There is a lack of empirical studies examining end-users' explainability needs and behaviors around AI explanations in real-world contexts. This is important to address in order to make AI systems accessible, usable, and understandable for more people. In

this work, we conducted a qualitative, descriptive, and empirical study with 20 end-users of the Merlin bird identification app. Our questions revolved around real-world needs and usage, with a goal to surface insights which can be utilized to improve the design of AI explanations. We found interesting insights into the actionability with which participants collaborate with the AI system and desire to use explanations to improve their collaboration. We also gathered concrete feedback on four representative XAI approaches that could be potentially embedded into Merlin, finding that participants prefer part-based explanations that resemble human reasoning and explanations. Further, participants' feedback revealed a creator-consumer gap in XAI, highlighting a need of involving end-users in the XAI design process. Based on our findings, we provided recommendations for future XAI research and design.

Acknowledgments

We foremost thank our participants for generously sharing their time and experiences. We also thank Tristen Godfrey, Dyanne Ahn, and Klea Tryfoni for their help in interview transcription. Finally, we thank members of the Princeton HCI Lab and the Princeton Visual AI Lab, especially Amna Liaqat and Fannie Liu, and the anonymous reviewers for their thoughtful and helpful feedback. This material is based upon work partially supported by the National Science Foundation (NSF) under Grant No. 1763642 awarded to OR. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. We also acknowledge support from the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award (OR), Princeton SEAS Project X Fund (RF, OR), Princeton Center for Information Technology Policy (EW), Open Philanthropy (RF, OR), and NSF Graduate Research Fellowship (SK).

Chapter 4

Understanding end-users' trust in a real-world AI application

Trust is an important factor in people's interactions with AI systems. However, there is a lack of empirical studies examining how real end-users trust or distrust the AI system they interact with. Most research investigates one aspect of trust in lab settings with hypothetical end-users. In this chapter, we provide a holistic and nuanced understanding of trust in AI through a qualitative case study of a real-world computer vision application. We report findings from interviews with 20 end-users of a popular, AI-based bird identification app where we inquired about their trust in the app from many angles. We find participants perceived the app as trustworthy and trusted it, but selectively accepted app outputs after engaging in verification behaviors, and decided against app adoption in certain high-stakes scenarios. We also find domain knowledge and context are important factors for trust-related assessment and decision-making. We discuss the implications of our findings and provide recommendations for future research on trust in AI.

This chapter contains materials published in the FAccT 2023 paper, Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision

Application [191], in collaboration with Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández.

4.1 Introduction

Trust is an important factor in people’s interactions with Artificial Intelligence (AI) systems. For the effective adoption and use of these systems, people must trust them appropriately. Both unwarranted trust (trusting when the AI system is not trustworthy) and unwarranted distrust (distrusting when the AI system is trustworthy) can hurt the quality of interactions [24, 167, 270, 413]. To better understand trust and foster it appropriately in human-AI interactions, recent works have started to investigate questions such as: What does it mean to trust an AI system? [136, 167, 373] How is trust established and developed? [177, 230] What factors influence people’s trust and how? [77, 146, 212, 332, 425, 426, 437].

Trust in AI research, however, is still in a nascent stage. As noted in recent surveys [379, 386], papers often use different definitions of trust, making their results difficult to compare. There is also little agreement on how to empirically study trust, e.g., when to use subjective vs. objective measures. Finally, there is a lack of research that approaches trust holistically. Most papers study one specific aspect of trust (e.g., whether explainability increases people’s trust in AI [26, 57, 77, 190, 198, 212, 287, 288, 332, 425, 432]) in lab settings with hypothetical end-users. While they provide valuable insights, they do not capture the complex nuances of trust in real-world contexts.

The FAccT community is increasingly focusing on trust because understanding and measuring it overlap with evaluating fairness, accountability, and transparency of algorithmic systems. Work around those principal interests has begun to mature, as scholars begin to conduct more empirical research on factors and hu-

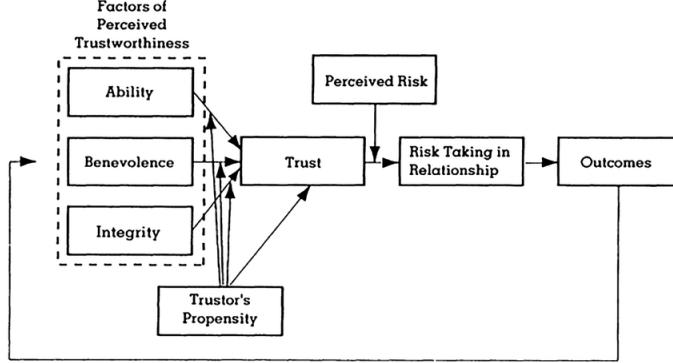


Figure 4.1: **Trust model of Mayer et al. [261]**. Based on this model, we separate *trust* from *trustworthiness* perceptions that precede it, and two trust-related behaviors that proceed from it: (1) AI *adoption*, which corresponds to *risk taking in relationship* in the model, and (2) AI *output acceptance*, which corresponds to *outcomes* evaluation in the model. In this work, we describe both *general* trustworthiness perceptions and trust attitudes, and *instance-specific* trust-related behaviors. See Section 4.2.1 for further discussion. This figure is replicated from the work of Mayer et al. [261].

man perceptions of fairness [97, 144, 180], and going even further, to recognize the influence of sociotechnical context on factors of fairness, ethics, and accountability [140, 266, 329, 338, 353]. However, papers on trust recently published at FAccT remain largely theory-focused [118, 167, 197, 230, 369, 373]. Our work contributes to the community by maturing the understanding of trust through an in-depth empirical study, moving the conversation as AI systems move from theoretical, lab-based projects out into the real world.

The goal of this work is to provide a more holistic and nuanced understanding of trust in AI through a qualitative case study of a real-world AI application. We ground our study in Merlin [368], a free mobile phone app that uses computer vision AI models to identify birds in user-uploaded photos and audio recordings (Figure 4.2). We chose Merlin because it is a widely-used app that allows us to connect with a diverse set of active end-users with varying levels of domain (bird) and AI background, and satisfies the requirements of commonly-used trust definitions [220, 261]. Concretely, we conducted semi-structured interviews with 20 Merlin end-users and

inquired about their trust in the app from several angles.¹ Since we were one of the first to talk to actual end-users about their trust relationship with the AI application, we focused on exploring what factors influence trust and how, rather than quantifying the importance of certain pre-specified factors. Afterwards, we applied Mayer et al.’s theoretical definition and model of trust [261] (Figure 4.1) to the collected empirical data, to delineate and describe multiple aspects of trust and their influencing factors.

We make three key contributions in this work: (1) We further the FAccT community’s understanding of trust, to date still heavy in theory, through an in-depth qualitative empirical study. Specifically, we study how end-users trust AI in a real-world context and what factors influence their trust. In doing so, we synthesize theoretical and empirical trust research by applying Mayer et al.’s theory [261] to the empirical data we collected. This approach will, we hope, yield insights into how readily existing theories of trust can be operationalized for empirical research. (2) We provide a more holistic and nuanced understanding of trust, as compared to the current state-of-the-art in the field. We find general trustworthiness perceptions and trust attitudes are distinct from instance-specific trust-related behaviors. While our participants told us they perceived the app as trustworthy and trusted it, they also described how they selectively accepted the app outputs after engaging in verification behaviors, and sometimes decided against app adoption in certain high-stakes scenarios. Domain knowledge and context were particularly important factors for participants’ trust-related assessment and decision-making. (3) Finally, we discuss the implications of our findings and provide practical recommendations for future research on trust in AI. Most critically, we advocate for researchers to define and delineate trust from related constructs, and to consider human, AI, and context-related factors of trust together.

¹In the same interviews, we also inquired about participants’ AI explainability needs, intended uses of AI explanations, and perceptions of existing explanation approaches, and analyzed that portion of the data in another paper [192].

4.2 Background and related work

4.2.1 Definitions and models of trust in AI

Trust has a long history of research. Before “trust in AI” was researched, disciplines ranging from psychology to philosophy to human factors had studied trust in the context of relationships between humans or between humans and machines. As such, there are numerous definitions and models of trust, but most have their roots in two papers: Mayer et al.’s 1995 paper on organizational trust [261] and Lee and See’s 2004 paper on trust in automation [220].

Within trust in AI research, many works do not state a definition of trust, according to recent review papers of the field [136, 379, 386]. Among works that do, the most commonly used definitions come from the aforementioned papers: (1) “the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party” by Mayer et al. [261] and (2) “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” by Lee and See [220].

Both definitions share the same key elements, as described by Vereschak et al. [386]: (1) *vulnerability*: the situation involves uncertainty of outcomes and potential negative consequences; (2) *positive expectations*: the trustor thinks that negative outcomes associated with trusting do not exist or are very unlikely; and (3) *attitude*: the general way of thinking and feeling, typically reflected in a behavior, although not a behavior itself. These elements also distinguish trust from other related constructs. It is not *trust* but: *confidence* when there is no vulnerability; *distrust* when there is no positive expectation; *compliance* or *reliance* when referring to a behavior; and *perceived trustworthiness* when referring to a perception of a trustee’s characteristics upon which trustors form their trust. For example, when a

study measures whether participants follow AI’s advice, which is a directly observable behavior, it is measuring *reliance* not *trust*. When a study asks participants to rate their trust level on a survey scale, it is measuring attitude, except when there is no vulnerability (e.g., lab experiment with no incentives or risks), it is measuring *confidence* not *trust*. We hope these distinctions reduce ambiguity and confusion around the terms.

Recently, scholars have proposed specific definitions and models for “trust in AI” [167, 230]. Jacovi et al. [167] formalized trust in AI as “contractual”: to trust an AI system is to believe that it is trustworthy to uphold some contract. Their formalization disentangles trust and trustworthiness, and defines “warranted trust” as trust that is “caused” by the AI’s trustworthiness. Liao and Sundar [230], on the other hand, took a communication perspective and proposed a model that describes how the trustworthiness of AI systems is communicated through trustworthiness cues and how those cues are processed by people to make trust judgments.

In this work, we adopt the model of trust by Mayer et al. [261], despite it being developed for organizational trust, because the model’s definition and process orientation fit our work’s objective of holistically understanding trust, in the context of human-AI interaction. Mayer et al. [261] delineate trust from its antecedents, context, and products, and describe how different components influence each other as the trustor interacts with the trustee (Figure 4.1). Based on their model, we separate trust from trustworthiness perceptions (trustworthiness being trust’s antecedent) and trust-related behaviors, i.e., output acceptance and adoption decisions. The models by Jacovi et al. and Liao and Sundar [167, 230] are less fitting for our work because Jacovi et al. [167] focus on formalizing prerequisites, causes, and goals of trust in AI, and Liao and Sundar [230] focus on modeling the communication of trustworthiness.

4.2.2 Empirical studies of trust in AI

Trust in AI is a fast-growing research field with significant empirical work. However, there is surprisingly little research on how end-users trust AI in *real-world contexts*, and *what factors* influence their trust—two gaps our work aims to fill. Filling these gaps is an important endeavor because while AI systems may perform well in controlled lab settings, their take-up and use in the real world are subject to various factors, many of which are context-dependent and currently under-anticipated in research and system design.

Much of prior work focus on understanding the effect of certain *pre-specified factors* on trust [26, 77, 146, 190, 198, 212, 287, 288, 312, 332, 399, 425, 426, 437]. Most utilize *lab experiments*, usually with participants recruited from crowdsourcing platforms (e.g., MTurk [26, 77, 190, 212, 287, 288, 312, 332, 425, 437], Prolific [287, 288], internal platform [198]). Typically, these works start with a hypothesis (e.g., explainability will increase trust in AI). To investigate the hypothesis, they choose a measure of trust (e.g., self-reported rating on a 1-7 scale [172]), make a change to the factor of interest in the design of the AI system (e.g., show an explanation of the AI’s output), and then quantify the effect of that change on participants’ trust. Based on the results, they conclude the effect of the factor of interest on trust.

The most commonly studied factors in the literature are transparency and explainability. However, researchers operationalize these factors in several ways. For instance, *transparency* is operationalized as providing model internals (e.g., learned coefficients in a linear regression model) in [312], overall performance measures (e.g., accuracy) in [198, 212, 332, 425, 426], confidence scores on individual outputs in [287, 437], and visualizations of input data distributions and feature engineering process in [102]. Similarly, while [26, 57, 77, 190, 198, 212, 287, 288, 332, 425, 432] all study the effect of *explainability* on trust, the operationalized explanations of AI’s be-

havior and outputs greatly vary in approach (e.g., feature attribution, counterfactual examples) and form (e.g., heatmap-based, part-based).

These works provide insights into the relationship between trust and the factor of interest, as operationalized in a specific and controlled way. However, they do not capture the contextual aspects of trust, and the design of these studies does not allow for discovering new trust-influencing factors. To address these two gaps, we conducted a qualitative case study of a real-world computer vision AI application and interviewed its end-users about their trust relationship with the AI. While resource-intensive, interviews enabled us to explore multiple aspects of trust in depth and identify trust-influencing factors in a bottom-up manner. The value of qualitative case studies has been demonstrated in recent works [114, 340, 410]. In one example, Widder et al. [410] conducted a case study investigating what factors influence engineers' trust in an autonomous software engineering tool in a high-stakes workspace. They found that trust, in their study setting, was influenced by the tool's transparency, usability, social context, and the organization's associated processes. Widder et al.'s work laid groundwork for our own qualitative study, as we applied their methods to ask similar questions about a different population in a different domain: end-users of an AI-based bird identification app. Our work provides complementary insights, and we encourage the community to conduct more qualitative case studies of trust in AI.

4.3 Methods

In this section, we describe our study methods. All were reviewed and approved by our Institutional Review Board.

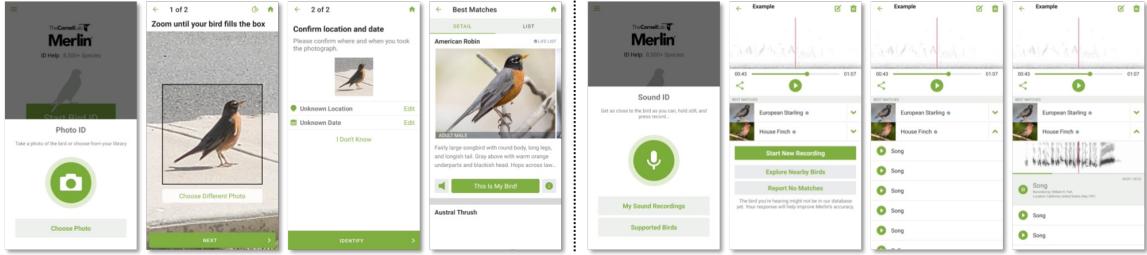


Figure 4.2: **Our study application Merlin [368]**. Merlin is a popular AI-based bird identification mobile phone app. Users upload photos on the Photo ID feature (left) or audio recordings on the Sound ID feature (right), with optional location and season data, and get a list of bird(s) that best match the input. See Section 4.3.1 for more details about the app.

4.3.1 Study application: Merlin Bird ID app

To study trust in AI in a realistic setting, we looked for a research setting that first, involves real-world AI use by end-users who range in their domain and AI background, and second, satisfies the requirements of widely-accepted trust definitions [220, 261]. We found Merlin [368] (Figure 4.2) to satisfy both conditions. First, Merlin is a mobile phone app that identifies bird species from user-input photos and/or audio recordings. It is an *expert* application with expertise that most people do not have, i.e., knowledge and skill to identify thousands of birds. As a free app with over a million downloads, it is used by people with diverse domain (bird) and AI backgrounds, thus satisfying our first requirement. Second, while Merlin is generally a *low-stakes* application, there are some amount of *vulnerability* and *positive expectations* in its use, as we verify in Section 4.4.1. This allows us to characterize end-users' *attitude* toward the app as trust, and study “trust in AI” and its influencing factors.

4.3.2 Participant recruitment and selection

We recruited participants who are active end-users of Merlin Photo ID or Sound ID, the app’s AI-based bird identification features, with considerations for diversity in the domain and AI background. Concretely, we created a screening survey with

questions about the respondent’s background and app usage pattern (e.g., regularly used features, frequency of use). We posted the survey on various channels: Birding International Discord, AI for Conservation Slack, several Slack workspaces within our institution, and Twitter. On Twitter, in addition to posting the survey, we reached out to accounts with tweets about Merlin via @mentions and Direct Messages. Based on the screening survey responses, we selectively enrolled participants to maximize the diversity of the study sample’s domain and AI background (Table 4.1). All participants were active end-users of Merlin who could provide vivid anecdotes of when the app worked well and when it did not. Regarding the frequency of use, 11 participants used it several times a week, 8 used it once a week, and one used it once a month.

4.3.3 Interview sessions and analysis

We interviewed 20 participants, over a Zoom video call, from July to August 2022. The interviews lasted one hour on average, and we compensated participants with their choice of a 30 USD gift card or a donation to a bird conservation organization made on their behalf. In the same interviews, we also inquired about participants’ explainable AI needs, uses, and perceptions, and analyzed that portion of the data in another paper [192]. In this work, we focus on understanding what factors influence participants’ trust in AI and how. Below we describe the relevant part of the interview protocol. See the supp. mat. for the interview questions.

We began each interview by introducing the study to the participant, communicating that we were not affiliated with the Merlin development team, and receiving consent for participation in research. We then asked about their domain and AI background, as well as their goals and stakes in their app use.

Next, we inquired about the participant’s perception of, experience with, and trust in the app. Regarding trust, we adopted Benk et al.’s trust enablement paradigm [36] and asked participants to describe their trust relationships with the app in their own

terms. We asked about general perceptions and attitudes, such as how accurate and trustworthy they find the app, as well as specific instances, such as how they assess the correctness of the app outputs, and in what circumstances they decide to use the app and not. Scoping down our unit of analysis, from the system as a whole to the “instance” of use, provided a way to gather dynamic data from our participants about which contextual factors they considered during their trust-related decision-making. Finally, we asked the participant whether they would adopt the app in hypothetical high-stakes scenarios with health-related and financial outcomes:

1. *Sick bird scenario*: “Suppose you find a sick bird and take it to the vet. The vet is not sure what bird it is. Would you recommend Merlin to identify the bird species so that the vet can determine the course of treatment?” We asked the participant to view Merlin as a decision-support tool as the participant and the vet will make the final call.
2. *Game show scenario*: “Suppose you enter a game show where you can win or lose money based on how well you can identify birds from photos or audio recordings. You can only use one resource among Merlin, books (e.g., field guides), the Internet (e.g., search engine, online birder community), and so on. Which resource would you use? Does your answer change depending on certain factors?”

We designed these scenarios to introduce high stakes into the AI adoption decision. These scenario-based inquiries allowed us to observe how participants’ trust-related assessment and decision-making differ across usage contexts.

We transcribed the interviews and analyzed the transcripts using abductive coding. Tavory and Timmermans [365] describe abductive coding as an iterative process, moving between empirical data and available theory, in order to ensure findings are informed by, in dialogue with, and can contribute to, social-science literature. We first

Table 4.1: **Participants’ domain (bird) and AI background.** See the supp. mat. for a description of the categories.

	Low-AI	Medium-AI	High-AI
Low-domain	P7, P12, P16	P8, P14	P11, P13
Medium-domain	P2, P20	P1, P4, P10	P6
High-domain	P5, P17	P3, P9, P15	P18, P19

read five transcripts to identify an initial set of empirical observations through which we could develop a theoretical hunch about the nature of trust-related perceptions, attitudes, and behaviors. We then consulted the literature on trust and trustworthiness, at which point we found Mayer et al.’s trust model [261], which provided a high-level framework and starting point. Using this model, we collectively developed a codebook with which we could analyze our initial observations and discern which of our theoretical hunches were novel contributions. Finally, we coded all of our data, discussed the results, and drew out themes.

4.4 Results

We start by discussing the definition of trust in our study context (Section 4.4.1) and then present multiple aspects of participants’ trust in the app: trustworthiness perception and trust attitude (Section 4.4.2); acceptance of individual AI outputs (Section 4.4.3); and AI adoption decision (Section 4.4.4). Throughout, we note the factors influencing participants’ trust, and close with a summary of the results (Section 4.4.5).

4.4.1 Prelude: Is “trust” the right term?

Before diving into the findings, let us first examine if “trust” is the right term for describing participants’ attitudes toward the app. Recall that trust is defined as an *attitude* and requires *positive expectations* and *vulnerability* in the trustor-trustee relationship [220, 261, 386]. It is easy to see that participants had *positive expectations*:

they were actively using the app because they expected it to help them achieve their goal of accurately identifying birds. However, is there *vulnerability* in use of this everyday app for bird identification? We answer yes because the app is used in situations involving *uncertainty of outcomes* and *potentially negative consequences*, satisfying [386]’s definition of *vulnerability*.

First, bird identification is a challenging task that requires the selection of a species among approximately 10,000 existing bird species, some of which are markedly similar to each other. Even though the app has been developed by bird and AI experts and trained on a large database of expert-annotated bird photos and audio recordings, it is not foolproof. There is always *uncertainty* about whether it would return an accurate identification, which participants were aware of. See Section 4.4.2 for detailed accounts of how participants perceived the app’s ability and trustworthiness.

There are also *potential negative consequences* when the app makes a misidentification. We heard the following responses when we asked participants what they gain and lose when the app succeeds and fails on the task. As gains, participants mentioned curiosity satisfaction (All), joy (P1, P7, P9, P12), bird knowledge (P4, P5, P8, P9, P10), and improved birding experience (P1, P2, P3, P4, P10). As losses, although several participants said “*nothing material*” (P1, P3, P11, P12, P13, P15, P16, P17), many expressed that they feel “*disappointed*”, even “*frustrated*”, when the app fails because they really care about correctly identifying birds and would like to gain accurate knowledge about birds (P1, P4, P5, P6, P9, P10, P13, P15, P18, P19, P20). Some noted that misidentifications can lead to people gaining wrong knowledge, (unintendedly) sharing misinformation by reporting wrong bird sightings, and negatively impacting science and conservation efforts (P4, P5).

In summary, there were *positive expectations* and *vulnerability* in participants’ use of the app, although there were individual differences in the amount of stakes participants placed in their use. Hence, we conclude “*trust*” is the right term for

describing participants' *attitudes* toward the app. With this established, we now proceed to describe participants' trust in AI in three parts.

4.4.2 Trustworthiness perception and trust attitude

We begin unpacking participants' trust in AI by explaining how they assessed the app's trustworthiness, a key antecedent to trust in Mayer et al.'s model [261]. Overall, participants assessed the app to be trustworthy and trusted it. We draw this conclusion based on participants' responses regarding the app's ability, integrity, and benevolence—the three factors of perceived trustworthiness in Mayer et al.'s model (Figure 4.1). Participants assessed that the app possesses all three, based on their prior experience with it, its popularity, and the reputation of the domain and the developers.

Participants assessed ability based on their prior experience with the AI and the AI's popularity

Ability refers to the trustee's skills and competencies [261]. For automation systems, Lee and See [220] describe it as performance, i.e., how well the automation is performing. Participants were overall impressed with the app and judged it to have high ability. Most described the app as very successful and that it seemed to be correct 9-10 out of 10 times, based on their **prior experience** with the app.² Exceptions were P2, who said Sound ID often made mistakes, and P6, who was disappointed with Photo ID. Most other participants were very impressed and described the app as “*pretty insane*” (P15), “*perfect*” (P11), and gave high praise despite having ob-

²Recent works [24, 336] suggest there may be a gap between the actual and perceived trustworthiness of an AI system. In our study context, however, participants' assessment of the app's ability seems reasonably accurate. Our judgment is based on public knowledge, as we are not affiliated with the app development team. Regarding Photo ID, one of the developers said in 2016 that its “accuracy is around 90 percent if the user's photo is of good quality” [307], and we expect the performance would have increased. Regarding Sound ID, recent research from the Cornell Lab of Ornithology suggests that AI models are highly capable of sound-based bird identification [174].

served mistakes: “*I’ve had one or two times where I’ve thought I don’t believe that’s really that bird? [...] But I trust it. I trust it*” (P14). Intriguingly, some participants mentioned they could not accurately assess the app’s ability due to their lack of **domain knowledge** (P11, P12, P13). For instance, P11 said: “*As far as I know, it’s been perfect, but I don’t know enough to know if it would be making mistakes.*” We discuss this point further in Section 4.4.3, where we describe how participants assessed the correctness of individual app outputs. Finally, while most assessed the app’s ability based on their own prior experience, P12 made an assessment based on the app’s **popularity**, which is an external factor: “*I imagine that if it has such a wide user base, it would be pretty accurate*” (P12).

Participants assessed integrity based on the developers’ reputation

Integrity refers to the degree to which the trustee adheres to a set of principles that are acceptable to the trustor [261]. For automation systems, Lee and See [220] describe it as a process, i.e., in what manner and with which algorithms it is accomplishing its objective. We found that participants believed in the app’s integrity because of the **reputation of the developers**, the Cornell Lab of Ornithology, which is a respected institution with a long history of bird science and conservation efforts. Most participants were well aware that the app was developed by this lab (P1, P3, P4, P5, P8, P9, P11, P12, P13, P14, P15, P17, P18, P19). Participants were also familiar with the lab’s other apps (e.g., eBird, BirdNET, iNaturalist) and resources (e.g., All About Birds, Macaulay Library), describing these and the app as their “*go-to*” (P15) when they want to learn about a specific bird. P14 specifically said they trusted the app because it was developed by this lab: “*I know that Cornell Ornithology Lab does excellent, excellent stuff. I mean, if you’re going to try and learn anything about a bird, just go there. Don’t try anything else. Don’t even bother the Audubon Society. Just go straight to Cornell. So I trusted it [the app] for that reason.*” Participants

did not know how the app was developed or how it works, since such information is not publicly available. Nonetheless, they believed in the app's integrity because they believed in the authority and expertise of the app developers.

Participants assessed benevolence based on the domain's reputation

Benevolence refers to the extent to which the trustee's motivations are aligned with the trustors' [261]. For automation systems, Lee and See [220] describe it as purpose, i.e., why the automation was built originally. We found that participants believed in the app's benevolence because of the positive **reputation of the domain**, i.e., the birding community that they and the app developers are part of. For instance, P18 described the birding community as a place where everyone tries to be accurate and do good: "*I think birders, in general, are a community where there's very few people who try and do adversarial attacks because it doesn't benefit anybody [...] the value of the birding community is that everybody is trying to be accurate.*" Some participants contrasted the app with other AI applications. For instance, P2 described the app as not having "*malicious intent*" compared to advertisements. P9 contrasted the app with other AI applications they found "*creepy*" and "*harmful*," such as voice assistants that may be "*monitoring*" user behavior. Although the app may also collect personal location data, participants seemed less concerned overall. P18 and P19 even wanted their data to go somewhere and be used (e.g., for science and conservation efforts or improving the app) so that it is "*contributing something to society*" (P19).

4.4.3 Output acceptance

Participants described the app as trustworthy and trusted it; however, they did not accept its outputs as true in every single instance of use. To the extent possible, participants carefully assessed the app's outputs, using their knowledge about the domain, and then made acceptance decisions. Our findings reveal a gap between *general trust*-

worthiness perceptions and trust attitudes, and *instance-specific* trust-related behaviors, highlighting the importance of considering both aspects in trust in AI research.

Participants assessed AI outputs based on likelihood and task difficulty

Participants had developed heuristics for judging the correctness of app outputs (P1, P2, P6, P16, P17, P19). One heuristic was assessing the **likelihood** of spotting a bird species in a given area. Participants described they were more trusting of the app's output when the identified species is common for the area, and less trusting when it is rare. For instance, P6 said they judge the output's correctness based on recent sightings in the area and rarity: “*If it's a common bird or even just a rare bird, uncommon or something like that, then maybe [it is correct]. But if it's a super rare bird, then definitely not.*” Another heuristic concerned **task difficulty**. P1’s response well explains this heuristic: “*I trust it [the app] more when I know that I'm looking at something that should be relatively unambiguous. If I'm looking at something that's like a Female Warbler or a Female Sparrow, which might just be a little brown bird, then I'm a little bit more skeptical of the result.*” For context, “*little brown bird*” is a term used by birders to describe a large number of species of small brown passerine birds, which are known to be notoriously difficult to distinguish. P1 described them as “*really hard to ID, even for a human ornithologist.*” We note that both heuristics require **domain knowledge** as participants must know whether the identified species is common or rare for the area, and which birds are difficult and easy to identify.

Participants verified AI outputs with input-output comparisons and information from other sources

Some participants verified the outputs via **input-output comparisons**, i.e., they compared input photos and audio recordings to reference photos and audio recordings of the identified species, which are provided in the app (P1, P10, P20). For instance,

P10 verified Photo ID outputs as follows: “*I go back and look at reference photos of that bird and then try to map field markings. So I’ll try to see if the bill shape lines up.*” P20 described a similar process for verifying Sound ID outputs: “*They have that drop down [menu] that has the different sounds, and I will literally just play one until I find one that I’m like: oh it’s that, that’s the exact sound.*” In short, participants looked for the resemblance between the inputs and the references, and then assessed the outputs’ correctness. This verification does not require **domain knowledge** per se; however, participants with it could more easily verify the output as they would know what to check.

Participants also described using **information from other sources** (P1, P4, P10, P15, P18). If the app identified a bird based on sound, participants tried to confirm it with their own visual identification, and vice versa (P10, P18). For instance, P10 said they try to visually confirm the bird when using Sound ID: “*I’ll also look for that bird and see if I can see it as well. Or see if it matches a bird that I have seen.*” Some participants took a step further and consulted other birders, through their personal networks or online communities (P1, P4, P15). P1 said they often share the app’s outputs with friends and birders online. They also remarked, “*If there was something that I knew was more of an ambiguous result, I would usually consult somebody*” (P1). P4 was an especially active member of online communities. They said they ask questions on Discord and Reddit and consult expert birders in the area: “*I do oftentimes reach out to people whose names I find on e-bird. I see local checklists and I’ll either find them on Facebook or LinkedIn or something and send them a message.*” Note that cross-checking requires **domain knowledge** for identifying birds on their own, whereas consulting other birders does not.

Some participants disregarded AI outputs when they could not verify

For some participants, verification was a crucial and necessary step for output acceptance (P3, P4, P15, P18). When unable to verify, they disregarded the app output. For example, P15 said they've never only relied on the app when identifying a bird they have not seen before. They almost always sent the output to more experienced birders and received their confirmation. P18 was also strict about when they accept app outputs, stating, “*I never, I never count on my bird registry anything that Sound ID says that I can't kind of confirm either through the facts of it or through a visual ID*” (P18). These participants disregarded unverifiable app outputs, despite their positive assessment of the app’s ability and trustworthiness, revealing a gap between general trustworthiness perceptions and instance-specific trust-related behaviors.

Not all participants had the ability to assess the correctness of AI outputs

So far we described various processes through which participants decided whether or not to accept app outputs. However, not all participants had the **ability to assess** the correctness of app outputs. In Section 4.4.2, we described how some participants with little **domain knowledge** said they could not accurately assess the app’s ability (P11, P12, P13). These participants also said that because they “*know so little about birds*” (P12), they could not “*validate or reject*” (P11) app outputs, especially if they can’t get information from other sources. P13 said, “*If it's misidentifying a bird that I can't see, then I have no way to know that.*” This finding suggests that domain knowledge is a key factor in appropriate trust calibration and has a wide influence on participants’ interactions with the app.

4.4.4 Adoption

The final aspect of trust in AI we analyze is how participants made AI adoption decisions. We compared participants’ decision-making process between their actual use

setting and two hypothetical high-stakes scenarios (see Section 4.3.3 for the scenario details). We found that while participants always used the app in their actual use setting, they made different adoption decisions for the high-stakes scenarios based on various factors: the app’s ability, familiarity, and ease of use (AI-related factors); participants’ ability to assess the app’s outputs and use the app (Human-related factors); and finally, task difficulty, perceived risks and benefits of the situation, and other situational characteristics (Context-related factors).

In their actual use setting, participants never decided against using the AI

We found that participants always use the app when opportunities arise. It is not that participants absent-mindedly used the app. Participants were aware of when the app works well and not, and knew how to help the app be more successful, e.g., by supplying better input photos and audio recordings. However, when we asked how they make app adoption decisions, they only described situations where they decided to use the app, and never situations where they decided against using it.

There could be several reasons for this finding. First, the app has a low cost of use. Since the app is free, the only use costs are the time and effort involved in taking photos or audio recordings and inputting them into the app, and perhaps a small amount of phone battery. Second, the risks of use are also low. There are potential negative consequences when the app misidentifies, e.g., gaining wrong knowledge, as described in Section 4.4.1. However, end-users can mitigate these risks by verifying the output and rejecting it if needed. Finally, we only interviewed active end-users of the app, who are likely to continue to use the app because they are satisfied with it. Past or non-users may provide different responses.

In hypothetical high-risk scenarios, participants carefully considered the AI's ability and various contextual factors

When we presented participants with hypothetical high-risk scenarios, we observed a different decision-making process around app adoption. Participants considered the app's **ability** with respect to various **situational characteristics**. For example, for the sick bird scenario, some participants judged the app is worth a try. P15 described using the app as "*something that wouldn't hurt*" since they are in a situation where both they and the vet could not identify the bird. They expressed some degree of confidence in the app's ability: "*I feel like Merlin's not gonna tell you that a baby hawk is a chickadee*" (P15). Similarly, P17 said they "*would definitely recommend it [the app] to get into the right ballpark.*" Still, they recommended consulting other scientific resources and doing a "*triple check*" of the app's output, since the **risk** of misidentification, i.e., the sick bird receiving the wrong treatment, is higher than the risk in their actual use settings, e.g., gaining wrong knowledge.

Other participants were skeptical that the app could identify the sick bird (P3, P6, P15, P19). P4 did not think the app could identify birds that they and the vet could not: "*Assuming that I don't know what the bird is and they [vet] don't know what the bird is, this bird is some ambiguous-looking bird. In those cases [...] I don't think Merlin would be able to know.*" P15 pointed out that sick birds are often "*fledglings, juveniles*" which are "*harder to ID for everybody in real life and presumably harder for Merlin.*" P6 noted that sick birds may be "*out-of-distribution*" for the app due to their underrepresentation in the training data: "*I assume Merlin is not trained on sick birds, so I can totally see it doing something crazy.*" These participants weighted the app's ability against the **task difficulty** and decided against adopting the app in the sick bird scenario.

Similarly, for the game show scenario, participants jointly considered the app's ability and the situation's characteristics. P1 and P4 said they would choose the app

Table 4.2: **Factors that influenced participants' trust in AI.** See Section 4.4.5 for a discussion.

Human-related factors	AI-related factors	Context-related factors
Domain knowledge	Ability	Task difficulty
Ability to assess the AI's outputs	Integrity	Perceived risks and benefits
Ability to assess the AI's ability	Benevolence	Situational characteristics
Ability to use the AI	Popularity	Domain's reputation
	Familiarity	Developers' reputation
	Ease of use	

if there are **time constraints**, but otherwise choose “*a good quality field guide*” (P1) or “*Discord*” (P4). Similarly, P15 said they would choose the app if they need to give an answer “*quickly, within 30 seconds*” but otherwise consult other birders. Others said they would choose the app if they have to do **sound-based identification** on the game show (P6, P10, P12, P20), describing difficulties with text-based referencing of sound: “*some books saying ‘it goes da-da-da’ is not helpful*” (P6). P20 explained their reasoning in detail: “*If I’m on this game show and it plays a sound, I would definitely want to use the app [...] but if it shows me a bird, I might just want to google it because I have enough knowledge personally that I could probably guess what type [...] and then search by colors. So I guess it comes down to what I think the app does the best, which is sound, versus what I think I can get away without it.*” Participants considered the app’s ability not only on its own, but also in comparison to other resources. Again, participants carefully made app adoption decisions as the **perceived risks and benefits** of the scenario, i.e., loss and gain of money, are higher than those of their actual use setting.

Some participants adopted the AI due to familiarity and ease of use

Two other factors that impacted participants’ app adoption decisions were **familiarity** and **ease of use**. For the sick bird scenario, P2 said they would definitely used the app because it “*feels kind of like second nature*.” P16 also chose to identify the

sick bird with the app because using the app “*would be the easiest.*” Similarly, for the game show scenario, P2 picked the app as their top choice because “*it’s so easy [...] it doesn’t take all that much time to look through everything.*” P16 mentioned both familiarity and ease of use: “*I think Merlin would make the most sense since I’m familiar with it.*” They described other resources as requiring more “work” by end-users, compared to the app where end-users can just input bird photos and/or audio recordings: “*You still have to do a lot of work to do like a Google search compared to this [app]*” (P16).

P4 described another aspect of familiarity: their **ability to use the app**. They said, “*I definitely would use Merlin because I’m familiar with it. And I trust my ability, like I know how to operate it pretty well*” (P4). We found this response particularly interesting because the ability to use the AI has not been explored much in the trust in AI literature. However, we expect it will become an important topic in trust and human-AI interaction research, as AI applications grow in complexity and require end-users to develop skills for effective use of AI.

Finally, some participants with little **domain knowledge** said they would adopt the app because other resources lack familiarity and ease of use (P11, P17). For instance, P17 described field guides as a more advanced and less accessible resource than the app “*because of the way they’re structured and organized.*” They said the app will get them closer to an answer “*a lot quicker*” (P17). Similarly, P11 mentioned they could not effectively use the Internet search engine because they do not know enough about birds to effectively describe the bird they have to identify. They added, “*Maybe someday, when I know a lot about birds, I would feel comfortable using another resource that’s more of an expert than Merlin is, but I think Merlin is the right level of expertise for what I know right now*” (P11).

4.4.5 Summary of the results

In short, we found that end-users' trust relationship with AI is complex. Overall, participants found the app trustworthy and trusted it. Still, they carefully assessed the correctness of individual outputs and decided against app adoption in certain high-stakes scenarios. This finding illustrates that trust is a multifaceted construct that must be approached holistically. To get a full and accurate picture of trust, it is crucial to examine both *general* aspects such as trustworthiness perceptions and trust attitudes and *instance-specific* aspects such as AI output acceptance and adoption decisions.

We also highlight that trust in AI is influenced by many factors. In Table 4.2, we organize the factors we identified based on whether they are related to the human trustor, the AI trustee, or the context, following prior work [146, 168, 178]. **Human-related factors** include domain knowledge and other factors influenced by it, such as the ability to assess the AI's outputs, the ability to assess the AI's ability, and the ability to use the AI. **AI-related factors** include internal factors such as ability, integrity, and benevolence; external factors such as popularity; and user-dependent factors such as familiarity and ease of use. **Context-related factors** include task difficulty, perceived risks and benefits of the situation, other situational characteristics, and the reputation of the domain and the developers. We emphasize that this is not a complete set of factors that can influence trust in AI, but what we *observed* in our case study in a bottom-up manner.

4.5 Discussion

In this section, we discuss the implications of our findings (Section 4.5.1), reflect on the applicability of Mayer et al.' trust model [261] (Section 4.5.2), examine the

limitations of our work and opportunities for future work (Section 4.5.3), and provide practical recommendations for future research on trust in AI (Section 4.5.4).

4.5.1 Key findings and their implications

Insights from instance-specific trust-related behaviors

Participants' instance-specific decisions about AI output acceptance and adoption were particularly useful for understanding what factors influence trust in AI and how. In Section 4.4.3, we described how participants trusted the app output more when the task is easy (e.g., “*relatively unambiguous*” bird) and less when the task is difficult (e.g., “*little brown bird*”). Similarly in Section 4.4.4, we described how some participants were hesitant to use the app to identify the sick bird in the first hypothetical scenario because they judged the task would be too difficult for the app. These examples illustrate the rich reasoning behind participants' trust-related behaviors, where factors of trust interact with each other. Participants used their domain knowledge (human-related factor) to assess task difficulty (context-related factor) and weighted it against the app's ability (AI-related factor) to decide whether or not to accept the app output or adopt the app in the given situation.

In earlier sections, we also described domain knowledge's influence on participants' ability to assess the app's ability (Section 4.4.2) and outputs (Section 4.4.3). Participants with domain knowledge assessed the correctness of an app output by, for example, cross-checking it with their own identification and judging its likelihood based on their knowledge of what birds are common and rare in the area and what birds are easy and difficult to identify. Participants without domain knowledge, however, had difficulties in assessing the correctness of app outputs and, consequently, the app's overall ability.

Taken together, these findings have two important implications. First, they imply that domain knowledge can be a widely influential factor of trust. In the above

examples, participants assessed task difficulty based on their domain knowledge of which birds are difficult and easy to identify. Further, participants' domain knowledge directly influenced their ability to verify the app's outputs and overall ability, which can have far-reaching influences on participants' trust and interaction with the app. Second, they imply that participants expected the app to struggle on similar tasks as humans do, when current research suggests AI models do visual recognition differently from the human brain [380] and make different mistakes than humans [201, 377, 419]. However, we cannot rule out the possibility that the app struggles on similar bird identification tasks as humans do, since we do not have access to its underlying AI models. This implication calls for further research on the relationship between end-users' domain knowledge, perceptions and expectations toward AI, and trust calibration ability.

The surprising impact of domain knowledge

As aforementioned, participants' domain knowledge had a wide impact, influencing all aspects of their trust in the app. We found this surprising because domain knowledge has not been discussed much as a factor of trust in the trust in AI literature. Upon reflection, we speculate this is because most AI applications are either (1) non-expert applications that do not have a notion of domain expertise (e.g., image search) or (2) expert applications that are developed to be used by domain experts only (e.g., clinical decision-support tool). Our study application is unique in that it is an expert application that is used by both domain experts and non-experts. Further, our study design, specifically our choice to recruit participants with varying background (Table 4.1), allowed us to observe group differences with respect to domain knowledge.

This observation has particularly important implications for expert AI applications that support high-stakes decisions about people. Using clinical decision-support tools as an example, we expect domain knowledge differences to lead to very differ-

ent trust-related assessments and decision-making between experts (clinicians) and non-experts (patients, regulators, and other stakeholders). As we saw in our case study, non-experts may not be able to spot the AI’s mistakes and assess its trustworthiness as accurately as domain experts. Even though non-experts are not the intended end-users of this AI (clinical decision-support tool), it is important that they have mechanisms to appropriately calibrate their trust in the AI. Hence, we urge the community to consider domain knowledge when designing AI applications and trust calibration interventions. For example, a system could assess the verification behaviors used by domain experts, and build these options into the system so that they are accessible to experts and non-experts alike.

The importance of contextual factors and contextually-grounded studies

Finally, we highlight the importance of contextual factors and contextually-grounded studies for understanding their influences on trust. When participants were describing the app’s trustworthiness, we observed that the positive reputation of the domain (birding community) and the developers (Cornell Lab of Ornithology) led them to positively assess the AI’s ability, integrity, and benevolence (Section 4.4.2). We have two points of discussion on this finding. First, it shows that external contextual factors (reputation of the domain and the developers) influence internal AI factors (ability, integrity, and benevolence), underlining the impact of contextual factors on trust. It also reiterates that factors influence each other, and calls for research that studies the interactions between factors. Second, while this specific finding is context-dependent, it provides generalizable insights. For example, we can anticipate end-users to have doubts about an AI application’s ability if the AI is not developed by a well-known institution; benevolence if the AI seems to have a different goal from them (e.g., recommendation systems trying to sell unneeded products); and integrity

if the AI seems to make decisions with wrong reasons (e.g., decision-making systems discriminating based on protected attributes).

4.5.2 Adapting existing trust models to AI

In this work, we used Mayer et al.’s model for organizational trust [261] to analyze the empirical data we collected on trust in AI. Overall, we found the model applicable and useful for understanding trust in AI. In particular, we found helpful the way in which it breaks down “trust” into multiple components and delineates trust from its antecedents, context, and products (Figure 4.1). However, as with any model, there were some limitations and challenges in its application. First, since the model was originally developed for *trust between people*, we had to make adaptations to apply it to *trust in AI*. For instance, when describing participants’ trustworthiness perceptions (Section 4.4.2), instead of using Mayer et al.’s [261] definitions of ability, integrity, and benevolence, we used Lee and See’s [220] automation-friendly translations of these factors: performance, process, and purpose. Second, Mayer et al.’s model [261] is by no means a comprehensive trust model. This is expected as the work’s goal was not to list all possible antecedents of trust. Hence, we drew from other works [146, 152, 168, 178] to categorize the factors we identified into human, AI, and context-related factors. Third, in our study, we did not observe the influence of trustor’s propensity, one of the model components. However, this result does not imply that trustor’s propensity is an unimportant factor of trust in AI. Future research, in particular survey and experimental studies, are needed for such conclusions.

4.5.3 Limitations and future work

Our work has the following limitations. First, as with any case study, our findings are context-specific. However, the gained insights, e.g., the trust-influencing factors we identified in Table 4.2, may generalize to other settings. Further, our holistic

approach to trust may aid future research on other types of AI applications. Another limitation is that all participants were active end-users of the app. Those who just started using it, stopped using it, or chose to not use it are not represented in the study. We encourage a more comparative study design for future work so that findings can be compared across non-users and user subgroups. Finally, due to the highly multifaceted and dynamic nature of trust in AI, there are aspects of it that our work does not cover. More research needs to be done, especially on how trust is initially developed and changes overtime, and how trust relationships with AI vary between stakeholder groups.

We highlight more important areas for future work. In our view, the overarching goal of trust in AI research is to establish *warranted and calibrated trust in AI*, where people's trust in AI systems matches their actual trustworthiness. We see three steps of research to achieving this goal. The first is to deepen the understanding of trust in AI, e.g., what aspects there are to trust and what factors influence it. Our work falls into this first step. The second is to understand how different factors influence trust, likely with quantitative methods such as surveys and behavioral experiments. We urge the field to move from studying one or a few factors in lab settings with hypothetical end-users, to studying multiple factors in real-world settings with actual end-users. This shift is necessary for understanding the interactions between factors, as well as the contextual influences on trust. The third step is to design effective trust calibration interventions, based on the gained insights on trust and its influencing factors. We point to Wischnewski et al.'s work [413] for a survey of the state-of-the-art trust calibration interventions and suggestions for future directions.

4.5.4 Practical recommendations

We conclude with a set of practical recommendations for future research on trust in AI.

1. *State a definition of trust.* Trust is a multifaceted construct that carries different meanings to different people. Explicitly stating a definition of trust can help remove confusion around the term and encourage accurate interpretation and comparison of study results.
2. *Examine if trust is the construct being studied.* Oftentimes what's being studied is not trust, but other related constructs such as confidence and reliance. We recommend that researchers carefully examine their study design and context to ensure trust is the construct being analyzed. We hope our Section 4.4.1 serves as a helpful example of such an examination.
3. *Approach trust holistically and study its antecedents, context, and products.* Our contextually-grounded study of general trustworthiness perceptions and trust attitudes, as well as instance-specific trust-related behaviors, revealed a comprehensive picture of end-users' trust relationships with AI that cannot be gained by studying only one aspect of trust. Hence, we recommend studying trust together with its antecedents, context, and products, to the extent possible.
4. *Consider human, AI, and context-related factors and their interactions.* As observed in this work, trust is influenced by many factors. To prevent surprises and gain a thorough understanding of trust in a given context, we recommend anticipating as many factors as possible and studying their interactions. We found it particularly helpful to consider factors along the dimensions of human, AI, and context.

4.6 Conclusion

We conducted a qualitative descriptive empirical study of end-users' trust in AI in a real-world context. We interviewed 20 end-users of a widely-used AI-based app for

bird identification [368] and inquired about their trust in the app from many angles. Using a process-oriented trust model [261], we elaborated on multiple aspects of trust in AI. Notably, we found a discrepancy between participants' *general* trustworthiness perceptions and trust attitudes, and *instance-specific* trust-related behaviors, adding nuances to existing understandings of trust in AI. We also identified human, AI, and context-related factors of trust, finding that domain knowledge had a particularly big influence on participants' trust and interaction with AI. Finally, we discussed the implications of our findings and provided recommendations for future research toward establishing warranted and calibrated trust in AI. We hope our work aids future research on other AI applications and the various contexts into which they are integrated.

Acknowledgments

We foremost thank our participants for generously sharing their time and experiences. We also thank Tristen Godfrey, Dyanne Ahn, and Klea Tryfoni for their help in the interview transcription. Finally, we thank the anonymous reviewers and members of the Princeton HCI Lab and the Princeton Visual AI Lab (especially Angelina Wang, Vikram V. Ramaswamy, Amna Liaqat, and Fannie Liu) for their helpful and thoughtful feedback. This material is based upon work partially supported by the National Science Foundation (NSF) under Grants No. 1763642 and 2145198 awarded to OR. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. We also acknowledge support from the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award (OR), Princeton SEAS Project X Fund (RF, OR), Princeton Center for Information Technology Policy (EW), Open Philanthropy (RF, OR), and NSF Graduate Research Fellowship (SK).

Chapter 5

Examining the impact of large language models' uncertainty expression on user reliance and trust

Widely deployed large language models (LLMs) can produce convincing yet incorrect outputs, potentially misleading users who may rely on them as if they were correct. To reduce such overreliance, there have been calls for LLMs to communicate their uncertainty to end users. However, there has been little empirical work examining how users perceive and act upon LLMs' expressions of uncertainty. In this chapter, we explore this question through a large-scale, pre-registered, human-subject experiment (N=404) in which participants answer medical questions with or without access to responses from a fictional LLM-infused search engine. Using both behavioral and self-reported measures, we examine how different natural language expressions of uncertainty impact participants' reliance, trust, and overall task performance. We find that first-person expressions (e.g., "*I'm not sure, but...*") decrease

participants’ confidence in the system and tendency to agree with the system’s answers, while increasing participants’ accuracy. An exploratory analysis suggests that this increase can be attributed to reduced (but not fully eliminated) overreliance on incorrect answers. While we observe similar effects for uncertainty expressed from a general perspective (e.g., “*It’s not clear, but...*”), these effects are weaker and not statistically significant. Our findings suggest that using natural language expressions of uncertainty may be an effective approach for reducing overreliance on LLMs, but that the precise language used matters. This highlights the importance of user testing before deploying LLMs at scale.

This chapter contains materials published in the FAccT 2024 paper, “I’m Not Sure, But...”: Examining the Impact of Large Language Models’ Uncertainty Expression on User Reliance and Trust [193], in collaboration with Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan.

5.1 Introduction

Large language models (LLMs) are transforming our daily lives. Today millions of people already incorporate LLMs into everyday tasks like searching for information [238, 265], writing [132, 427], and programming [10, 281, 315]. However, the use of LLMs raises significant risks [35, 44, 405]. Notably, like all models, LLMs are imperfect. They are widely recognized to produce outputs that are fluent and plausible, yet ultimately wrong [171, 173, 232]. This can lead to disastrous outcomes through *overreliance* [58, 75, 304, 384], when people take actions based on incorrect outputs. This concern garnered much public attention in 2023 when a lawyer included fake judicial opinions generated by ChatGPT in a legal brief presented in court [407]. Such risks have been at the forefront of regulators’ minds when drafting new frameworks for governing AI including the Draft AI Act in the European Union [303] and the

NIST AI Risk Management Framework in the United States [362]. In fact, Article 14 of the Draft EU AI Act explicitly requires developing and evaluating approaches to prevent overreliance on AI systems. However, overreliance is notoriously difficult to mitigate, as many mitigations, such as explanations, are found to be ineffective or even can backfire to increase overreliance [26, 304, 312, 400, 437].

As one approach to reduce overreliance, the research community has called for LLMs and LLM-infused applications to express the uncertainty of outputs to end users [22, 226, 231, 268, 385, 443]. The idea of conveying AI uncertainty is not new; in AI-assisted decision-making settings, communicating (un)certainty has been shown to support trust calibration [437], increase vigilance [313], and improve task performance [26]. But because of their open-ended outputs, wide-ranging use cases and user bases, and shifting public perception, LLMs raise new questions around how to both estimate and express uncertainty [226, 385]. For estimation, “default” approaches are often found to be overconfident [76, 268, 417], and a new line of work has emerged on improving their calibration [13, 76, 86, 204, 231, 233, 268, 370, 443]. For expression, LLMs open up a new design space; instead of presenting uncertainty numerically or visually, LLMs can present natural language expressions of uncertainty — for instance, hedging phrases like “*I’m not sure, but...*” — embedded in their outputs. Still, there is little understanding about how to effectively express uncertainty in natural language to end users.

To deploy LLMs responsibly, it is necessary to understand how users react to uncertainty expression before implementing approaches at scale since it may have unintended negative consequences — potentially even increasing overreliance if it causes the system to appear more trustworthy than it is. Best practices for uncertainty expression will play a critical role in ensuring that requirements like those in the Draft EU AI Act serve their intended purpose. To that end, we study how people perceive and act upon an LLM’s expression of uncertainty when seeking

medical information using a fictional LLM-infused search engine. We choose to study this setting because search (unlike, for example, creative writing) is an application in which the correctness of responses is fundamental — especially for potentially high-stakes medical queries — making overreliance a serious concern. Additionally, LLM-infused search engines are already used by millions of people.¹

We choose to focus on natural language expressions for several reasons. First, LLM-infused search engines already include hedging language [226, 238, 268]. Second, social science research shows that, in human communication, expressing (un)certainty through natural language is often preferred and perceived as more intuitive than numerical expressions [95, 236, 392, 412, 448]. Third, this allows uncertainty to be expressed seamlessly within the natural language interactions of LLM-infused applications, rather than on the side or in onboarding materials [67, 274, 295] that users might overlook.

Taking inspiration from the uncertainty communication literature — both in the context of AI systems and person-to-person [292, 434] — we also explore the impact of the perspective used to express the uncertainty, comparing expressions in the first person (e.g., “*I’m not sure, but...*”) with expressions from a general perspective (e.g., “*It’s not clear, but...*”).

Concretely, we conduct a large-scale, pre-registered, human-subject experiment ($N=404$) in which participants answer medical questions with or without access to responses from a fictional LLM-infused search engine, referred to as “AI System A.” We randomly vary whether participants have access to the system’s responses as well as the presence (present/not present) and perspective (first-person/general) of uncertainty expressed in these responses. We measure the impact of these experimental conditions on factors including participants’ accuracy, the amount of time they take,

¹In March 2023, Microsoft reported Copilot in Bing served 45 million chats in the first month of its public preview [265]. Perplexity AI reported its service had reached 2 million monthly active visitors in four months [9, 359].

Task 3 / 8

Question: Can an adult who has not had chickenpox get shingles?

AI system A's answer:

No, an adult who has not had chickenpox cannot get shingles without having had chickenpox before. Both chickenpox and shingles are caused by the varicella-zoster virus [1]. When a person is exposed to the virus, they develop chickenpox and may develop shingles later in life [1].

[1] microsoftstart.msn.com. https://microsoftstart.msn.com/en-us/health/ask-professionals/expert-answers-on-shingles/hp-shingles?questionid=u9hgatt3&type=condition&source=bingmainline_conditionqa

How confident are you in AI system A's answer?

Not confident at all Not very confident Neither Fairly confident Very confident

What is your final answer to the question: Can an adult who has not had chickenpox get shingles?

Yes No

How confident are you in your final answer?

Not confident at all Not very confident Neither Fairly confident Very confident

What is your final answer based on? (Select all that apply)

AI system A's answer
 Your own reading of the linked sources in AI system A's answer
 Your own knowledge
 Your own internet search
 Other (please specify)

Figure 5.1: **Screenshot of the information-seeking task** (shown in the CONTROL, UNCERTAIN1ST, and UNCERTAINGENERAL conditions). Additional screenshots, including the variant shown in the NO-AI condition, can be found in the supp. mat.

their reliance on the system’s responses versus other sources of information, and their self-reported trust in the system.

We find that participants who are shown first-person expressions of uncertainty are less confident in the system’s answers, agree with the system’s answers less often, and submit more correct answers compared with participants who see no expression of uncertainty. An exploratory analysis suggests that the increased accuracy can be attributed to reduced (but not fully eliminated) overreliance on the system’s incorrect answers. While we observe similar effects for uncertainty expressed from a general perspective, these effects are weaker and not statistically significant. These results suggest that expressing uncertainty through natural language can be an effective way to reduce overreliance and (over)trust in LLM-infused search engines. Still, we advocate for teams building and deploying LLMs to evaluate approaches to mitigate overreliance, including language choices, carefully with end users before release and for policymakers to embrace diverse and flexible approaches.

5.2 Related work

5.2.1 Uncertainty expression

Uncertainty expression has been studied extensively both in the context of AI and in the context of human communication. Estimates of uncertainty can be expressed in different ways, including numerically (e.g., “*a probability of 0.2*”), visually (e.g., displaying error bars), and through natural language (e.g., “*with high uncertainty...*”). While numerical expressions and visualizations allow for high precision, they are notoriously difficult for people to understand and are often misinterpreted, even by experts [33, 155, 176, 360]. In contrast, while less precise, natural language expressions of uncertainty are often perceived to be more intuitive and favored by people [95, 236, 392, 412, 448]. Because of this, and since LLMs already produce natural language outputs, we focus on natural language expressions in our study.

Different forms of natural language uncertainty expression have been studied by researchers in disciplines ranging from psychology [84, 390, 409] and human-computer interaction [18, 121, 374] to communication [247] and marketing [104, 292]. Our study design builds on this literature. Most notably, in the context of marketing, Oba and Berger [292] found that different types of hedges — a form of uncertainty expression — have different levels of persuasion, with the most persuasive being those that suggest a high likelihood of occurrence (e.g., “*probably*” as opposed to “*possibly*”) and those that take a personal, first-person perspective (e.g., “*I feel like...*”) as opposed to a general perspective (e.g., “*It feels like...*”). Indeed, taking a first-person perspective is often found to increase the engagement in the persuasion literature [73]. This motivated our exploration of the effect of AI uncertainty expressions taking a first-person or general perspective, though we explore the *negative* persuasive effect (i.e., deterring overreliance) of expressing uncertainty.

Many methods have been proposed for estimating and communicating uncertainty in AI systems [39]. Most relevant, several studies have shown that communicating uncertainty can reduce overreliance. In the context of house valuation, Prabhudesai et al. [313] found that visualizing a system’s uncertainty using quantile dot plots forced participants to slow down and think analytically. In the context of sentiment classification and question answering, Bansal et al. [26] found that displaying a system’s numerical confidence improved participants’ task performance.

A few studies have looked specifically at natural language expressions of uncertainty in AI systems. They have been studied in the chatbot literature as a design strategy to prevent user frustration from conversational breakdowns [18, 121]. Radensky et al. [317] studied communicating the confidence of a conversational music recommendation system through language and color coding. Experimenting with different levels of calibration, they found that accurate confidence signals led to the greatest increase in trust without encouraging overreliance, though underreliance may still occur. Zhang et al. [434] explored different ways of having an AI system (named “ShapeBot”) express confidence in the reasoning behind its recommendation, examining the effect of point of view (first-person “*I think...*” vs. third-person “*ShapeBot thinks...*”) and strength of the belief expressed (“*ShapeBot thinks...*” vs. “*ShapeBot knows...*”). They found that both factors affected user reliance, highlighting the importance of carefully considering the language used to express (un)certainty.

5.2.2 Uncertainty in LLMs

Obtaining accurate numerical estimates of uncertainty for LLMs is an active line of research. One way to estimate an LLM’s uncertainty is by the likelihood of generating a specific output given the context. However, this “generation probability” may not reflect what end users expect or want when they think of uncertainty [385]. A more useful notion of uncertainty might be one that captures how likely it is that

the LLM’s output is factually correct or correctly meets the user’s needs. This notion of uncertainty may apply to a full output or to sentences, phrases, or words within the output. Many researchers are working on evaluating how calibrated existing uncertainty estimates are and proposing new techniques to improve calibration [13, 76, 86, 204, 231, 233, 268, 370, 443]. Current findings suggest that LLMs are often overconfident [76, 268, 417], which may give a false impression of their capabilities and exacerbate overreliance [96].

There is also a growing interest in LLMs’ ability to directly generate natural language expressions of (un)certainty [231, 268, 417, 443]. Notably, Mielke et al. [268] observed that LLMs regularly express confidence (e.g., “*Obviously...*”) and doubt (e.g., “*I’m not sure, but...*”) through the language used in their outputs, but these expressions are poorly calibrated. Zhou et al. [443] “taught” OpenAI’s GPT-3 [55] model to express (un)certainty through prompt engineering, but also found that the generated expressions were not well calibrated, especially those suggesting high certainty.

Despite this active research, there has been little empirical work examining the impact of uncertainty expression on users of LLM-infused systems. Notable exceptions are the works of Vasconcelos et al. [385] and Spatharioti et al. [357], who explored the effect of highlighting uncertain parts of LLM outputs in the context of code completion and search, respectively, and the concurrent work of Zhou et al. [444], who explored the effect of LLMs’ natural language expressions of (un)certainty in the context of trivia question answering. These studies’ results support uncertainty expression as a promising technique to encourage appropriate reliance, particularly when uncertainty estimates are well calibrated. Our work adds empirical knowledge on this topic through a large-scale, pre-registered experiment studying natural language uncertainty expressions in the context of LLM-infused search.

To avoid making assumptions of calibration or tying our experiment to a particular uncertainty estimation approach, we design our study to include both instances

in which the system expresses uncertainty when it is incorrect and instances in which it expresses uncertainty when it is correct. By randomly varying whether or not uncertainty is expressed on any particular response, we are able to directly compare participants’ behavior when uncertainty is and is not expressed.

5.2.3 Measuring reliance and trust

We hypothesize that whether or not an AI system expresses uncertainty — and if it does, the perspective in which the uncertainty is expressed — impacts user reliance and trust. We note that there are many definitions, measures, and factors of reliance and trust [78, 191, 379, 386, 436]. We use a combination of dependent variables and a mix of behavioral and self-reported measures to capture aspects most relevant to our research setting of LLM-infused search.

In the AI-assisted decision-making literature, many experiments follow the judge-advisor paradigm [45], presenting a participant with an AI-generated answer to a question and then asking the participant to provide their own answer, a set-up we adopt in our work. In this set-up, *agreement* between a participant’s answer and that of the AI system is a commonly studied behavioral measure of reliance and trust [58, 69, 212, 237, 250, 277, 424, 437]. While this does not capture reliance or trust directly — the participant may have come up with the same answer on their own even without the AI system — comparing how often participants agree with the AI system’s answers across experimental conditions gives a way of measuring whether they rely on the system differently across conditions. We use this measure in our work. We note that in some prior work, the participant is asked to provide an initial answer first before seeing that of the AI system [239, 240, 312, 335]. In these cases, other metrics like weight of advice can be used to more directly capture reliance. We do not adopt this set-up because we use questions we do not expect participants to be able to answer on their own.

To complement *agreement*, we examine participants' *confidence*, *source usage*, *trust intentions*, and *trust beliefs*. First, as in prior work [69, 81, 190, 250, 312], we ask participants to report their *confidence* both in the answer output by the AI system and in their own answer. Second, as an indirect measure of reliance and trust, we capture participants' *source usage* by both tracking whether or not participants click on the linked sources in the system's responses and asking them to self-report the resources that they based their final answer on. Finally, using responses collected in an exit questionnaire, we measure participants' *trust intentions* and *trust beliefs* using the scales developed by McKnight et al. [145]. Trust intentions refer to a participant's desire to use the system, while trust beliefs refer to their perceptions about the system's trustworthiness such as the system's perceived ability, benevolence, and integrity [261]. In general, the two are positively related [145], but Radensky et al. [317] found that they can be differently impacted by a system's expressed confidence. We measure both in our experiment to better understand the impact of the system's uncertainty expression.

We also measure two system facets that are known to impact trust: *perceived anthropomorphism* and *perceived transparency*. Participants may view the expression of uncertainty (especially first-person) as an inherently human behavior, leading to increased anthropomorphism. Recent work has expressed concern around anthropomorphism leading to over-trust [3, 344], a potential path for uncertainty expression to backfire. Uncertainty expression can also increase the system's perceived transparency [39], which is generally shown to enhance trust, whether or not appropriate [226].

In addition to reliance and trust, we also consider task performance as a dependent variable, measured as *correctness* of participants' answers and *time on task*. Both of these have been studied in prior work on AI-assisted decision making [215], as well as specifically for studying the effect of uncertainty expression [384].

We formally define the dependent variables in Section 5.3.2 and articulate hypotheses in Section 5.3.3.

5.3 Methods

As described in Section 5.2.3, our experiment is designed to measure the impact of natural language expressions of an LLM’s uncertainty on user reliance and trust. We do this in the context of information seeking in the medical domain. We pre-registered our experimental design, hypotheses, analysis plan, and data collection procedures before collecting data.² To complement our pre-registered analyses, we include exploratory analyses and a qualitative analysis of participants’ free-form responses. The study was approved by our internal Institutional Review Board (IRB).

5.3.1 Procedure and experimental conditions

We designed a between-subjects experiment with some within-subjects comparisons, which we conducted on Qualtrics. Participants complete a set of information-seeking tasks. Each task involves determining the correct yes-or-no answer to a challenging, factual question in the medical domain with or without access to responses from a fictional LLM-infused search engine, “AI system A.” The presence and form of system responses provided to participants depend on their experimental condition. Specifically, participants are randomly placed into one of four experimental conditions:

- CONTROL: Participants see AI responses without any expression of uncertainty.
- UNCERTAIN1ST: Participants see AI responses and half of the time these responses include uncertainty expressed in the first person, with personal pronouns (e.g., “*I’m not sure, but it seems...*”).

²Our pre-registration is viewable at <https://osf.io/mnrf9>.

- **UNCERTAINGENERAL:** Participants see AI responses and half of the time these responses include uncertainty expressed in a general perspective, without personal pronouns (e.g., “*There is uncertainty, but it seems...*”).
- **No-AI:** Participants are not told about the AI system and do not see AI responses.

CONTROL is a baseline to which we compare the conditions **UNCERTAIN1ST** and **UNCERTAINGENERAL** to understand the impact of uncertainty expressions. **No-AI** is a second baseline to understand the impact of access to the AI system.

The experiment is divided into three components. In the first, participants are introduced to the study and to AI system A (if applicable). They are given several task comprehension questions and are asked to complete an example task.

In the second component, participants answer a total of eight questions (details in Section 5.3.4). They are told they can use any resources they want as in natural settings. For each question, participants, except for those in the **No-AI** condition, are provided with responses from AI system A (Figure 5.1). The system’s yes-or-no answers within their responses are correct for only half the questions. In the **UNCERTAIN1ST** and **UNCERTAINGENERAL** conditions, the AI system expresses uncertainty in its answers for half the questions — we refer to them as *uncertain* answers versus *not uncertain* answers. We chose this breakdown to have sufficient data for each of the four possible scenarios of correct/incorrect answers with/without uncertainty expression. We randomize the order in which questions are presented, as well as the set of questions for which the AI system expresses uncertainty. However, since we based the AI system’s answers on real responses from Copilot in Bing (see Section 5.3.4), the set of answers and their correctness are fixed.

In the final component, participants fill out an exit questionnaire about their experience with and perception of the AI system (if applicable), their background on LLMs, and basic demographic information; see the supp. mat. Lastly, participants

are debriefed and reminded that some of the AI responses they saw may have contained inaccurate information.

5.3.2 Dependent variables

We now formally define the dependent variables (DVs) that we measured, motivated in Section 5.2.3. First, for each of the eight questions, we measured the following DVs based on participants' observed behavior:

- **Agree:** TRUE if the participant's final answer is the same as the AI system's answer; FALSE otherwise.
- **Correct:** TRUE if the participant's final answer is correct; FALSE otherwise.
- **Time:** Number of minutes from when the participant saw the task to when they clicked next.
- **LinkClick:** TRUE if the participant clicks on one or more links in the system's answer; FALSE otherwise.

We additionally measured the following DVs based on participants' self-reported ratings:

- **UseAI:** TRUE if the participant selected "AI system A's answer" in the question "What is your final answer based on? (Select all that apply)"; FALSE otherwise.
- **UseLink:** TRUE if they selected "Your own reading of the linked sources in AI system A's answer" in the above question; FALSE otherwise.
- **UseInternet:** TRUE if they selected "Your own Internet search" in the above question.; FALSE otherwise.
- **ConfidenceAI:** Rating on the question "How confident are you in AI system A's answer?" on a 5-point scale.

- **ConfidenceAnswer:** Rating on the question “How confident are you in your final answer?” on a 5-point scale.

Finally, based on responses to the exit questionnaire, we calculated the following indexes (all on a 5-point scale):

- **TrustBelief:** Average rating on six statements adapted from the trust scale by McKnight et al. [145].
- **TrustIntention:** Average rating on four statements adapted from the trust scale by McKnight et al. [145].
- **Anthropomorphism:** Average rating on four items from the Godspeed Questionnaire Series [28].
- **Transparency:** Average rating on two statements: “I feel I had a good understanding of what AI system A’s answers were based on” and “I feel I had a good understanding of when AI system A’s answers might be wrong.”

Full details are in the supp. mat. Note that some DVs were not applicable for the No-AI condition, where we measured only **Agree**, **Correct**, **Time**, **UseInternet**, and **ConfidenceAnswer**. Here we made one (and only one) minor deviation from our pre-registration by including **Agree**. Although participants in this condition do not see the AI system’s answers, this gives us a baseline for how often participants would arrive at the same answer on their own.

5.3.3 Hypothesis and analysis

We expected the presence and perspective of uncertainty expression to impact participants’ reliance, trust, and performance. Formally, for each DV, we hypothesized that condition affects DV. For each repeatedly measured DV, we additionally hypothesized that whether or not uncertainty was expressed in a particular AI response

affects DV. We tested our hypotheses with the following pre-registered, confirmatory analyses, for which we present results in Section 5.4.

We first test the main effect of the conditions with a **between-condition analysis**. For repeatedly measured DVs, we fit the model $DV \sim Condition + (1|participant) + (1|question)$ with CONTROL as the reference level for Condition. Then to compare the effects of the two conditions with uncertainty, we used a Wald test to test the equality of the corresponding coefficients. For DVs measured once in the exit questionnaire, we use analysis of variance (ANOVA) to compare means across the conditions. If significant, we conduct pairwise comparisons with a post-hoc Tukey test.

Next, we test the effect of uncertainty being expressed or not in a particular response with a **within-condition analysis** for conditions with uncertainty expression. For repeatedly measured DVs, we fit the model $DV \sim AIUncertain + (1|participant) + (1|question)$, where AIUncertain is TRUE if the AI response is uncertain and FALSE otherwise. We fit this model once for data from the UNCERTAIN1ST condition and once for data from UNCERTAINGENERAL.

We complement the confirmatory analyses with two additional analyses. First, we conduct an exploratory analysis of the effect of AI’s uncertainty expression on over- and underreliance by separately analyzing cases where the AI system gave correct versus incorrect answers. Analysis details and results are presented in Section 5.5.1. Second, we conduct a thematic analysis [48, 51] of free-form responses from participants in the UNCERTAIN1ST and UNCERTAINGENERAL conditions describing their experience with and perception of the AI system. The first author drafted the codebook and conducted the initial coding, then discussed the results with all authors and refined the coding together. We describe how the AI system’s uncertainty expression affected participants in Section 5.4 along with the quantitative results, and describe how participants interpreted the system’s uncertainty expression in Section 5.5.2.

5.3.4 Questions and AI responses used

We selected a set of factual questions for participants to answer according to the following criteria: (1) most lay people should not know the answer; (2) the question and answer should not directly show up when using popular search engines; and (3) the answer can be objectively and automatically assessed. To satisfy the criteria, we constructed a set of yes/no medical questions. We began with questions from the MedQuAD dataset [34] and made minor modifications to some to increase the difficulty of finding an answer. We verified that each question does not show up as it is and can not be immediately answered using popular search engines, and consulted multiple sources to confirm the correct answer.

To create AI responses that are realistic and reflect the state-of-the-art in LLM-infused search, we input the selected questions into Microsoft’s Copilot in Bing. All responses were obtained in July 2023. To keep the fluency, style, and content of responses as realistic as possible, we made only minor modifications such as presenting in-line citations using square brackets instead of superscripts (see Figure 5.1) and starting each response with “Yes” or “No” for consistency (most responses from Copilot in Bing did this already). We did not make substantive changes to the content.

To create the uncertain versions of the responses, we drew on Oba and Berger [292]. For UNCERTAIN1ST, we replaced the leading “Yes” or “No” with an expression of the form “*I’m not {certain, sure}, but {it seems to me, it seems like, I would guess, I’d guess that}*,” selecting one phrase from each set of brackets. Similarly, for UNCERTAINGENERAL, we used an expression of the form “{*It’s unclear, It’s not clear, There is uncertainty*}, but *it seems like*.” Current LLMs already output expressions such as “*I’m not sure, but*” and “*I’m not sure, but my guess is*” [268], so we believe these insertions preserve the realisticness of responses. The only difference between the three versions of the AI system’s responses is the presence and perspective of uncertainty expression; their information content is otherwise identical.

Table 5.1: **Between-condition analysis.** We compare DVs across conditions. We report the model-estimated means (and standard errors) from our confirmatory analysis. The rightmost column shows pairs of conditions with statistically significant differences with significance marked as * ($p < 0.05$) or ** ($p < 0.01$).

[†]Note that we did not compare all possible pairs of conditions (see Section 5.3.3). For repeatedly measured DVs, we compared (CONTROL vs. UNCERTAIN1ST/UNCERTAINGENERAL/No-AI) and (UNCERTAIN1ST vs. UNCERTAINGENERAL) but not (No-AI vs. UNCERTAIN1ST/UNCERTAINGENERAL) to reduce the number of hypothesis testing. For DVs measured once in the exit questionnaire, we first compared the means of (CONTROL, UNCERTAIN1ST, UNCERTAINGENERAL) using ANOVA, then if significant, conducted pairwise comparisons.

Sec.	DV	CONTROL	UNCERTAIN1ST	UNCERTAINGENERAL	No-AI	Significant differences [†]
5.4.1	Agree (%)	80.9% (5.5)	74.8% (6.7)	77.6% (6.2)	58.4% (8.5)	No-AI <** CONTROL UNCERTAIN1ST <* CONTROL
5.4.2	ConfidenceAI (1-5)	3.95 (0.17)	3.66 (0.17)	3.80 (0.17)		UNCERTAIN1ST <** CONTROL
	ConfidenceAnswer (1-5)	4.30 (0.08)	4.34 (0.08)	4.27 (0.08)	4.22 (0.08)	
	LinkClick (%)	2.7% (2.1)	7.2% (4.8)	3.9% (3.1)		
	UseAI (%)	77.3% (5.1)	64.8% (6.9)	72.3% (6.0)		
5.4.3	UseLink (%)	74.7% (5.5)	85.5% (4.0)	81.7% (4.8)		
	UseInternet (%)	19.1% (5.6)	27.0% (7.4)	23.2% (6.7)	92.9% (2.5)	CONTROL <** No-AI
	TrustBelief (1-5)	3.90 (0.06)	3.86 (0.07)	4.00 (0.07)		
5.4.4	TrustIntention (1-5)	3.25 (0.10)	2.91 (0.10)	3.36 (0.10)		UNCERTAIN1ST <* CONTROL <* UNCERTAINGENERAL
	Anthropomorphism (1-5)	3.07 (0.10)	3.00 (0.11)	3.13 (0.11)		
	Transparency (1-5)	4.04 (0.06)	3.93 (0.07)	4.01 (0.07)		
5.4.5	Correct (%)	63.9% (8.6)	72.8% (7.4)	67.9% (8.1)	74.2% (7.1)	CONTROL <** UNCERTAIN1ST CONTROL <** No-AI
	Time (min)	2.13 (0.22)	2.10 (0.23)	2.03 (0.22)	1.57 (0.21)	No-AI <* CONTROL

We selected the final eight questions such that: (1) four questions have a correct answer of “Yes” and four have a correct answer of “No” so that always selecting “Yes” is no better or worse than random guessing; (2) the AI answers are correct for four and incorrect for four, so that always agreeing with the AI system is no better or worse than random guessing; (3) questions are not too easy to answer without access to the AI system (determined via piloting). The final set of questions used is provided in the supp. mat. along with the original responses from Copilot in Bing and our modified responses.

5.3.5 Data collection and participants

We conducted our experiment on Amazon Mechanical Turk (MTurk), a crowdsourcing platform widely used for human-subject experiments. Research has shown data from MTurk workers is comparable to data from other pools (e.g., commercial panels, social media, colleges) [37, 61, 71, 89, 260, 301], but recently there has been a decrease in data quality [80, 185, 249, 259, 387, 403]. Indeed, in pilot studies we found that a strong requirement on qualification was necessary to obtain meaningful data, with the highest quality data obtained when requiring a “Masters” qualification (granted by Amazon based on past performance). Since the pool of available participants with a Masters qualification is limited, we pre-registered a recruitment plan in which we would initially aim to recruit 432 (determined via a power analysis) U.S.-based MTurk workers with a Masters qualification, 99% or higher approval rating, and at least 2000 completed human intelligence tasks (HITs), removing the Masters requirement after 7 days if we were unable to meet our target sample size. In parallel, we implemented best practices to mitigate the use of bots and improve data quality, including CAPTCHAs, honeypot questions, speed checks, attention checks, and open-ended questions, which we used to define data exclusions.

We collected 656 complete responses over the course of two weeks in September 2023, of which we excluded 252 (38.4%) based on five pre-registered exclusion criteria. Our final sample consists of 404 responses: 104 in CONTROL, 92 in UNCERTAIN1ST, 94 in UNCERTAINGENERAL, and 114 in No-AI. See the supp. mat. for more on our data collection procedures.

Participants were paid \$5 USD. The payment was determined based on the expected experiment duration of 20 minutes (estimated from pilot studies) and target hourly wage of \$15. The actual median experiment duration was 20.5 minutes, so on average, participants were paid \$14.80 per hour. See the supp. mat. for more information about participants.

5.4 Results: Confirmatory analysis

We now present the results of our pre-registered, confirmatory analyses. Tables 5.1 and 5.2 contain the between-condition and within-condition analysis results, respectively. We refer the reader to Section 5.3.2 for definitions of all DVs. We present the estimated means (and standard errors) from the fitted models, calculated without conditioning on the random effects, and represent binary variables in percentages. We use *significance* to refer to statistical significance at the level of $p < 0.05$.

5.4.1 Agreement with AI

We begin with the results of agreement, a commonly used behavioral measure of reliance and trust. Our first finding is that **people tend to agree with the AI system when its responses are provided**. Our between-condition analysis suggests that participants with access to the AI system are significantly more likely to submit the same answer as the system than those who do not have access (80.9% CONTROL vs. 58.4% No-AI). We next find that **AI's uncertainty expression decreases agreement with the AI system**. Compared to CONTROL (80.9%), Agree is significantly lower in UNCERTAIN1ST (74.8%). It is also lower in UNCERTAINGENERAL (77.6%), although the difference is not significant. Our within-condition analysis suggests AI's uncertainty expression decreases agreement at the instance level as well. The estimated means of Agree for not uncertain vs. uncertain AI responses are 84.7% vs. 70.9% for UNCERTAINGENERAL (significantly different) and 79.5% vs. 73.4% for UNCERTAIN1ST (not significantly different). We find support for this finding in the qualitative data as well. 10 participants (out of 186 in UNCERTAIN1ST and UNCERTAINGENERAL) stated that when they disagreed with the system it was, as one put it, “*because of the uncertainty of the answers provided by AI*”

Other reasons participants mentioned for disagreement included the system’s answer being different from their own knowledge (e.g., “*A couple of the AI’s answers didn’t make sense so from my own common sense I had to make my own judgement*”) or the information in other resources (e.g., “*The answer seemed to contradict the links given, or I could not find how they came to that answer with the information from the links given*”) and having lower trust in the system.

5.4.2 Confidence in answers

Looking at participants’ self-reported confidence in answers, we find that **AI’s uncertainty expression decreases people’s confidence in its answer**. Compared to CONTROL (3.95 on a 5-point scale), ConfidenceAI is significantly lower in UNCERTAIN1ST (3.66). It is also lower in UNCERTAINGENERAL (3.80), although the difference is not significant. Our within-condition analysis provides further evidence for this finding and suggests that **AI’s uncertainty expression decreases people’s confidence in their final answer at the instance level**. For both UNCERTAIN1ST and UNCERTAINGENERAL, ConfidenceAI and ConfidenceAnswer are significantly lower on instances with uncertain (vs. not uncertain) AI responses. Indeed, one participant in the UNCERTAIN1ST condition stated, “*If the AI didn’t seem confident, I would like [sic] on the links. If the AI seemed confident I assumed he was correct most of the time.*”

5.4.3 Source usage

Looking at source usage, we find that **people with access to the AI system conduct their own Internet search less frequently than those without access**. UseInternet is notably lower in CONTROL (19.1%) than No-AI (92.9%). However, we see no significant differences in source usage between the three conditions in which AI responses are present, meaning **there is no evidence that the presence and**

Table 5.2: **Within-condition analysis.** For each of the two conditions with uncertainty, we compare DVs measured on instances with AI answers that are not uncertain vs. uncertain. We report the model-estimated means (and standard errors) from our confirmatory analysis. $>$ and $<$ note statistically significant differences with significance marked as * ($p < 0.05$) or ** ($p < 0.01$).

[†]The model does not fit properly due to large individual variance which we discuss in an exploratory analysis in Appendix C.1.

Sec.	DV	UNCERTAIN1ST		UNCERTAINGENERAL	
		Not Uncertain	Uncertain	Not Uncertain	Uncertain
5.4.1	Agree (%)	79.5% (7.2)	73.4% (8.6)	84.7% (5.0)	>** 70.9% (7.8)
5.4.2	ConfidenceAI (1-5)	3.88 (0.20)	>** 3.44 (0.20)	3.95 (0.16)	>** 3.65 (0.16)
	ConfidenceAnswer (1-5)	4.42 (0.09)	>** 4.26 (0.09)	4.36 (0.07)	>** 4.18 (0.07)
	LinkClick (%)	14.3% (6.9)	9.9% (5.1)	<i>See table caption[†]</i>	
5.4.3	UseAI (%)	73.2% (7.2)	>** 57.4% (8.9)	79.7% (4.7)	>** 62.9% (6.5)
	UseLink (%)	86.5% (4.2)	84.3% (4.7)	<i>See table caption[†]</i>	
	UseInternet (%)	23.0% (6.3)	<** 34.1% (7.8)	19.2% (6.5)	26.9% (8.1)
5.4.5	Correct (%)	73.6% (8.9)	75.7% (8.4)	69.4% (10.1)	70.3% (9.9)
	Time (min)	2.00 (0.29)	2.19 (0.29)	1.84 (0.25)	<* 2.23 (0.25)

perspective of AI’s uncertainty expression affect people’s source usage behavior. On the other hand, from our within-condition analysis, we find that AI’s uncertainty expression has a significant effect on the self-reported DVs at the instance level. On instances with uncertain AI responses (vs. not uncertain responses), UseAI is significantly lower in both UNCERTAIN1ST and UNCERTAINGENERAL, and UseInternet is significantly higher in UNCERTAIN1ST. These results suggest that **at the instance level, AI’s uncertainty expression decreases the use of the system’s answer and increases the use of other resources.** In their free-form responses, 11 participants stated that the system’s uncertainty motivated them to verify information using the links provided in the answer (e.g., “*I clicked on the links provided when the AI was uncertain of the answers*”) while 13 mentioned it motivated them to perform their own search (e.g., “*When the AI wasn’t certain, I searched on Google*”).

We note that both LinkClick and UseLink are intended to capture whether or not participants read the linked sources provided in the AI system’s responses, but the estimated means of LinkClick (2.7% – 7.2%) from the between-condition analysis are much lower than those of UseLink (74.7% — 85.5%). There are several factors that

might contribute to this discrepancy. First, these estimated means are from different models that include participants and questions as random effects. The intercepts for random effects in these models are quite high. Looking at the raw data, there is still a gap, but not as large: the actual means are 30.6% – 34.9% for **LinkClick** and 64.2% – 71.6% for **UseLink**. Second, there could have been measurement error from self-report bias for **UseLink** [98]. Finally, some participants could have considered their answers to be based on their “own reading of the linked sources” if they read the list of links, even if they didn’t click to open them. For example, one participant wrote, “*If the link title gave the same answer as the AI answer, then I assumed it was the right answer.*” This discrepancy emphasizes the value of including both behavioral and self-reported measures.

5.4.4 Trust and perception of AI

Moving onto trust and perception of AI, we find that **the presence and perspective of uncertainty expression neither affect people’s trust beliefs nor the perceived anthropomorphism and transparency of the system**. There are no significant differences in **TrustBelief**, **Anthropomorphism**, and **Transparency** between CONTROL and the two conditions with uncertainty. Overall, participants had **somewhat positive trust beliefs** about the system (**TrustBelief** is around “4: Somewhat agree” for all conditions); reported that they had a **somewhat good understanding of what the AI system’s answers were based on and when they might be wrong** (**Transparency** is around “4: Somewhat agree”); and had **neutral perceptions of anthropomorphism** (**Anthropomorphism** is around “3: Neutral”).

In contrast, we find that **first-person expressions of uncertainty decrease trust intentions while expressions from a general perspective do not**. **TrustIntention** is significantly lower in **UNCERTAIN1ST** (2.91) compared to both **CONTROL** (3.25) and **UNCERTAINGENERAL** (3.36), indicating a lower desire to use

the system. Illustrative of this, one participant in the UNCERTAIN1ST condition stated, “[*The AI system*] was very non-committal in its answers so I didn’t feel I could trust it.” This suggests that frequent first-person expressions of uncertainty can lead people to view the system as less trustworthy and decrease their desire to use it. More about the used scales and item-level results is in the supp. mat.

5.4.5 Task performance

Finally, we analyze participants’ task performance. From our between-condition analysis of **Correct**, we first find that **having access to the AI system decreases people’s accuracy**. Participants with access to the system have significantly lower accuracy than those without (63.9% CONTROL vs. 74.2% No-AI). However, this result should be interpreted in the context of the AI system’s low overall accuracy (50.0% in our experimental setup). Second, we find that **AI’s uncertainty expression increases people’s accuracy**. **Correct** is significantly higher in UNCERTAIN1ST (72.8%) than CONTROL (63.9%). It is also higher in UNCERTAINGENERAL (67.9%), but the difference is not significant.

From our between-condition analysis of **Time**, we find that **having access to the AI system increases task time**. **Time** is significantly higher in CONTROL (2.13 min) than No-AI (1.57 min). Together with the results on correctness, this suggests that **having access to the AI system decreases overall performance in our experimental setup**. Between CONTROL and the two conditions with uncertainty, there are no significant differences in task time. However, our within-condition analysis suggests that **AI’s uncertainty expression increases task time at the instance level**. The estimated means of **Time** for not uncertain vs. uncertain AI responses are 1.84 min vs. 2.23 min for UNCERTAINGENERAL (significantly different) and 2.00 min vs. 2.19 min for UNCERTAIN1ST (not significantly different). An

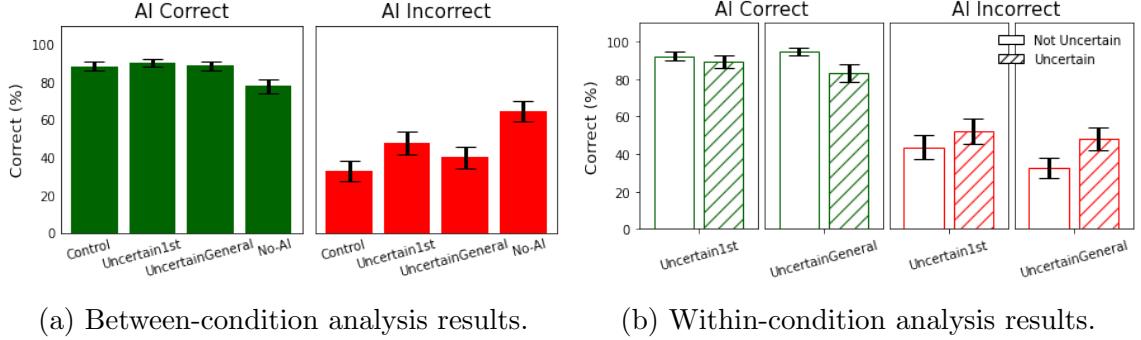


Figure 5.2: **Exploratory analysis of over- and underreliance (Section 5.5.1).** We analyze **Correct (%)** separately for questions the system answered correctly vs. incorrectly. We show the model-estimated means and standard errors for each condition (Figure 5.2a) and for not uncertain vs. uncertain responses in the conditions with uncertainty (Figure 5.2b).

explanation of this result is that AI’s uncertainty expression slows people down to use more caution when completing the task.

5.5 Results: Additional analyses

5.5.1 Effect of uncertainty expression on over- and underreliance

In Section 5.4.1, we analyzed participants’ agreement with the AI system as a measure of reliance. Agreement can be appropriate or inappropriate, depending on the correctness of the AI system’s answers; agreeing with the system when it is incorrect is a sign of overreliance, whereas disagreeing with it when it is correct is a sign of underreliance. To better understand the extent to which uncertainty expression leads to either of these phenomena, we take inspiration from the analysis of Chen et al. [75] and separately analyze participants’ agreement on questions the AI system answers correctly and questions it answers incorrectly. Note that, having conditioned on the (in)correctness of the AI system, analyzing agreement is equivalent to analyzing correctness of people’s final answers. We present the results here in terms of correctness

to focus attention on whether reliance is appropriate (i.e., beneficial to the user) or not.

Concretely, we run similar analyses to those presented in Section 5.4.5, but fit the regression models once on data from the four questions the system answered correctly and once on data from the four questions the system answered incorrectly (see Section 5.3.4). We show the estimated means (and standard errors) from these models in Figure 5.2.

First, by comparing the CONTROL and No-AI conditions, we observe that having access to the AI system’s answer increases accuracy when the system is correct (the estimated mean of **Correct** is 88.5% in CONTROL vs. 77.9% in No-AI), but decreases accuracy when it is incorrect (33.0% in CONTROL vs. 64.7% in No-AI), as in Figure 5.2a. Comparing CONTROL with UNCERTAIN1ST and UNCERTAINGENERAL, we see that having the AI system express uncertainty improves accuracy on questions that the system answers incorrectly without reducing accuracy when the system is correct. In line with our earlier results, expressing uncertainty in the first-person perspective leads to a bigger improvement in accuracy when the AI system is incorrect compared with expressing uncertainty in the general perspective.

To better understand how expressions of uncertainty drive accuracy, we break down the results further, comparing task accuracy on questions for which the system expresses uncertainty and those for which it does not (Figure 5.2b). We find that expressing uncertainty about a particular question leads to some reduction in accuracy when the AI system is correct (92.2% to 89.2% for UNCERTAIN1ST, 94.8% to 83.1% for UNCERTAINGENERAL), but a greater increase in accuracy when the AI system is incorrect (43.6% to 52.0% for UNCERTAIN1ST, 32.8% to 48.0% for UNCERTAINGENERAL).

While these results provide some evidence that expressions of uncertainty help reduce overreliance, we note that participants in the UNCERTAIN1ST and UNCER-

TAINGENERAL conditions still have substantially lower accuracy on questions where the AI system is incorrect compared with participants in the No-AI condition.

5.5.2 Participants' interpretations of AI's uncertainty expression

While our quantitative results shed light on whether expressions of uncertainty impact reliance and trust, they cannot tell us why. We next explore participants' interpretations of the expressed uncertainty via a thematic analysis of free-form responses to the question "When and why do you think AI system A expresses uncertainty?" in the exit questionnaire.

The majority of participants (N=102 of the 186 in the conditions with uncertainty) attributed the system's expressed uncertainty to its inability to answer a particular question, for example because it could not find an answer, found conflicting or unreliable answers, or could not understand the information it found. Three suggested the system was programmed to express uncertainty, as in "*I would guess there's some sort of certainty variable and if the score is below a level, an uncertainty message is included in the result.*" Another ten suggested the expressed uncertainty is due to the inherent difficulty of the question. As one put it, "*It could be a question that is very hard to come up with a simple yes or no answer.*"

These interpretations are all in line with the goal of reducing overreliance: if the system is unable to answer a question or the question is inherently difficult, users should verify the answer for themselves. Five participants explicitly interpreted the expressed uncertainty as a way of encouraging users to check their answers, for example, "*I suppose the AI wanted us to do further research in those cases where it could not be 100% sure of the answer.*"

A small number of participants attributed the uncertainty to other reasons, such as impression management ("*to appear more human, encourage confidence, and*

appear thoughtful”), maintaining credibility (“*It doesn’t want to risk being wrong on something and ruining its credibility*”), avoiding liability (“*it expresses uncertainty to absolve it of responsibility in the event it is wrong*”), or restrictions on answering medical questions (“*the AI is programmed not to dispense medical advice which could potentially be harmful without a caveat*”). These interpretations may not necessarily reduce overreliance.

One might ask whether participants interpreted the two types of uncertainty expression differently. Indeed, we found that participants in the UNCERTAINGENERAL condition were more likely than those in UNCERTAIN1ST (51.5% vs. 41.3%) to attribute the uncertainty to the AI system finding conflicting or unreliable information or the question being inherently hard, whereas those in UNCERTAIN1ST were more likely (20.7% vs. 7.4%) to attribute it to limitations of the AI system itself.

5.6 Discussion

Our results suggest that expressing uncertainty through natural language can be an effective way to reduce overreliance and (over)trust in LLM-infused search engines. Expressions of uncertainty led to more cautious behaviors, from taking longer to arrive at an answer to reporting more reliance on outside sources. However, it did not fully eliminate overreliance; the participants with the highest task performance were those without access to AI responses.

We find that perspective matters: uncertainty expressions in first-person show stronger effects than general perspective. This is consistent with prior findings that first-person messages increase recipients’ involvement and engagement [73, 434] compared with general or third-person messages. This persuasive effect of first-person expressions should be interpreted with caution: while it helps heighten the warning effect of a negative message such as uncertainty, it might amplify a positive message,

even if unjustified. For example, one may want to avoid first-person expressions of *confidence* because they may exacerbate overreliance and over-trust, as found in prior work [444]. There are also concerns around harms from anthropomorphism of AI systems that may stem from over-trust, deception, threats to human agency, and propagation of stereotypes [3]. While we did not observe that first-person uncertainty expression increases perceived anthropomorphism, people can start assigning social attributes to machines without conscious awareness [283]. Future research and practices should further explore the long-term effects of interacting with AI systems expressing uncertainty in a first-person perspective and consider other potential negative effects of anthropomorphism.

Our research has implications both for those building and deploying LLMs and LLM-infused applications and for policymakers regulating the use of AI. Most critically, any approach to reducing overreliance should be validated through empirical research. There may also be tradeoffs when balancing over- and underreliance. In our study, the most successful approach to reducing overreliance was to use first-person uncertainty expression, but this also decreased participant trust in the AI system, which may be undesirable in settings where people already under-trust the AI system. We believe there is no one-size-fits-all approach to implementing natural language uncertainty expression. For these reasons, given that the issue of how to manage overreliance is of particular importance to regulators, we advocate for raising awareness of the complexities of mitigating overreliance and for customized, evidence-based solutions, rather than universal ones.

There are limitations to our research. The widespread deployment of LLMs is still relatively new and the human-computer interaction and broader research communities are still grappling with the question of how to design effective studies to understand how end users perceive and interact with them. There is always a tradeoff between the controllability of the experiment and the generalizability of the conclusions to

user behaviors in their day-to-day tasks [263]. In order to be able to measure agreement and correctness, we adopted an experimental set-up inspired by the AI-assisted decision-making literature in which study participants provide simple yes/no answers to questions. This approach does not allow exploring how the expression of uncertainty would impact people’s behavior when completing more complex tasks, like writing an article or planning a trip. Further, our measurements of time and source usage are less reliable than they would have been had we opted for an in-person lab study. We chose questions from the medical domain, where overreliance is particularly concerning, but people may behave differently when seeking information about their own medical symptoms rather than answering a pre-defined set of questions. People may also behave differently when given the chance to interact with the system repeatedly as opposed to in a single session. The AI system in our study exhibited low accuracy and expressed uncertainty often, in a poorly calibrated manner. These design choices may have impacted our results — particularly the lower task performance when given access to the AI system. Also, there may be differences across cultures and languages in how people interpret or react to uncertainty. Our study was conducted in English with U.S.-based participants and results may not generalize to other cultural and linguistic contexts.

For all of these reasons, while our findings suggest that natural language expressions of uncertainty could be an effective approach to reducing overreliance, we caution against overgeneralizing from our study. Instead, we view our results as evidence that language choices matter in how people perceive and act on the outputs of LLMs, and teams building and deploying LLMs should therefore evaluate them carefully with end users before release.

5.7 Ethical considerations and positionality

We conclude with a reflection on the ethical considerations of our work and our positionality.

Mitigating harms to human subjects. We recruited U.S.-based participants on MTurk, which many people rely on as a primary source of income. As discussed in Section 5.3.5, we aimed to provide an hourly wage of \$15 USD. We came close to this goal, with participants receiving an estimated \$14.80 per hour on average. (This is likely an underestimate of average wage, since we have no way to know if workers spent time on other activities between accepting the task and completing it.) This is substantially higher than the U.S. federal minimum wage of \$7.25 per hour, though a few states have recently adopted a minimum wage of \$15/hour or higher. As discussed in the FAccT 2023 panel “The Humans Behind the Intelligence: Speaking with Data Workers,” our choice to limit participation to workers with a 99% or higher approval rating, at least 2000 completed tasks, and in some cases, a Masters qualification, prevented workers who are new to MTurk from participating. We made the decision to include these qualifications after piloting several versions of the study with less restrictive qualifications and finding that the data quality was too poor to use. We paid and approved the work of everyone who completed the study, regardless of whether their responses passed our quality checks. At the end of the study, we debriefed participants, reminding them that the medical information output by the AI system was sometimes incorrect. We did not collect personally identifiable information except for MTurk IDs, which were used to ensure that workers who participated in pilots of our study did not participate in the main study. These were deleted when no longer needed. Our procedure was reviewed and approved by our internal IRB and we obtained participant consent.

Potential negative societal impact. While our results provide evidence for the effectiveness of natural language expressions of uncertainty for reducing overreliance, generalizing too heavily from our findings could lead to potential harms. Given the limitations of our research (see Section 5.6), teams deploying LLMs or LLM-infused applications should not make decisions about how to express uncertainty to end users without extensive user testing in their own contexts. They also should not assume that they have addressed overreliance by expressing uncertainty. (Indeed, in our study we see that participants still have higher task performance with no access to the AI system.) Likewise, regulators should avoid making blanket requirements on uncertainty expression, at least until more research has been done.

Separate from these potential unintentional misuses of our research, there is a possibility that bad actors could strategically incorporate the expression of uncertainty into an LLM’s output to make them more persuasive, regardless of whether they represent objective fact, potentially contributing to the spread of misinformation.

Positionality. Our research questions and design were influenced by our position as employees of a U.S.-based technology company. Members of our research team have first-hand experience observing and participating in discussions regarding the responsible development and deployment of LLM-infused applications, which has shaped our understanding of gaps in knowledge and other practical challenges that arise in attempting to meet responsible AI principles and proposed regulatory requirements. We had access to sufficient budget to run large-scale experiments, which is not an option for some research teams. Our view that there are potential benefits of responsibly deployed LLM-infused applications is likely influenced by our experience in industry, yet we acknowledge some fundamental limitations of using

LLMs for information retrieval [341]. We encourage future research on overreliance from research teams in academia and civil society.

Acknowledgements

We thank the participants for their time and effort and the anonymous reviewers for helpful comments. We also thank Jake Hofman, Dan Goldstein, Sophie Spatharioti, Amy Heger, Samir Passi, Harsh Kumar, Wesley Deng, Snehal Prabhudesai, Abhraneel Sarma, Yeji Park, Nina Grgić-Hlača, Julia Stoyanovich, Jessica Hullman, and the Microsoft Research New York City lab and FATE group for thoughtful discussion and feedback. SK acknowledges support from the NSF Graduate Research Fellowship and the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award to her PhD advisor.

Chapter 6

Fostering appropriate reliance on large language models

Large language models (LLMs) can produce erroneous responses that sound fluent and convincing, raising the risk that users will rely on these responses as if they were correct. In this chapter, we present two studies aimed to foster appropriate user reliance on LLMs. Through a think-aloud study in which participants use an LLM-infused application to answer objective questions, we identify several features of LLM responses that shape users' reliance: *explanations* (supporting details for answers), *inconsistencies* in explanations, and *sources*. Through a large-scale, pre-registered, controlled experiment ($N=308$), we isolate and study the effects of these features on users' reliance, accuracy, and other measures. We find that the presence of explanations increases reliance on both correct and incorrect responses. However, we observe less reliance on incorrect responses when sources are provided or when explanations exhibit inconsistencies. We discuss the implications of these findings for fostering appropriate reliance on LLMs.

This chapter contains materials published in the CHI 2025 paper, Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources,

and Inconsistencies [194], in collaboration with Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky.

6.1 Introduction

Large language models (LLMs) are powerful tools, capable of a wide range of tasks from text summarization to sentence completion to code generation. Technology companies have leapt at the unprecedented opportunity to build LLM-infused applications that help users with information retrieval and search, learning new things, and performing everyday tasks more efficiently. Many such applications, such as LLM-infused search engines and chatbots, are predicated on LLMs’ ability to provide intricate responses to complex user questions. Already millions of people use LLMs to find answers to their questions about health, science, current events, and other domains, and the use of LLMs is widely predicted to grow [179, 343, 446]. However, the responses produced by LLMs are often inaccurate, sometimes in subtle ways [92, 171, 331, 348]. While such *overreliance* on AI systems is not a new problem [26, 304, 312, 400, 437], it may be exacerbated by the introduction of LLMs, since LLM responses are often fluent and convincing even when wrong and public excitement around LLMs is high.

When asked to answer a question, LLMs and systems based on them typically provide a response that contains both an answer to the question and some supporting details or justification for this answer [222, 418]. For example, when asked a math question, an LLM may provide a step-by-step derivation for its answer [87, 149]. In line with everyday usage and much of the psychology literature [183, 241, 245], we refer to such supporting details as an *explanation* of the answer. (We note that this differs from how the term explanation is often used within the explainable AI community in that we do not make any assumptions about the extent to which it faithfully describes

the way that the model arrived at its answer. That is, the explanation describes why the answer is correct, not necessarily why the model output the answer that it did.) Some authors have argued that such explanations should help users spot incorrect answers, potentially mitigating overreliance [62, 137, 212, 384]. However, prior work suggests that in many settings, the very presence of an explanation can increase trust and reliance, whether or not it is warranted [26, 120, 299, 312, 400, 437]. To avoid such unintended negative consequences, it is necessary to understand how users interpret and act upon explanations from LLMs, and how explanations and other features of LLM responses might be adjusted to encourage appropriate reliance.

To explore these questions, we first conduct a think-aloud study with 16 participants with varying knowledge of and experience with LLMs. In this study, participants answer objective questions with the use of the popular LLM-infused application ChatGPT via multi-turn interactions. The goal of this preliminary study is to understand how people perceive LLM responses and which features of a response shape their reliance. We observe that participants interpret *inconsistencies* in explanations — that is, sets of statements that cannot be true at the same time [164] — as a cue of unreliability. Participants also seek out *sources* to verify supporting details in LLM responses and are less likely to rely on incorrect answers when the sources provide helpful information.

Building on the findings from this study, we next conduct a large-scale, pre-registered, controlled experiment ($N = 308$) in which participants answer difficult objective questions with access to LLM responses, i.e., responses from a hypothetical LLM named “Theta.”¹ These responses were created in advance using state-of-the-art LLM-infused applications ChatGPT and Perplexity AI so that we can fully control their features. Specifically, we employ a $2 \times 2 \times 2$ within-subjects

¹We note that the line between what we would call an “LLM” as opposed to an “LLM-infused system” can be blurry, especially when the system takes the form of a chatbot such as Theta or ChatGPT. Throughout the paper, use the term LLM for readability in places where the distinction is not important.

design, varying three features of the LLM responses: accuracy of the LLM’s answer to the question (correct/incorrect), presence of an explanation (absent/present), and presence of clickable sources (absent/present). Further, we capitalize on the natural inconsistencies that arise in LLM responses to investigate the effects of inconsistencies. We examine the impact of these variables on participants’ reliance, accuracy, and other measures, such as confidence, source clicking behavior, time on task, evaluation of LLM responses, and likelihood of asking follow-up questions.

We find that when either or both an explanation and sources are present, participants report higher confidence in their answer, rate the LLM response higher in terms of the quality of the justification it provides for the answer and the actionability of its response, and are less likely to ask follow-up questions. However, explanations and sources differ in their effects on reliance. Explanations increase reliance on both correct LLM answers and incorrect LLM answers. In contrast, sources increase appropriate reliance on correct LLM answers, although less effectively than explanations, while decreasing overreliance on incorrect LLM answers. Finally, when explanations have inconsistencies, we observe less overreliance on incorrect LLM answers compared to when there are no inconsistencies or when explanations are not provided at all. We complement these quantitative findings with qualitative insights and close with a discussion of implications and future research directions for fostering appropriate reliance on LLMs.

Together, our approach and findings offer a number of contributions. (1) Our studies tackle the timely and critical issue of fostering appropriate reliance on LLMs. Since research on user reliance on LLMs is relatively new, we take a mixed-methods approach, first (via the think-aloud study) identifying features of LLM responses that shape user reliance, and then (via the controlled experiment) isolating and studying the effects of the identified features. (2) Through our two studies, we identify which combinations of features help people achieve appropriate reliance and high task

accuracy, providing actionable insights on how to adjust LLM response features. We also contribute a more holistic and nuanced understanding of user reliance on LLMs with insights on people’s interpretation of explanations from LLMs, source clicking behavior, and interaction effects between explanations and sources. (3) We provide an in-depth discussion of the implications of our findings, limitations of our work, and future research directions. In particular, we identify providing (accurate and relevant) sources and highlighting inconsistencies and other unreliability cues in LLM responses as promising strategies for fostering appropriate reliance on LLMs. However, such approaches should always be tested with users before deployment.

6.2 Related work

6.2.1 Appropriate reliance on AI

Despite the rapid progress of technology, AI systems still frequently and unexpectedly fail. Without knowing when and how much to rely on a system, a user may experience low-quality interactions or even safety risks in high-stakes settings. Prior work has investigated how providing information about an AI system’s accuracy [147, 424, 426] and (un)certainty [26, 58, 62, 139, 437], explanations of outputs [26, 58, 62, 137, 139, 212, 437], and onboarding materials [68, 213] impact user reliance, as well as the roles played by human intuition [75], task complexity [327, 328], and other human, AI, and context-related factors [191]. However, fostering appropriate reliance on AI remains difficult. Findings on the effectiveness of proposed methods are mixed, and more research is needed on how reliance is shaped in real-world settings.

While most prior work on AI reliance has been in the context of classical AI models (e.g., specialized classification models), there is a growing body of work looking at reliance on systems based on LLMs or other modern generative AI models [193, 222, 349, 357, 385, 444]. For example, several recent studies explored the effect

of communicating (un)certainty in LLMs by highlighting uncertain parts of LLM responses [357, 385] or inserting natural language expressions of uncertainty [193, 444], finding that some but not all types of (un)certainty information help foster appropriate reliance.

Contributing to this line of work, we first take a bottom-up approach to identify the features of LLM responses that impact user reliance in the context of answering objective questions with the assistance of a popular LLM-infused application ChatGPT (Section 6.3). In line with findings from prior work [349], we see that reliance is shaped by the content of *explanations* provided by the system, particularly whether or not these explanations contain *inconsistencies*. We also observe that participants seek out *sources* to verify the information provided in responses. We then design a large-scale, pre-registered, controlled experiment to isolate and study the effects of these features (Section 6.4). We discuss the relevant literature on these features and their impact on AI reliance next.

6.2.2 Explanations and inconsistencies

The impact of *explanations* on human understanding and trust of AI systems has been studied extensively within the machine learning and human-computer interaction communities, often under the names explainable AI or interpretable machine learning [27, 192, 225, 323, 414]. Explanations are often motivated as a way to foster appropriate reliance and trust in AI systems, since in principle they provide clues about whether a system’s outputs are reliable. However, empirical studies have shown mixed results, with a large body of work suggesting that providing explanations increases people’s tendency to rely on an AI system even when it is incorrect [26, 312, 400, 437]. One potential reason for this is that study participants do not make the effort to deeply engage with the explanations [60, 129, 182, 224, 384]. That is, instead of encouraging deep, analytical reasoning (System 2 thinking [175, 176]),

study participants may resort to heuristics, such as the explanation’s fluency or superficial cues to expertise [375], and defer to the system’s response on this basis. People may also be more likely to assume an AI system is trustworthy simply because it provides explanations [111]. Further, some clues of unreliability may be difficult to pick up on without existing domain knowledge [75].

Adopting the broad definition of an explanation as an answer to a why question [53, 127, 245, 408], LLMs often provide explanations by default; when asked a question, LLMs rarely provide the answer alone. For factual questions, they provide details supporting the answer [222, 418], and for math questions, they provide detailed steps to derive the answer [87, 149]. This default behavior is likely due to human preference for verbose responses [79, 325, 440]. Research in psychology has shown that explanations are often sought spontaneously [128, 254], favored when they are longer, more detailed, or perceived to be more informative [16, 32, 235, 406, 430], and used to guide subsequent judgments and behaviors [243, 244]. Since LLMs are often fine-tuned on human preference data via approaches such as Reinforcement Learning from Human Feedback (RLHF) [82, 296, 447], such preferences would shape the form of their outputs. We note that the default explanations that LLMs present typically provide evidence to support their answers, but do not necessarily reflect the internal processes by which the LLM arrived at the answer. This distinguishes these explanations from those traditionally studied in the explainable AI literature.

Explanations generated by LLMs are widely known to contain inaccurate information and other flaws [92, 171, 331, 348]. We direct readers to recent surveys for comprehensive overviews [162, 393]. In our studies, we found *inconsistencies* in explanations to be an important unreliability cue that shapes participants’ reliance. As documented in prior work, inconsistencies can occur within a response; they are sometimes referred to as logical fallacies or self-inconsistency in the NLP community [161, 401]. Inconsistencies can also occur between responses; many studies have

demonstrated that LLMs often change their answer to a question when challenged, asked the question in a slightly different way, or re-asked the exact same question [113, 209, 222]. Such inconsistencies, when noticed, may impact people’s evaluation of explanations and reliance on LLMs.

We contribute to this line of work in several ways. We first conduct a qualitative, think-aloud study to understand what features of LLM responses shape people’s reliance, and find that reliance is shaped by explanations, inconsistencies in explanations, and sources. We then conduct a larger-scale, pre-registered, controlled experiment to quantitatively examine the effects of these features. While a previous work by Si et al. [349] has studied the effects of LLM-generated explanations and inconsistencies on people’s fact-checking performance through a small-scale study (16 participants per condition), our work provides a more holistic picture by studying what (else) might contribute to reliance and how the identified features affect a wider range of variables including people’s evaluation of the LLM response’s justification quality and actionability and likelihood of asking follow-up questions. As for the findings, first, consistent with Si et al. [349], we find that explanations increase people’s reliance, including overreliance on incorrect answers, and that inconsistencies in explanations can reduce overreliance. Additionally, we find that clickable sources — which were not studied by Si et al. [349] — increase appropriate reliance on correct answers, while reducing overreliance on incorrect answers, adding empirical knowledge on user reliance on LLMs. Lastly, our work also contributes nuanced insights on people’s interpretation of LLMs’ explanations, source clicking behavior, and interaction effects between explanations and sources.

6.2.3 Sources

The final feature of LLM responses that we study is the presence of *sources*, i.e., clickable links to external material.² Sources are increasingly provided by LLM-infused applications, including general-purpose chatbots (e.g., ChatGPT, Gemini) and search engines (e.g., Perplexity AI, Copilot in Bing, SearchGPT). Sources are commonly sought by users, as found in prior work [195] and supported in our studies. Similar to explanations, however, sources in LLM responses can be flawed in various ways [11, 238]. For instance, Liu et al. [238] conducted a human evaluation of popular LLM-infused search engines and found that their responses frequently contain inaccurate sources and unsupported statements. Alkaissi and McFarlane [11] conducted a case study of ChatGPT in the medical domain and found that it generates fake sources. These issues were observed in our studies as well. Currently there is active research on techniques such as Retrieval Augmented Generation (RAG) [130, 223] to help LLMs provide more accurate information and sources.

It is well known that the presence and quality of sources impact how credible people find given content in other settings [321, 402]. However, there has been little work studying how people make use of and rely on sources in the context of LLM-infused applications. On the one hand, the presence of sources might reduce overreliance if people click on the provided links to verify the accuracy of the LLM’s response. On the other hand, the presence of sources might increase reliance if people interpret them as signs of credibility and defer to the system without verifying the answers themselves. Indeed, in one study of uncertainty communication in LLM-infused search, participants were found to rarely click on source links [193]. Through a large-scale, pre-registered, controlled experiment (Section 6.4), we study how the presence of clickable sources impacts people’s reliance, task accuracy, and

²One might consider sources to be a component of an explanation. To simplify the exposition of our results, we treat them as a distinct component of LLM responses throughout this paper.

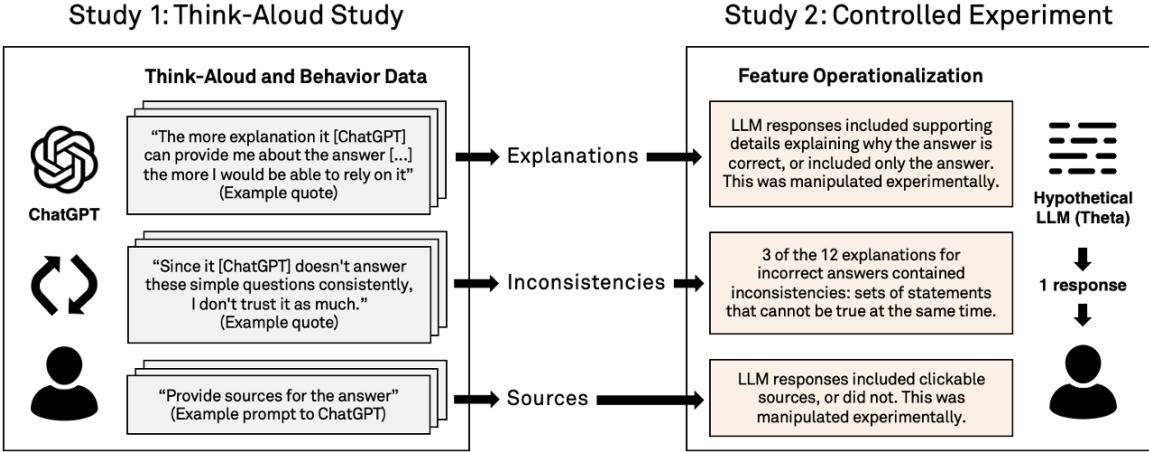


Figure 6.1: **Overview of our studies.** In Study 1, participants engaged in multi-turn interactions with ChatGPT to arrive at correct answers to objective questions. Based on a thematic analysis of think-aloud and behavioral data, we identified *explanations*, *inconsistencies*, and *sources* as three features of LLM responses likely to influence user reliance. These three features were then investigated in a controlled experiment (Study 2), with features operationalized as indicated in the schematic illustration. Similar to Study 1, participants solved question-answering tasks. However, this time, they had access to one LLM response whose features we experimentally manipulated.

other measures, and how this interacts with the presence of explanations and inconsistencies. In our studies, we use realistic explanations and sources, generated by state-of-the-art LLM-infused applications ChatGPT and Perplexity AI, and provide insights for fostering appropriate reliance on LLMs.

6.3 Study 1: Think-aloud study

Towards the goal of identifying features of LLM responses that can help foster appropriate reliance, we first take a bottom-up approach and conduct a think-aloud study in a relatively natural setting. Specifically, we observe how participants solve question-answering tasks with ChatGPT in multi-turn interactions, and explore how they perceive ChatGPT’s responses and what helps them arrive at correct answers despite incorrect answers from ChatGPT.

6.3.1 Study 1 methods

In this section, we describe our study methods, all of which were reviewed and approved by our Institutional Review Board (IRB) prior to conducting the study.

Procedure

The study session had two parts. In Part 1 (Base), participants were introduced to the study and asked to complete three question-answering tasks while thinking aloud. Each task involved determining the correct answer to an objective question using ChatGPT³ and reporting confidence in their final answer on a 1–7 scale. As in natural settings, participants could exchange as many messages with ChatGPT as they wished. Participants could also check the sources provided in ChatGPT’s responses, but were asked not to conduct their own internet search.

Each participant was given three questions: a general domain factual question (e.g., “Has Paris hosted the Summer Olympics more times than Tokyo?”), a health or legal domain factual question (e.g., “Is it illegal to collect rainwater in Colorado?”), and a math question (e.g., “Sue puts one grain of rice on the first square of a Go board and puts double the amount on the next square. How many grains of rice does Sue put on the last square?”). The factual questions were binary questions. The math questions were not binary, but had one correct numerical answer. The specific question was randomly selected from a set of questions we created in advance based on examples of real user-LLM interactions [297] and prior work [346, 416]. Before beginning the tasks, we also asked each participant if they knew the answer to any of the questions so that we could switch to a different question if they did, but this did not happen.

³We created a research account with a Plus subscription. Participants logged into our account and used ChatGPT-4o — the latest version at the time (June 2024) — through the web interface with Browsing allowed and Memory disallowed.

In Part 2 (Prompting), we asked participants to complete the same three tasks again, but this time while employing follow-up prompts in their engagement with ChatGPT. We designed Part 2 to explore whether certain prompts can help participants more appropriately rely on ChatGPT and succeed on the tasks. Since participants had different levels of familiarity with prompting, we provided examples of prompts they could use, such as asking for a different type of explanation (e.g., “Explain step by step” and “Explain like I’m five”), asking for more information (e.g., “Provide an explanation with supporting sources” and “Explain how confident you are in the answer”), and challenging the previous response (e.g., “Explain why your answer may be wrong” and “I think you are wrong. Try again”). Participants could use whichever and as many prompts as they wished. As in Part 1, participants reported their final answer and confidence in their final answer at the end of each task.

In between Part 1 and Part 2 and before concluding the study, we asked interview questions about participants’ perception of and experience with ChatGPT. Details are in the supp. mat.

Participant recruitment and selection

To recruit participants, we posted a screening survey on Mastodon, X (previously Twitter), and various mailing lists and Slack workspaces within and outside the first author’s institution. The survey included questions about the respondent’s knowledge and use of LLMs. Based on the survey responses, we selectively enrolled participants to maximize the diversity of the study sample’s LLM background. See below for a summary of participants’ knowledge and use of LLMs. We manually reassigned two participants to different categories than what they selected in their survey when their survey responses did not line up with their described experience (high to low knowledge for one participant and low to high knowledge for another). We refer to individual participants by identifier P#.

- *Low-knowledge*: “Slightly familiar, I have heard of them or have some idea of what they are” (P6, P9, P13, P15) or “Moderately familiar, I know what they are and can explain” (P2, P3, P11, P14).
- *High-knowledge*: “Very familiar, I have technical knowledge of what they are and how they work” (P1, P4, P8, P10, P16) or “Extremely familiar, I consider myself an expert on them” (P5, P7, P12).
- *Low-use*: “Never” use LLMs (P5, P13, P15, P16) or use LLMs “Rarely, about 1–2 times a month” (P4) or use LLMs “Sometimes, about 3-4 times a month” (P3, P6, P8).
- *High-use*: Use LLMs “Always, about once or more a day” (P1, P2, P7, P9, P10, P11, P12, P14).

Conducting and analyzing studies

We collected data from 16 participants in June 2024, each over a Zoom video call. The study lasted one hour on average, and participants were paid \$20 for their participation. All sessions were video recorded and transcribed for data analysis. We used a mix of quantitative and qualitative methods to analyze the study data. On the quantitative side, we analyzed the accuracy of participants’ answers and their self-reported confidence in their answers measured on a 1–7 scale for each task. Since each participant solved three tasks, once in Part 1 and again in Part 2, there are 6 accuracy and 6 confidence numbers for each participant. On the qualitative side, we conducted a thematic analysis [48, 51] of participants’ think-aloud data and their responses to interview questions to identify features of LLM responses that shaped participants’ reliance. The first author performed the initial coding, discussed the categories with other authors, and then refined the coding.

6.3.2 Study 1 results

We first provide some descriptive statistics about participants' accuracy, over- and underreliance, and confidence across the two parts of the study (Section 6.3.2). We then discuss which LLM response features participants reported as influences on their reliance (Section 6.3.2). We emphasize that this study was not intended to provide statistically significant results, but to identify features that may help foster appropriate reliance. Given the small sample size, we report the quantitative results only to provide context.

Accuracy, reliance, and confidence

In Part 1 (Base), we collected data on 48 task instances (16 participants \times 3 tasks). For 34 of these instances, ChatGPT gave a correct answer in its first response. (ChatGPT sometimes changed its answer over the course of the interaction, either due to stochasticity or in response to participants' follow-up messages.) Among these, participants' final answer agreed with ChatGPT's correct answer in 33 instances (average confidence 5.97 on the 1–7 scale) and disagreed in only a single instance (confidence 4.5), indicating that **underreliance was not prevalent**. In 13 instances, ChatGPT gave an incorrect answer in its first response. Among these, participants' final answer agreed with ChatGPT's incorrect answer in 9 instances (average confidence 6.15) and disagreed in only 4 instances (average confidence 5.61), indicating **widespread overreliance**. In a single instance, ChatGPT did not answer the question in its first response, and the participant submitted an incorrect answer with a confidence of 2.

We did not find meaningful differences in participants' accuracy between the two parts of the study. That is, **follow-up prompting did not increase participants' accuracy**, at least based on our small sample of quantitative data. For 44 out of 47 instances in which the participant completed Part 2 (Prompting) (one participant had to skip a task instance due to lack of time), the participant submitted the same answer

in both parts. In 3 instances, participants submitted an incorrect answer in Part 1 and a correct answer in Part 2. In 2 of these 3 instances, ChatGPT gave an incorrect answer in Part 1, but gave a correct answer in Part 2. In the other instance, ChatGPT gave incorrect answers in both parts, but the participant arrived at the correct answer in Part 2 after engaging in multiple rounds of interaction with ChatGPT.

Finally, we compared participants' confidence in their answers for the same task between the two parts, finding that it increased in Part 2 in 19 instances, decreased in 8 instances, and stayed the same in 20 instances. However, **changes in confidence do not correspond to changes in answers**. As mentioned above, participants changed their answers in only 3 out of 47 instances. In these 3 instances, participants' confidence stayed the same or increased slightly as their answer changed from being incorrect to correct. Participants' self-described reasons for increased confidence included seeing and checking sources, seeing ChatGPT give the same answer multiple times, and receiving more information in general. Reasons for decreased confidence included experiencing issues with sources (e.g., links were broken or sources were not reputable) and seeing ChatGPT change answers.

LLM response features shaping reliance

From a thematic analysis of participants' think-aloud data and responses to interview questions, we found **explanations**, **inconsistencies**, and **sources** to be key features of LLM responses that participants reported as influences on reliance. First, consistent with our discussion in Sections 6.1 and 6.2.2, we observed that ChatGPT provided **explanations** of its answers by default. Participants found these explanations important for judging the reliability of ChatGPT's answers. For example, P14 (low-knowledge, high-use) described explanations as "*very important for having reliability on the answer*" and said "*the more explanation it [ChatGPT] can provide me about the answer [...] the more I would be able to rely on it.*" P11 (high-knowledge, high-

use) added that they judge the response by “*how well ChatGPT explains the answer.*” This participant judged ChatGPT’s explanation in one task to be very high quality, noting “*I would put this on my homework and submit it [...] the quality is very high*”.

However, in another task, P11 submitted a different answer from ChatGPT after observing **inconsistencies**: “*Since it [ChatGPT] doesn’t answer these simple questions consistently, I don’t trust it as much.*” Sometimes inconsistencies occurred within a response (e.g., ChatGPT saying Paris hosted the Summer Olympics more times than Tokyo while also saying both have hosted twice). At other times inconsistencies occurred across multiple responses (e.g., ChatGPT changing its answer when asked the same or similar questions, or when challenged). In either case, **when participants observed inconsistencies, they often asked follow-up questions and engaged more with the system to resolve the inconsistencies**. For example, when P8 (high-knowledge, low-use) was considering the question “Did Tesla debut its first car model before or after Dropbox was founded?” ChatGPT initially stated that Tesla debuted its first car model in 2008 then later changed the year to 2006. After noticing the inconsistencies, P8 engaged in three more rounds of interaction with ChatGPT to verify individual pieces of information, and arrived at the correct answer.

Finally, participants frequently sought and used **sources** to determine whether or not to rely on ChatGPT. More often than not, ChatGPT did not provide sources as part of its responses, even though participants were using the latest version at the time of the study (4o) with browsing capabilities. Participants had to explicitly ask for them using prompts like “Provide sources for the answer.” Participants rarely did this in Part 1, and as such, sources were provided in only 17 out of 48 instances. However, in Part 2, participants asked for sources more often after seeing prompt examples and were provided sources in 30 instances. **When participants checked sources, they were often able to avoid overreliance on ChatGPT.** For example, out of 11 instances in which participants submitted correct answers despite

incorrect answers from ChatGPT (both parts combined), 7 were instances in which participants checked sources. (In the other 4 instances, sources were not provided, but participants were able to submit correct answers through other strategies, such as repeatedly asking ChatGPT about a piece of information.) For example, when P2 (low-knowledge, high-use) was solving the question “Sue puts one grain of rice on the first square of a Go board and puts double the amount on the next square. How many grains of rice does Sue put on the last square?” ChatGPT built on an incorrect assumption about the size of a Go board and gave an incorrect answer. P2 initially judged it as correct, but after checking sources, realized ChatGPT’s error and was able to submit a correct answer.

As discussed in Section 6.3.2, sources also influenced participants’ confidence in their answers. **The presence of sources increased confidence in general, except when there were issues with sources.** For example, P1 (high-knowledge, high-use) said their confidence increased in Part 2 for one task when they received sources and were able to verify information in ChatGPT’s responses. But they said their confidence decreased for another task when some of the source links did not open or did not contain relevant information, highlighting the importance of source *quality* in addition to *presence*. Finally, we emphasize that **checking sources did not always eliminate overreliance.** Out of 30 instances in which participants checked sources (both parts combined), in 4 instances, participants’ final answer still agreed with ChatGPT’s incorrect answer, which is a sign of overreliance.

6.4 Study 2: Large-scale, pre-registered, controlled experiment

Based on the insights from Study 1, we designed a large-scale, pre-registered, controlled experiment to study the effects of different features of LLM responses on

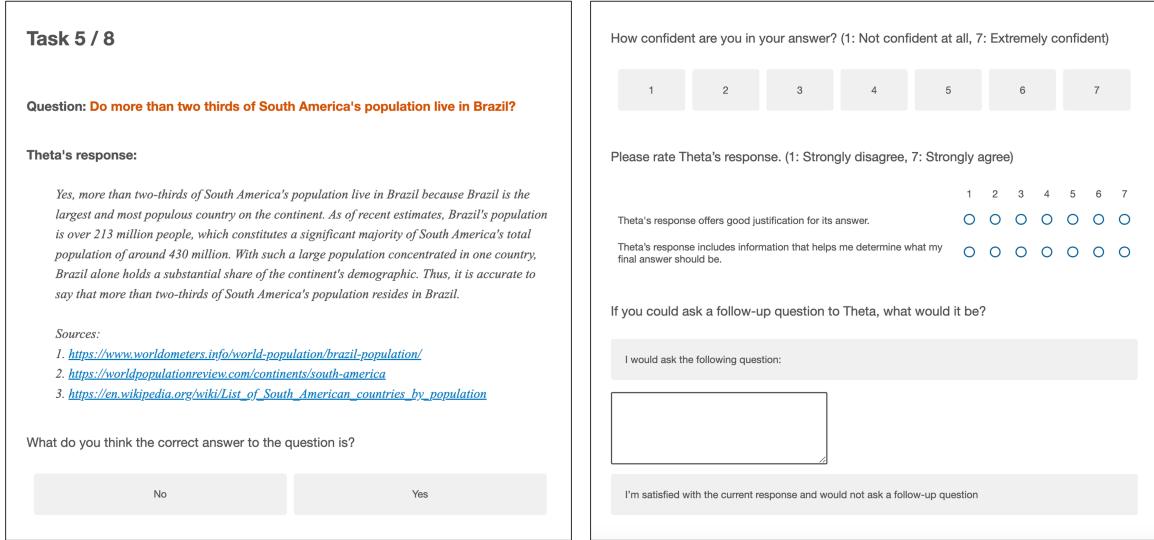


Figure 6.2: **Screenshots of Study 2’s experimental task.** Here the LLM response provides an incorrect answer, includes sources, and includes an explanation (with inconsistencies). See Figure 6.3 for responses with a correct answer for the same task question.

people’s reliance, task accuracy, and other measures including confidence, source clicking behavior, time on task, evaluation of LLM responses, and asking of follow-up questions. The goal of the study was to test whether the findings from Study 1 apply at scale and identify which combinations of features help people achieve appropriate reliance and high task accuracy.

6.4.1 Study 2 methods

In this section, we describe our study methods. Before collecting data, we obtained IRB approval and pre-registered our experimental design, analysis plan, and data collection procedures.⁴

Procedure

We designed a within-subjects experiment in which participants completed a set of question-answering tasks with LLM responses. Each task involved determining the

⁴Our pre-registration is viewable at <https://aspredicted.org/bg22-yfw7.pdf>.

correct answer to a binary factual question with access to a response from a hypothetical LLM named “Theta” (hereafter we occasionally refer to it as “the LLM”). See Figure 6.2 for an example. Our experiment had a $2 \times 2 \times 2$ design where we varied three variables in Theta’s responses: accuracy of Theta’s answer to the question (correct/incorrect), presence of an explanation (absent/present), and presence of clickable sources (absent/present). In total, there were 8 types of responses. Participants completed 8 tasks in the experiment and saw one of each type. This makes Theta’s accuracy 50%, but participants were not given this information: participants did not receive feedback on whether their answer or Theta’s answer was correct after solving a task. See Figure 6.3 for examples of different types of responses.

The experiment had three parts. In the first part, participants were introduced to the study and to Theta. Theta was described as an LLM-based AI system prototype that uses similar technology to OpenAI’s ChatGPT, is connected to the internet, and can answer a wide range of questions. In the second part, participants answered a total of eight questions. For each question, participants were provided with a response from Theta and were asked to submit their answer, report their confidence in their answer, and rate Theta’s response. They were told that they could click on source links in Theta’s responses, but asked not to conduct their own internet search. Participants could also optionally write a follow-up question, but they did not see Theta’s response to it. We made this choice to fully control the number and content of responses, while being able to collect data on when and what types of follow-up questions participants ask. We acknowledge that showing one controlled response instead of allowing free-form interaction has limitations (see Section 6.5.3). However, we adopt this method from prior work studying LLMs [193, 222, 349] as a valid approach for capturing user perceptions and behaviors around LLM responses with the advantage of controlling unwanted noise from free-form interactions (for instance, LLMs making different mistakes across participants in follow-up interactions).

<p>Neither</p> <p>No, not more than two-thirds of South America's population live in Brazil.</p>	<p>Sources only</p> <p>No, not more than two-thirds of South America's population live in Brazil.</p> <p>Sources:</p> <ol style="list-style-type: none"> 1. https://www.cia.gov/the-world-factbook/countries/brazil/ 2. https://www.macrotrends.net/global-metrics/countries/BRA/brazil/population 3. https://www.worldometers.info/world-population/south-america-population/
<p>Explanation only</p> <p>No, not more than two-thirds of South America's population live in Brazil. Although Brazil is the largest country in South America by both land area and population, it does not encompass the majority of the continent's people. As of recent estimates, Brazil's population is roughly 213 million, while South America's total population is about 440 million. This means Brazil's population constitutes roughly 48% of the continent's total, which is significantly less than two-thirds. Hence, while Brazil has a large population, it does not exceed the two-thirds threshold relative to the entire continent's population.</p>	<p>Explanation & Sources</p> <p>No, not more than two-thirds of South America's population live in Brazil. Although Brazil is the largest country in South America by both land area and population, it does not encompass the majority of the continent's people. As of recent estimates, Brazil's population is roughly 213 million, while South America's total population is about 440 million. This means Brazil's population constitutes roughly 48% of the continent's total, which is significantly less than two-thirds. Hence, while Brazil has a large population, it does not exceed the two-thirds threshold relative to the entire continent's population.</p> <p>Sources:</p> <ol style="list-style-type: none"> 1. https://www.cia.gov/the-world-factbook/countries/brazil/ 2. https://www.macrotrends.net/global-metrics/countries/BRA/brazil/population 3. https://www.worldometers.info/world-population/south-america-population/

Figure 6.3: **Types of LLM responses used in Study 2.** We vary three variables in the LLM responses: accuracy of the LLM’s answer to the question (correct/incorrect), presence of an explanation (absent/present), and presence of clickable sources (absent/present). In total there are 8 types of responses. Here we show 4 types of responses with a correct answer to the question: “Do more than two thirds of South America’s population live in Brazil?” See Figure 6.2 for a response with an incorrect answer.

We randomized the order in which questions were presented, as well as the assignment of the 8 response types to the questions. In the final part, participants filled out an exit questionnaire about their experience with and perception of Theta, their background on LLMs, and basic demographic information. Lastly, participants were debriefed and reminded that some of the responses they saw may have contained inaccurate information.

Dependent variables

We formed a set of dependent variables (DVs) using a mix of behavioral and self-reported measures to capture participants’ reliance and accuracy, as well as related behaviors and judgments. First, we measured the *agreement* between a participant’s

answer and that of Theta; this is a commonly used behavioral measure of reliance [58, 69, 212, 237, 250, 277, 424, 437]. Second, we measured the *accuracy* of a participant’s answer to assess the task outcome. These are our main two DVs. To complement them, we also examined participants’ *confidence* and *source clicking behavior* as indirect measures of reliance, as well as *time on task*, since efficiency is also an important aspect of task outcome. These complementary measures have also been commonly studied in prior work [69, 81, 190, 193, 250, 312, 385].

Additionally, we had participants evaluate the individual LLM responses. First, we had participants evaluate the *justification quality* of a response, i.e., whether it offers a good justification for its answer. Based on prior work in psychology, we expected this to be correlated with reliance and confidence [101, 243], as well as whether participants ask follow-up questions [128, 235]. Second, we had participants evaluate the *actionability* of a response, as incorrect responses or responses with low justification quality can still be useful if they are actionable; recall that in Study 1, we observed that participants often treated an LLM response as a starting point for determining what action to take next to arrive at the correct answer. Finally, we measured whether participants wrote a follow-up question they would like to ask to Theta. This is in part a proxy for satisfaction: prior work in psychology has found that children are less likely to re-ask a question when they are satisfied with an initial response [50, 207, 272]. On the other hand, greater satisfaction with a response can increase curiosity about related content [235].

Formally, we measured the following DVs based on participants’ observed behavior:

- **Agreement:** TRUE if the participant’s final answer is the same as Theta’s answer; FALSE otherwise.
- **Accuracy:** TRUE if the participant’s final answer is correct; FALSE otherwise.

- **SourceClick:** TRUE if the participant clicked on one or more sources; FALSE otherwise.
- **Time:** Number of minutes from when the participant saw the question to when they clicked next.

We additionally measured the following DVs based on participants' self-reported ratings or selections:

- **Confidence:** Rating on the question "How confident are you in your answer?" on a 7-point scale.
- **JustificationQuality:** Rating on the statement "Theta's response offers good justification for its answer" on a 7-point scale.
- **Actionability:** Rating on the statement "Theta's response includes information that helps me determine what my final answer should be" on a 7-point scale.
- **Followup:** TRUE if the participant wrote a follow-up question they would like to ask instead of selecting "I'm satisfied with the current response and would not ask a follow-up question."

All DVs were measured once for each of the 8 tasks. See Figure 6.2 for screenshots of an example task.

Analysis

We hypothesized that the three features of LLM responses that we manipulated — the accuracy of the answer, the presence of sources, and the presence of an explanation — would affect each of the DVs. To examine this hypothesis, we used a mixed-effects regression model (logistic or linear depending on the data type), where each participant has a unique ID and each task question has a unique ID. Specifically,

for each DV except `SourceClick`, we fit the model $DV \sim AI_Correct * AI_Sources * AI_Explanation + (1|participant) + (1|question)$. For `SourceClick`, we fit the model $DV \sim AI_Correct * AI_Explanation + (1|participant) + (1|question)$ only looking at data points for which participants were provided with sources. `AI_Correct`, `AI_Sources`, and `AI_Explanation` are binary variables with `Correct Answer`, `No Sources`, and `No Explanation` as the reference levels.

We complemented the main analysis with several additional analyses. First, we conducted two pre-registered analyses exploring how participants reacted to inconsistencies in explanations (Section 6.4.3) and how participants' source clicking behavior relates to other DVs (Section 6.4.3). Analysis details and results are presented in the respective sections. Second, we conducted a thematic analysis [48, 51] of participants' free-form answers in the exit questionnaire. The results are presented in Section 6.4.2 alongside the quantitative results from the main analysis.

Materials

To simulate a realistic LLM usage scenario of users seeking answers to questions they don't know the answer to, we selected task questions according to the following criteria: (1) most lay people should not know the answer off the top of their head so that they will likely engage with the LLM response and (2) the answer can be objectively and automatically assessed. To satisfy the criteria, we first created 32 binary factual questions based on facts from the books *Weird But True Human Body* [186] and *Weird But True World 2024* [187] by National Geographic Kids. We then ran a short pilot study ($N = 50$) in which we asked participants to answer the 32 questions based on their knowledge and without consulting external sources. This allowed us to assess how commonly known the answers to the questions are in our sample. We selected questions with less than 50% accuracy (i.e., worse than random guessing) as our final set of task questions (12 in total) to satisfy our first selection

criterion. However, we acknowledge that focusing on difficult questions may affect the generalizability of our results. See Figures 6.2 and 6.3 for an example question and the supp. mat. for the full set.

To create LLM responses that are realistic and reflect the state-of-the-art, we used ChatGPT-4o with a Plus subscription and with Browsing allowed, Memory disallowed, and a new chat for each prompt. Initially, we inputted the selected task questions to ChatGPT without any system prompts. Consistent with prior work [222], we observed that ChatGPT’s responses generally follow the same structure: answer to the question (e.g., yes or no) followed by an explanation (supporting details). However, the responses greatly varied in form (e.g., the number of paragraphs and the use of bulleted or numbered lists) and length (ranging from 48 to 213 words). To reduce this variability, we used the system prompt “Provide a one paragraph response not exceeding 180 words” following the choices in prior work [222].

For each task question, we first created a pair of responses with explanations, one with a correct answer and one with an incorrect answer. To do so, we used the prompts “Why is [correct answer/incorrect answer] the correct answer to the question: [Task question]?” We note that the obtained responses were similar in nature to responses obtained by just asking the task question. We phrased the prompts this way to reduce any structural differences between responses for correct and incorrect answers. We made minor edits to improve readability and ensure all responses had the same structure (i.e., answer to the question followed by an explanation). We did not make substantive edits to the content. To create responses without explanations, we removed the explanation component from these responses.

To create responses with sources, we sent the same prompts to Perplexity AI, one of the most popular LLM-infused search engines, with a Plus subscription and with GPT-4o as the AI model. This is because none of the responses from ChatGPT-4o included sources, even with Browsing allowed. Perplexity AI’s responses included 5

to 10 sources. When we analyzed the sources, we found that all sources are real, relevant to the task question, and tended to provide accurate information, although we could not fact-check every single statement in these sources. To not overwhelm participants, we randomly selected 3 sources and appended them to the responses with/without explanations to create responses with sources. All responses from ChatGPT-4o and Perplexity AI were retrieved between July 29 and August 1 of 2024 using the latest version at the time.

After creating different types of LLM responses, we went through the responses again and coded the presence of *inconsistencies*, i.e., sets of statements that cannot be true at the same time [164], which we found to be an important unreliability cue in Study 1.⁵ We found that 3 out of 12 responses with explanations for an incorrect answer contained inconsistencies: (1) For the task question “Do more than two thirds of South America’s population live in Brazil?” the incorrect response (see Figure 6.2) says “yes” but later states Brazil’s population as around 213 million and South America’s total population as around 430 million. (2) For “Which body part has a higher percentage of water, lungs or skin?” the incorrect response says “skin” but later states skin has 64% and lungs have 83% water percentage. (3) For “Do all mammals except platypus give birth to live young?” the incorrect response says “yes” but later states echidnas as another mammal species that does not give birth to live young. In contrast, none of the 12 responses with an explanation for a correct answer contained inconsistencies. While the presence of inconsistencies is not something we control for or manipulate, we coded it to study its effects on reliance

⁵To code inconsistencies, we carefully read each LLM response and reasoned about every pair of statements (whether they can be true at the same time). This was doable because the responses are relatively short (less than 180 words) and do not require specialized knowledge to understand. For the same reasons, we expect most people to be able to notice these inconsistencies. We note that this may not always be the case. The presence of inconsistencies may have little to no effect if they are hard to detect, for example, because the LLM response is long, complex, and requires specialized knowledge to understand that two statements cannot be true at the same time.

and other measures. See Figure 6.3 for examples of different types of responses and the supp. mat. for more information on the study materials.

Participants

We aimed to collect a minimum of 300 responses post-exclusions. This number was determined based on a power analysis on pilot data using the simR package in R [141]. We conducted data collection using Qualtrics and Prolific in August 2024. Specifically, we collected responses from 320 U.S.-based adults on Prolific who had completed at least 100 prior tasks with a 95% or higher approval rating. We excluded 12 responses (3.75%) based on three pre-registered exclusion criteria (3 for response time under 5 minutes, 9 for less than 80% accuracy on the post-task attention check, and 1 for off-topic free-form response; 1 response was caught on multiple criteria). Our final sample consists of 308 responses. Regardless of inclusion or exclusion in the final sample for analysis, we paid all participants \$3.75. The median study duration was 15.3 minutes, so on average, participants were paid \$14.70 per hour. See the supp. mat. for more information about participants.

6.4.2 Study 2 results: Main analysis

We begin with the main analysis results. We report the raw data means (M) and standard deviations in Table 6.1 and the regression results (β, SE, p) in the text. We use *significance* to refer to statistical significance at the level of $p < .05$. Recall that we fit mixed-effects regression models with three variables and all possible interactions (see Section 6.4.1 for details). We did not find a significant three-way interaction for any DVs. Given our interest in the effects of explanation and sources in LLM responses, we report significant main effects and two-way interactions in the following order: main effects of explanation and interactions with LLM accuracy (Section 6.4.2), main effects of sources and interactions with LLM accuracy

Table 6.1: **Study 2 main results.** We report the raw data means (and standard deviations) for the eight types of LLM responses: {Correct ✓, Incorrect ✗ answer} \times {Neither, Explanation only, Sources only, Explanation & Sources}. See Section 6.4.2 for details.

DV	Answer	Neither	Explanation only	Sources only	Explanation & Sources
Agreement (%)	✓	67.2% (47.0%)	78.2% (41.3%)	73.4% (44.3%)	79.9% (40.2%)
	✗	78.2% (41.3%)	82.8% (37.8%)	68.2% (46.7%)	76.9% (42.2%)
Accuracy (%)	✓	67.2% (47.0%)	78.2% (41.3%)	73.4% (44.3%)	79.9% (40.2%)
	✗	21.8% (41.3%)	17.2% (37.8%)	31.8% (46.7%)	23.1% (42.2%)
Confidence (1-7)	✓	4.55 (1.68)	5.26 (1.47)	5.50 (1.58)	5.83 (1.28)
	✗	4.92 (1.73)	5.47 (1.44)	5.43 (1.51)	5.61 (1.28)
SourceClick (%)	✓	-	-	28.2% (45.1%)	25.0% (43.4%)
	✗	-	-	27.9% (44.9%)	22.1% (41.5%)
Time (min)	✓	1.05 (1.23)	1.08 (0.98)	1.24 (1.04)	1.30 (1.14)
	✗	0.89 (0.85)	1.05 (0.94)	1.39 (1.16)	1.44 (1.30)
Justification Quality (1-7)	✓	2.58 (1.89)	5.52 (1.48)	4.45 (2.13)	5.99 (1.32)
	✗	2.81 (2.09)	5.51 (1.62)	3.90 (2.17)	5.44 (1.75)
Actionability (1-7)	✓	2.56 (1.94)	5.14 (1.74)	4.90 (2.06)	6.13 (1.19)
	✗	2.91 (2.09)	5.32 (1.72)	4.59 (2.09)	5.62 (1.63)
FollowUp (%)	✓	71.4% (45.2%)	28.2% (45.1%)	34.4% (47.6%)	12.7% (33.3%)
	✗	65.3% (47.7%)	27.6% (44.8%)	47.1% (50.0%)	25.6% (43.7%)

(Section 6.4.2), interactions between explanation and sources (Section 6.4.2), and additional effects of LLM accuracy (Section 6.4.2).

Main effects of explanation and interactions with LLM accuracy

We find a significant main effect of explanation on most DVs (all except `SourceClick` and `Time`). Specifically, provided that the LLM answer is correct and there are no sources, providing an explanation leads to higher participant agreement with the LLM answer ($M = 78.2\%$ vs. 67.2% , $\beta = .60$, $SE = .19$, $p = .002$), accuracy ($M = 78.2\%$ vs. 67.2% , $\beta = .65$, $SE = .19$, $p < .001$), confidence in the final answer ($M = 5.26$ vs. 4.55 , $\beta = .74$, $SE = .10$, $p < .001$), rating of the LLM response's justification quality ($M = 5.52$ vs. 2.58 , $\beta = 2.94$, $SE = .13$, $p < .001$), and rating of its actionability ($M = 5.14$ vs. 2.56 , $\beta = 2.59$, $SE = .13$, $p < .001$). On the other hand, the likelihood

of asking a follow-up question is lower when an explanation is provided ($M = 28.2\%$ vs. 71.4% , $\beta = -2.38$, $SE = .21$, $p < .001$).

For participants' accuracy, however, we find a significant interaction between the presence of an explanation and the accuracy of the LLM answer ($\beta = -1.00$, $SE = .28$, $p < .001$). In the absence of sources, when the LLM answer is correct, participants' accuracy is higher when an explanation is provided ($M = 78.2\%$ vs. 67.2%). In contrast, when the LLM answer is incorrect, accuracy is lower when an explanation is provided ($M = 17.2\%$ vs. 21.8%). That is, in both cases, participants submitted the same answer as the LLM's more often when an explanation was provided.

We find support for these findings in the qualitative data as well. In their free-form answers in the exit questionnaire, 28 participants wrote that they submitted a different answer from the LLM's answer when there was no explanation. As put by one participant, “*One sentence answers felt incomplete and did not explain how Theta arrived at its conclusion.*” Another wrote the absence of explanation “*made the [LLM's] answer too hard to trust.*”

In summary, we find that **explanations tend to increase reliance, both appropriate reliance on correct answers and overreliance on incorrect answers**. Explanations also tend to increase participants' confidence in their answer and evaluation of the LLM response, and decrease their likelihood of asking a follow-up question. Intuitively, this suggests participants viewed LLM responses with explanations as more satisfying and reliable, regardless of their accuracy. These findings are consistent with prior research [26, 120, 299, 312, 349, 400, 437] and suggest explanations from state-of-the-art LLMs can also lead to overreliance and have unintended negative consequences.

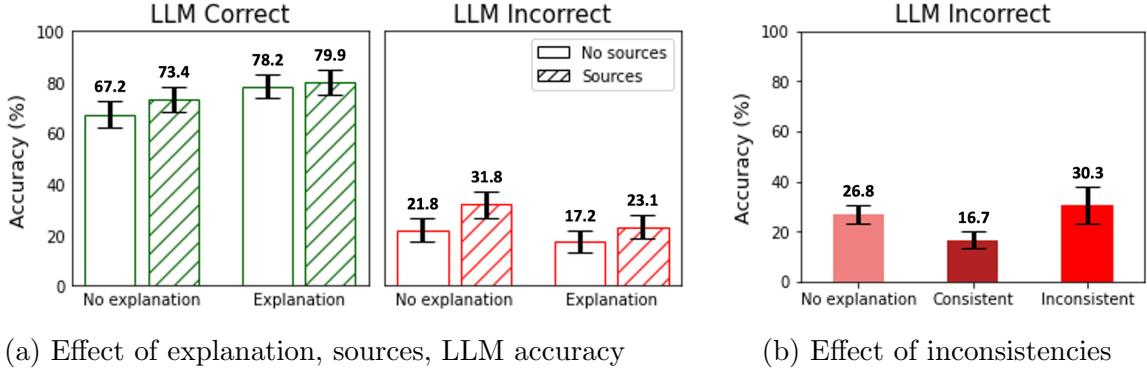


Figure 6.4: **Summary of participants' accuracy in Study 2.** We plot the raw data means and 95% confidence intervals for participants' accuracy when provided with different types of LLM responses. When the LLM's answer is correct, participants' accuracy is highest when the LLM response includes an explanation and sources (Figure 6.4a left). When the LLM's answer is incorrect, participants' accuracy is highest when the LLM response includes sources but not an explanation (Figure 6.4a right). When the LLM response includes an explanation for an incorrect answer, participants' accuracy is higher when the explanation is inconsistent (Figure 6.4b).

Main effects of sources and interactions with LLM accuracy

We find a significant main effect of sources on the time spent on the task, as well as on all self-reported variables. That is, when the LLM answer is correct and there is no explanation, providing sources leads to higher participant time on task ($M = 1.24$ min vs. 1.05 min, $\beta = .17, SE = .07, p = .027$), confidence in the final answer ($M = 5.50$ vs. 4.55, $\beta = .96, SE = .10, p < .001$), rating of the LLM response's justification quality ($M = 4.45$ vs. 2.58, $\beta = 1.88, SE = .13, p < .001$), and rating of its actionability ($M = 4.90$ vs. 2.56, $\beta = 2.34, SE = .13, p < .001$). In contrast, the likelihood of asking a follow-up question is lower when sources are provided ($M = 34.4\%$ vs. 71.4%, $\beta = -2.04, SE = .20, p < .001$).

However, we find a significant interaction between the presence of sources and LLM accuracy on many DVs. Beginning with agreement ($\beta = -.83, SE = .27, p = .002$), provided that there is no explanation, when the LLM answer is correct, agreement is higher when sources are provided ($M = 73.4\%$ vs. 67.2%). But when the LLM answer is incorrect, agreement is lower when sources are provided ($M = 68.2\%$ vs. 78.2%).

These results suggest that **sources tend to increase appropriate reliance on correct answers and reduce overreliance on incorrect answers.**

Significant interactions are also found for all self-reported variables: **Confidence** ($\beta = -.45, SE = .14, p = .002$), **JustificationQuality** ($\beta = -.79, SE = .19, p < .001$), **Actionability** ($\beta = -.65, SE = .19, p < .001$), and **Followup** ($\beta = 1.05, SE = .27, p < .001$). Provided that there is no explanation and the LLM answer is correct, providing sources increases **Confidence**, **JustificationQuality**, and **Actionability** while decreasing **Followup**. When the LLM answer is incorrect, these effects of sources are all attenuated. The fact that sources have different effect sizes for correct and incorrect LLM answers provides further (if indirect) support for the idea that sources can help foster appropriate reliance.

The final significant interaction between sources and LLM accuracy is found for time on task ($\beta = .33, SE = .11, p = .002$). Provided that there is no explanation, when the LLM answer is correct, time on task is higher when there are sources ($M = 1.24$ min vs. 1.05 min). But when the LLM answer is incorrect, this effect of sources on time on task is magnified ($M = 1.39$ min vs. $.89$ min). A possible reason for this result is that when the LLM answer is incorrect, in some instances participants may have found conflicting information between the LLM response and the sources and spent more time resolving the conflict and completing the task. For example, 45 participants wrote in their free-form responses that they submitted a different answer from the LLM's answer when it conflicted with the information in the sources, e.g., “*I trusted the information in the links more than I trusted Theta's answer. Therefore, if the information in the links differed, I submitted a final answer that was different from Theta's.*”

Participants also wrote that the mere presence of sources tended to increase the credibility of the LLM response (e.g., “*If Theta supplied sources for its answers, I felt the answers were more credible*”), while the absence of sources had the opposite effect

(e.g., “*Not having any links provided with [Theta’s] answer was a red flag to me to think something is wrong or can’t be found*”). 29 participants explicitly stated that they submitted a different answer from the LLM’s answer when there were no sources in the LLM response. Additionally, several participants wrote about how they were forced to rely on their intuition when there were no sources, e.g., “*Without being able to verify info, my gut was my best answer.*” They expressed frustration about this and said they would prefer to have sources since it is “*easiest to agree or disagree when the AI cited its sources.*”

Interactions between explanation and sources

In addition to the main effects of explanation and sources and their respective interactions with LLM accuracy, we find a significant interaction between explanation and sources for all self-reported variables: **Confidence** ($\beta = -.42, SE = .14, p = .004$), **JustificationQuality** ($\beta = -1.41, SE = .19, p < .001$), **Actionability** ($\beta = -1.36, SE = .19, p < .001$), and **Followup** ($\beta = .81, SE = .31, p < .001$). Provided that the LLM answer is correct, when there are no sources, providing explanations increases **Confidence** ($M = 5.26$ vs. 4.55), **JustificationQuality** ($M = 5.52$ vs. 2.58), and **Actionability** ($M = 5.14$ vs. 2.56), while decreasing **Followup** ($M = 28.2\%$ vs. 71.4%). When there are sources, however, providing explanations still increases **Confidence** ($M = 5.83$ vs. 5.50), **JustificationQuality** ($M = 5.99$ vs. 4.45), and **Actionability** ($M = 6.13$ vs. 4.90), while decreasing **Followup** ($M = 12.7\%$ vs. 34.4%), but all to a lesser extent than when there are no sources. In short, including both explanation and sources achieves the biggest effects in these measures, though their joint effects are subadditive, i.e., less than the sum of the individual effects.

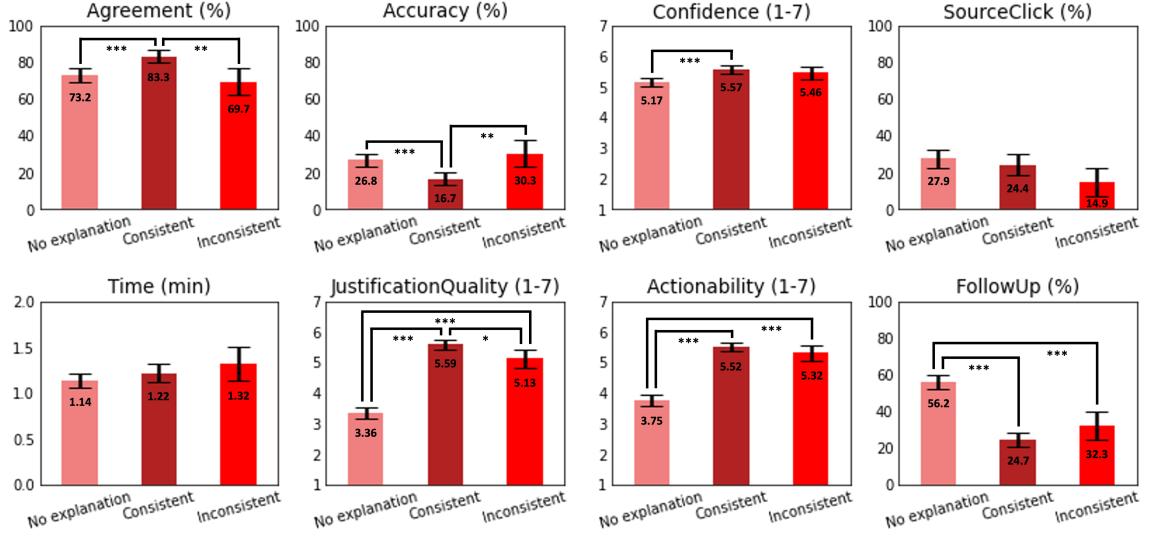


Figure 6.5: Study 2 results on inconsistencies. We plot the raw data means and 95% confidence intervals. Brackets indicate statistically significant differences between three types of incorrect LLM responses: No explanation, Consistent explanation, and Inconsistent explanation. Significance is marked as * ($p < .05$), ** ($p < .01$), and *** ($p < .001$). See Section 6.4.3 for details.

Additional effects of LLM accuracy

Finally, we find a significant main effect of LLM accuracy on many DVs, in addition to its interactions with explanation and sources reported above. Provided that there are no sources or explanation, when the LLM answer is incorrect compared to correct, agreement is higher ($M = 78.2\%$ vs. 67.2% , $\beta = .60$, $SE = .19$, $p = .002$), confidence is higher ($M = 4.92$ vs. 4.55 , $\beta = .37$, $SE = .10$, $p < .001$), and Actionability is higher ($M = 2.91$ vs. 2.56 , $\beta = .35$, $SE = .13$, $p = .007$), while accuracy is lower ($M = 21.8\%$ vs. 67.2% , $\beta = -2.07$, $SE = .19$, $p < .001$) and time on task is lower ($M = .89$ min vs. 1.05 min, $\beta = -.17$, $SE = .08$, $p = .025$). These results suggest that participants found incorrect answers more plausible than correct answers for the task questions used in the experiment. This is not surprising as we deliberately selected challenging questions, i.e., questions with less than 50% human accuracy in our pilot study. As such, this is likely a feature of our stimulus materials rather than a generalizable finding.

6.4.3 Study 2 results: Additional analyses

Finally, we report results from our additional pre-registered analyses on the effects of inconsistencies in explanations (Section 6.4.3) and the relationship between participants' source clicking behavior and other DVs (Section 6.4.3).

Inconsistencies in explanations

In Study 1, we found inconsistencies in explanations to be an important unreliability cue that participants often noticed. While the presence of inconsistencies is not something we control for or manipulate, we explore whether and how the natural inconsistencies that arose in LLM responses have effects on the DVs with a pre-registered analysis. Specifically, we used analysis of variance (ANOVA) to compare the means of DVs across three types of incorrect LLM responses: No explanation ($N = 616$), consistent explanation ($N = 461$), and inconsistent explanation ($N = 155$), where N indicates the number of instances for which participants received a given response type. If there was a significant difference across response types, we conducted pairwise comparisons with post-hoc Tukey tests. We only analyze responses with an incorrect answer, as none of the responses with a correct answer contained inconsistencies (as described in Section 6.4.1). We present the results in Figure 6.5.

For most DVs (all except `SourceClick` and `Time`) we find a significant difference across response types. Overall, overreliance on incorrect answers is most prevalent when participants receive consistent explanations, as evidenced by the highest agreement with the LLM answer, confidence in their final answer, and ratings of justification quality and actionability, as well as the lowest accuracy and likelihood of asking follow-up questions. In comparison, when participants receive inconsistent explanations, agreement is significantly lower ($M = 69.7\%$ vs. 83.3% , $p = .002$), rating of the LLM response's justification quality is significantly lower ($M = 5.13$ vs. 5.59 , $p = .028$), while accuracy is significantly higher ($M = 30.3\%$ vs. 16.7% , $p = .002$).

While our study materials did not allow us to investigate the effect of inconsistencies when an LLM answer is correct, which may happen less naturally based on our observations, these results suggest that **inconsistencies can help reduce overreliance on incorrect answers induced by explanations.**

Consistent with the quantitative results, 19 participants stated in their free-form answers that they disagreed with Theta when “*Theta’s responses were contradictory.*” For example, several participants wrote about how Theta provided a logically inconsistent response to the question “Do more than two thirds of South America’s population live in Brazil?” (See Figure 6.2 for the response.) As one participant elaborated, “*The Brazilian and South American population answer contradicted itself. Two-thirds would imply 66% but given the number of Brazilians compared to the total population of South America given in the answer, the actual percentage is closer to 50%.*”

Source clicking behavior

From our main analysis (Section 6.4.2), we did not find any significant effect on when participants chose to click on the provided source links. We only found a marginally significant main effect of explanation such that participants’ source clicking likelihood is lower when there is an explanation than not ($M = 25.0\%$ vs. 28.2% , $\beta = -.62$, $SE = .36$, $p = .086$). However, there is high variance across individuals. According to our tracking, 189 out of 308 participants never clicked on sources, 33 participants clicked on sources in one task, 18 participants in two tasks, 23 participants in three tasks, and 45 participants in all four tasks for which sources were provided.

To better understand participants’ source clicking behavior, we conducted a pre-registered analysis to examine its relationship with other DVs. Specifically, we used ANOVA to compare the means of DVs between instances in which participants were provided sources but did not click on any ($N = 914$) and instances in which participants were provided sources and clicked on one or more ($N = 318$). Among

the latter, 164 are instances in which the LLM answer is correct and 154 are instances in which the LLM answer is incorrect.

We find that when participants click on sources, accuracy is higher ($M = 60.1\%$ vs. 49.2% , $p < .001$) and time on task is higher ($M = 2.11$ min vs. 1.08 min, $p < .001$), while rating of the LLM response's justification quality is lower (4.58 vs. 5.08 , $p < .001$). The accuracy and time on task results are intuitive. The sources in our study stimuli tended to provide accurate and relevant information (see Section 6.4.1 for details), so when participants clicked on sources, they likely found correct answers at the expense of spending more time on task. Indeed, we see that source clicking was helpful when the LLM gave an incorrect answer. The increase in accuracy is bigger when the LLM answer is incorrect ($M = 37.0\%$ vs. 24.2%) than when the LLM answer is correct ($M = 81.7\%$ vs. 74.8%). For reference, when LLM responses do not include sources, participants' answer accuracy is $M = 19.5\%$ when the LLM answer is incorrect and $M = 72.7\%$ when the LLM answer is correct.

There are multiple possible factors that might influence the finding that the rating of justification quality is lower when participants have clicked on the provided sources. Participants may have clicked on sources because they found the LLM response's justification quality to be low, or their rating may have decreased after examining the sources. Again breaking down the data into instances in which the LLM answer is correct and those where it is incorrect, participants' rating of the response's justification quality when they clicked on sources vs. not is $M = 5.04$ vs. 5.29 when the LLM answer is correct and $M = 4.08$ vs. 4.87 when the LLM answer is incorrect.

Together, these results suggest engaging with the content of (accurate and relevant) sources can be an effective way of improving decision outcomes. However, the presence of explanation may reduce users' natural tendency to examine sources, especially when they find the explanation to be of high quality. It could be helpful

to nudge users to pay more attention to sources by highlighting sources or placing sources above explanations.

6.5 Discussion

6.5.1 Implications of findings

Explanations

In our studies, we found that explanations play an important role in shaping users' reliance. In Study 1, we gained qualitative insights on how participants interpreted and used explanations to judge the reliability of LLM answers. In Study 2, we examined the effects of the presence of an explanation, as well as its interaction with other variables, and found that explanations increase reliance on both correct and incorrect responses. This is consistent with prior findings in HCI that explanations can increase overreliance [26, 120, 312, 400, 437], including explanations generated by LLMs [299, 349]. It is also consistent with prior work in psychology, which finds that explanations are often found compelling even when they contain little content [133, 217] or content that experts judge irrelevant [158], and that effects of superficial cues on explanation quality are more severe when time and prior knowledge are limited [159, 184]. In the absence of effort and expertise, users will inevitably rely on superficial cues to explanation quality, such as fluency [375], a characteristic that LLM explanations typically possess in spades. This suggests a potential tension in providing LLM explanations to lay users: the properties that make such explanations intelligible and compelling may be precisely those that lead to overreliance. As such, we encourage LLM explanations to be evaluated and optimized for appropriate reliance, in addition to other qualities such as fluency, justification quality, and satisfaction.

Sources

Our results offer some basis for optimism, however: sources helped reduce overreliance on incorrect answers and increase appropriate reliance on correct answers. One possibility is that sources encouraged participants to engage in slow and careful System 2 thinking, instead of quick and automatic System 1 thinking [175, 176]. In our study, participants spent significantly more time on task when provided with sources, especially when the LLM’s answer was incorrect. The qualitative data also supports this. Many participants wrote that they checked sources. Many also wrote that they submitted a different answer from the LLM’s answer when it conflicted with the information in the sources. We emphasize, however, that the sources provided in Study 2 were all real and tended to provide accurate and relevant information. This is not always the case. Recent work has found that popular LLM-infused applications frequently generate statements that are not supported by sources [238] and sometimes even generate fake sources [11]. If the provided sources are junk or just broken links, then presumably they will not help foster appropriate reliance. They could potentially even hurt by making the LLM response look more trustworthy, similar to how flawed and meaningless explanations have been found to increase people’s trust and reliance [112, 182, 334]. In addition to improving the quality of sources in LLM responses, future work should explore different issues with sources (e.g., fake, unreliable, conflicting sources and inaccurate summaries of sources), design choices (e.g., location of sources and amount of preview), and their effects on people’s perceptions and behaviors.

When it comes to choosing between providing sources only and providing sources and an explanation, there are benefits and drawbacks to each. When the LLM answer is incorrect, participants’ accuracy is highest on responses with sources only ($M = 31.8\%$), followed by responses with explanation and sources ($M = 23.1\%$), neither ($M = 21.8\%$), and explanation only ($M = 17.2\%$) — suggesting that providing sources only is most effective at reducing overreliance on incorrect answers. However,

it is not as effective at improving appropriate reliance when the LLM answer is correct. Here, participants' accuracy is highest on responses with explanation and sources ($M = 79.9\%$), followed by responses with explanation only ($M = 78.2\%$), sources only ($M = 73.4\%$), and neither ($M = 67.2\%$). In contexts where LLMs have much higher accuracy than users, providing sources only can lead to lower overall accuracy than providing sources and an explanation. Further, participants rated responses with sources only lower in terms of justification quality and actionability, compared to responses with sources and explanation, suggesting that people prefer the latter.

Inconsistencies and other unreliability cues

Finally, we found that LLM responses contain new forms of unreliability cues. Prior research, in particular the work by Chen et al. [75], found that people identify AI models' biases, inability to consider contexts or multiple features, and lower performance on rare instances as cues of unreliability. In our studies, we identified other cues such as inconsistencies in explanations and lack of explanation or sources — all of which are related to the particular characteristics of LLMs. For example, some inconsistencies occur due to the stochastic nature of LLMs: LLMs can generate different responses for the same input, unlike deterministic AI models. Even within a single response, inconsistencies occur because LLMs are not trained to generate only logically consistent statements. The other cues are connected to LLMs' natural language modality and ability to handle a wide variety of tasks, which lead to responses with much more diverse features and forms compared to classical AI models with fixed output spaces.

Intriguingly, we found positive effects of such unreliability cues when it comes to reducing overreliance. In Study 1, participants who noticed unreliability cues engaged with the LLM responses more thoroughly. In Study 2, participants relied less on incorrect LLM responses when they were provided with explanations containing inconsistencies than those without. These findings, along with prior findings on other

unreliability cues (e.g., inconsistencies between multiple responses [222, 349]), suggest that guiding people’s attention to these cues can be an effective approach to reducing overreliance. For example, we could apply computational methods to automatically detect inconsistencies (e.g., [94, 208]) then use highlighting to draw people’s attention to the detected inconsistencies. Other interventions (e.g., expressing uncertainty, displaying disclaimers, and encouraging source checking) could be applied jointly for cases in which inconsistency detection is difficult or where LLM responses are consistently inaccurate. We suggest future research to explore more thoroughly what unreliability cues exist in LLM responses and how to design interventions that help people notice and reason about these cues.

6.5.2 Explanation of the answer vs. Explanation of how the LLM arrived at the Aaswer

Throughout the work, we have used the term *explanation* to refer to supporting details in LLM responses that justify the LLM’s answer to the input question. This is different from how the term is often used within the explainable AI community in that we do not make any assumptions about the extent to which it faithfully describes the way that the model arrived at its answer. We emphasize that faithfulness is extremely difficult for users — or even model developers — to evaluate, especially without access to the model’s internals. Evaluating the faithfulness of model explanations is an active area of research [8, 21, 166, 404, 439].

Nevertheless, many participants in Study 1 interpreted ChatGPT’s responses as including somewhat faithful explanations of how the system arrived at its answer, especially when the responses had certain characteristics (P3, P6, P7, P8, P10, P11, P14, P15, P16). For some, the critical characteristic was the presence of sources (P6, P16). As P16 (high-knowledge, low-use) described, “*I would think of the citation itself as an explanation because it kind of implies ‘I’m giving you this information*

because it came from this source' and then me as a human can evaluate that source."

For others, it was the step-by-step form of responses, which are common for math questions (P3, P7, P8, P14). After seeing them, P14 (low-knowledge, high-use) said, "*I think it's very clear how did it [ChatGPT] provides me the answer.*"

In contrast, three participants, all with high knowledge of LLMs, were strongly opposed to the idea that ChatGPT could provide explanations of how it arrived at its answers (P5, P12, P13). P5 (high-knowledge, low-use) stated that "*it's provably false that ChatGPT's responses provide a description of how it arrives at its answers,*" emphasizing that ChatGPT's responses are "*definitely and empirically not explanations because there's no reflection in the model.*" Similarly, P12 (high-knowledge, high-use) said they don't think of ChatGPT as explaining anything to them, and that ChatGPT was just "*trained to provide answers that look like an explanation because that's what we would find most useful.*" P4 (high-knowledge, low-use) shared this view and emphasized that "*there's no way to interpret how the answer came from.*" They noted that the explanations ChatGPT offers describe "*how a normal person would reach the answer,*" and are not explanations of how ChatGPT arrives at its answers.

In sum, while there was considerable variability between individuals, we found that many participants, especially those without much knowledge of LLMs, viewed ChatGPT's responses as including somewhat faithful explanations for how the system arrived at its answer. This raises a concern because first, again, there is no reliable way for users or anyone to evaluate their faithfulness without access to the system's internals, and second, recent work has found explanations from LLMs are often not faithful to their process [251, 256, 378, 411, 439] and can easily be manipulated, e.g., to rationalize incorrect information [59, 300, 429]. Such assumptions can be strengthened by the increasing anthropomorphization of LLMs and lead to inappropriate reliance [85, 342, 405]. We suggest future research to explore strategies for improving people's understanding of LLMs [14, 246] and study how they are connected to reliance behaviors.

6.5.3 Limitations

There are several limitations of our work that are worth reflecting on. First and foremost, our studies were conducted in the context of objective question-answering and may not generalize to other contexts of LLM use (e.g., writing, idea generation, and task automation). We encourage the community to conduct more empirical studies on how user reliance is shaped in various contexts.

Each of our studies has a different set of strengths and limitations. Study 1 was a think-aloud study that offered descriptive examples of how users interpret and act upon different LLM response features in a relatively natural setting. However, prior work has pointed out that the set-up of a think-aloud study can cause people to behave differently than they would otherwise [46, 126, 280]. For example, we saw a much higher rate of source clicking in Study 1 ($M = 63.8\%$) than in Study 2 ($M = 25.8\%$) which was an online experiment. We also emphasize that the LLM response features identified in Study 1 are not comprehensive. We suggest future work to explore what other features influence users' reliance and can help them succeed in tasks despite inaccuracies from LLMs.

In Study 2, we employed a different research method (a controlled experiment), prioritizing the generalizability of findings by controlling as many other variables as possible. For example, in the experiment, participants saw exactly one response from Theta, created in advance using the state-of-the-art LLM-infused applications ChatGPT and Perplexity AI, instead of interacting with a real system in multiple rounds. While participants referred to Theta as "AI" or "LLM" in the exit questionnaire (e.g., "*I just trusted the AI when I didn't know the answer already*"), we did not measure participants' general perceptions of Theta or inquire about their experience. Hence, it is more accurate to view Study 2 as a study of people's perceptions and behaviors around specific LLM responses rather than a study of people's interactions with LLMs. While showing one controlled LLM response is a commonly used method

(e.g., [193, 222, 349]), people’s perceptions and behaviors may change over time, meriting further studies in more interactive settings.

Additionally, we set Theta’s accuracy to be 50% which is significantly worse than the state-of-the-art. While this choice allowed us to compare the effects of LLM response features on relying on correct vs. incorrect answers in a balanced fashion, future work should explore whether there are interaction effects between these features and the LLM’s accuracy. There are also implications of our experimental task, which was answering difficult factual questions (that less than 50% of pilot study participants knew the answer to). We chose this task to simulate realistic scenarios of people seeking answers to questions they don’t know the answer to. However, it is possible that our findings may not generalize to tasks where people have sufficient prior knowledge and can more deeply engage with the content of the LLM responses. Finally, there are many LLM response features that we did not study or control for (e.g., simplicity of explanations [242], quality of sources [321, 402], and presence of jargon [91]). We encourage future work to explore different features and methods to understand user interactions with LLMs, an emerging research area whose importance will only grow with time.

6.6 Conclusion

We conducted two empirical studies to understand how different features of LLM responses shape users’ reliance. We found that the presence of explanations increases reliance on both correct and incorrect responses. However, we observed less reliance on incorrect responses when sources are provided or when explanations exhibit inconsistencies. Our findings highlight the importance of evaluating LLM response features with users before deployment. Our findings also suggest that providing (accurate and relevant) sources and designing interventions that help users notice and reason about

inconsistencies and other unreliability cues in explanations can be promising directions for fostering appropriate reliance on LLMs.

Acknowledgements

We foremost thank the participants for sharing their time and experiences. We also thank the members of the Princeton Visual AI Lab, the Princeton HCI Lab, and the Princeton Concepts & Cognition Lab, as well as the anonymous reviewers for thoughtful feedback and discussion. We acknowledge support from the NSF Graduate Research Fellowship Program (SK) and the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award (OR).

Chapter 7

Conclusion

This dissertation presented a series of human-centered evaluation studies conducted to advance responsible development and deployment of AI technologies. First, it demonstrated how to introduce *human* evaluation into areas dominated by automatic evaluation such as explainable AI, and showed how insights from human evaluation can re-orient the research and design of AI technologies (Chapter 2). Next, it illustrated the importance of *context* in evaluation by presenting insights from contextualized studies with real users that cannot be gained from automatic evaluations or non-contextualized studies with hypothetical users (Chapters 3 and 4). Finally, it presented careful studies of human-AI *interaction* that shed light on how users perceive and act upon different features of AI technologies and contributed actionable recommendations for fostering appropriate user trust and reliance on AI (Chapters 5 and 6).

In this chapter, I conclude the dissertation with a discussion of key implications, practical recommendations, and future directions for advancing responsible AI.

7.1 Key implications

Always evaluate AI technologies before release. Evaluation is critical for assessing effectiveness, identifying strengths and weaknesses, and making informed de-

cisions about AI technologies (e.g., models, systems, and features). There have been many instances where evaluation revisited assumptions and uncovered blind spots in AI research and development. For example, our evaluation in Chapter 2 raised a warning that responsible AI efforts, such as providing an explanation of an AI output to help people understand and make better decisions about the output, may backfire and have unintended negative effects (e.g., make an incorrect AI output seem correct). Next, our evaluation in Chapter 5 revealed that language choices matter: even seemingly subtle choices such as the perspective in which uncertainty is expressed in AI outputs (e.g., first-person vs. general perspective) shaped how people perceive and act upon AI outputs. To minimize unexpected effects of AI technologies, evaluation should be a priority and not an afterthought in AI research and development.

Consider what different evaluation methods can capture and choose appropriately. How should evaluation be done? This is the million dollar question as there is no one-size-fits-all method for evaluation. Different methods have different strengths and weaknesses and suit different purposes and contexts. The method should be chosen based on the goal of the evaluation (what one hopes to learn) and available resources (money, time, effort, and expertise). Poor choice of method and over-extrapolation of results can misguide AI research and development. Many studies, including those presented in Chapter 2, have revealed that automatic evaluation metrics, proxy tasks, and subjective measures can be misleading in evaluating AI technologies, especially their usefulness to users [57, 160, 190]. In another example (Chapter 5), we found a tension between self-report and behavioral measures in evaluating LLM responses: LLM responses with explanations were rated higher (measured via self-report) but also increased overreliance (measured behaviorally), suggesting use of both types of measures and a careful balance of optimization objectives (e.g., user preference vs. appropriate reliance). Overall, embracing diverse evaluation methods will help choose the appropriate method for each situation.

7.2 Practical recommendations

Conducting human-centered evaluation is important, but how we do it matters as much. A well-conducted evaluation can provide valuable insights, but a poorly conducted one can be misleading [63]. Here I share a set of practical recommendations for future researchers based on lessons I learned from my dissertation work.

Clearly state definitions and explain operationalizations of constructs.

Evaluation of AI technologies' performance and impact in human-AI interaction often involves measuring complex constructs. One example of such construct that has gained a lot of traction in recent years is *trust*. However, recent surveys of trust in AI research [379, 386] note that many empirical work do not state a definition of trust or do not explain why they chose a specific operationalization. This makes it difficult for researchers to assess the validity of the work (e.g., whether there is a mismatch between the theoretical understanding of the construct and the operationalization [165]) and to accurately interpret and compare the results (e.g., whether a proposed approach is more effective than another approach studied in a different work). Clearly stating and explaining the chosen definition and operationalization will reduce unnecessary confusion and increase the rigor of the field.

Carefully design tasks and instructions. Designing doable tasks and providing clear instructions to human participants (evaluators) are challenging and important endeavors. Prior research has found that poorly designed tasks and unclear instructions hurt not only the results quality but also the relationship between requesters and workers in crowdsourced studies [196, 264, 290, 415]. Tasks (including instructions) should be designed in consideration of participants' backgrounds such as AI expertise and language and cultural context. In general, tasks should be clear (i.e., easy to understand and do) and concise (i.e., only contain information necessary for completing the task). Concrete recommendations include: (1) using everyday non-technical terms (e.g., “photo” instead of “image” and “developed to recognize

scenes” instead of “trained on a dataset to perform scene classification”); (2) simplifying technical content if necessary (e.g., in Chapter 2, one AI explanation we wanted to evaluate was a complex decision tree with 511 decision nodes and up to 10 decision steps which were too overwhelming for participants. To make the evaluation tractable, we significantly simplified the AI explanation and had participants evaluate only the final two decision steps of the explanation); and (3) designing introduction materials (e.g., in the same example, we helped participants understand decision trees by walking through a simple decision tree and presenting warm up exercises). Overall, well-designed tasks and instructions will lead to higher-quality results.

Employ sufficient data quality checks when using crowdsourcing. In AI research and development, crowdsourcing platforms (e.g., Amazon Mechanical Turk and Prolific) are commonly used for creating datasets, evaluating models, and running user studies as they provide access to a large pool of workers who can complete tasks and participate in studies. Crowdsourcing also has the benefit of being time- and cost-effective than traditional in-person recruitment methods. However, recent work suggests that LLMs are widely used in crowd work, which could threaten the validity of human judgment and behavior data obtained from crowdsourcing platforms [388]. It is more important than ever to employ sufficient data quality checks to filter out data that is likely compromised (e.g., produced using LLMs or from bots). In my experience (Chapters 5 and 6), the following data quality checks (exclusion criteria) have been useful : (1) answers to honeypot questions that are invisible to people but visible to bots; (2) same answers to all questions or tasks; (3) too short response time; (4) failed attention checks, especially questions about the main task; and (5) free-form answers that are clearly off-topic (e.g., answers are irrelevant, gibberish, spam) or identical across multiple responses. Using a combination of these checks and setting appropriate qualification criteria (e.g., minimum approval rating and number of completed tasks/studies on the platform) may help obtain quality human judgment

and behavior data from crowdsourcing platforms. See the work by Bylinskii et al. [63] and Schuff et al. [337] for additional practical recommendations.

7.3 Future research directions

There is much more that needs to be done for advancing responsible AI. In closing, I highlight several future research directions for the community.

Evaluate AI systems in interaction contexts. While AI systems are developed to interact with people, they are still primarily evaluated on static benchmarks. This creates a mismatch: AI systems that excel on these benchmarks often fail to meet user needs in real-world interactions. To reduce this mismatch, we need to evaluate AI systems in context and in interactive settings. Concretely, future research could pursue: designing evaluation metrics that capture user needs and contexts; developing new tools for collecting and analyzing rich human-AI interaction data; and partnering with organizations to conduct evaluations with real users. Together, these efforts will bring AI evaluation closer to real-world interaction contexts.

Incorporate participatory approaches to AI development and evaluation. Given the broad impact of AI in society, we need people with comparably broad sets of experiences and perspectives shaping its development. There are many benefits to broadening participation. For example, in Chapter 3, we observed users pointing out blind spots in state-of-the-art explainable AI methods and proposing solutions. Building on recent successes in participatory AI *design* [423, 431], we can innovate new ways for participatory AI *development* and *evaluation*. For example, we can explore creating fine-tuning datasets and designing evaluation metrics with users and those impacted by AI to incorporate their perspectives, while carefully considering the benefits and harms from the participation.

Increase public understanding of AI. Responsible AI is not only about technological advancements. Public excitement and fear around AI are currently high, and they are often fueled by misconceptions. Increasing public understanding of AI is crucial for responsible adoption and use of AI. It would be important to identify common misconceptions people have about AI, develop learning materials that correct these misconceptions, and study how people's understanding of AI shape their perceptions and interactions, building on the AI literacy and learning sciences literature [246, 363]. This line of work will provide implications for education and policy on how to teach and talk about widely-deployed AI technologies.

Overall, responsible AI efforts must take place at various levels. They should also be interdisciplinary as AI systems are complex and continuously evolving sociotechnical systems. Responsible AI efforts would be most effective when they fully engage with the complexity and the messiness of real-world human-AI interactions by integrating perspectives from AI, HCI, and other fields.

Appendix A

Supplementary material for Chapter 2

A.1 Details on the evaluation tasks

Agreement task. For each image, we show one model prediction-explanation pair and ask the participants how confident they are in the model’s prediction. We show 10 images in total (5 correct, 5 incorrect predictions in random order). Participants rate their confidence in the given prediction on a 4-point scale (1: confident prediction is incorrect, 2: somewhat confident prediction is incorrect, 3: somewhat confident prediction is correct, 4: confident prediction is correct).

Distinction task. For each image, we show four model prediction-explanation pairs for it (in random order) and ask the participants to identify the correct prediction based on the explanations. For GradCAM [339] and BagNet [52], participants are tasked with 10 sample images (5 correct and 5 incorrect predictions), each of which is shown with four heatmaps. On correctly predicted samples, the four heatmaps correspond to the top-4 predicted classes. On incorrectly predicted ones, we show heatmaps for the top-3 predicted classes and the heatmap of the ground-truth class.

For ProtoPNet [72], we show four correctly predicted samples in total. Each sample is presented with four explanations corresponding to the top-4 predicted classes. We reduce the total number of samples and focus on correctly predicted samples due to the complexity of the ProtoPNet explanations; even with this change, the ProtoPNet study duration is twice as long as that of GradCAM and BagNet. For ProtoTree [285], we show 10 correctly predicted samples in total and ask participants to select the correct decisions on the two final nodes which lead to four (2^2) different predictions.

Additionally for ProtoPNet [72] and ProtoTree [285], we ask participants to rate the similarity of prototype-region pairs in both tasks using a 4-point Likert scale (1: not similar, 2: somewhat not similar, 3: somewhat similar, 4: similar).

A.2 Details on the evaluated interpretability methods

GradCAM [339]. For our ImageNet [324] studies, we generate GradCAM explanations for the ResNet50 [148] model in the `torchvision` library which achieves 76.1% classification accuracy. For our CUB studies, we generate GradCAM explanations for a ResNet50 [148] model we trained on the CUB [391] training set. This model achieves 81.0% accuracy on the CUB test set. We used the code by Gildenblat et al. [134] to generate GradCAM visualizations.¹ For the *agreement* task, we generate the GradCAM heatmap for the model prediction and normalize it into the [0, 1] range. For the *distinction* task, we generate four GradCAM heatmaps for each image: for correct predictions, we generate heatmaps for the top-4 predictions; for incorrect predictions, we generate heatmaps for the top-3 predictions and for the ground-truth class. We identify the local minimum and maximum of the four heatmaps, and then normalize the heatmaps into the [0, 1] range. This way, we preserve the intensity

¹<https://github.com/jacobgil/pytorch-grad-cam>

difference between heatmaps for different predictions. See Figure A.1 for an example set of GradCAM explanations.

BagNet [52]. For our ImageNet studies, we use the BagNet33 model trained by the original authors which achieves 66.7% accuracy on ImageNet classification. For our CUB studies, we train a BagNet33 model on the CUB training set. This model achieves 74.2% accuracy on the CUB test set. For the *agreement* task, we use the authors' code as is and normalize each heatmap individually by clipping the values above the 99th percentile.² On the other hand, for the *distinction* task, we normalize the four heatmaps together so that we preserve the intensity difference. See Figure A.2 for an example set of BagNet explanations.

ProtoPNet [72]. For ProtoPNet, we used the ResNet34-based model trained by Hoffmann et al. [154]. We pruned 331 prototypes from this model to improve interpretability. The resulting model has 1669 prototypes and achieves 79.9% accuracy on the CUB [391] test set. For generating explanations, we used the code by the original authors with some modifications which we describe below.³ In our studies, given an explanation, participants are asked to rate the similarity of each prototype-region pair, then either rate the level of confidence in the prediction's correctness (*agreement*) or select the correct class (*distinction*). To make ProtoPNet's explanations more suitable for these tasks, we made the following modifications to the original explanation form.

²<https://github.com/wielandbrendel/bag-of-local-features-models>

³<https://github.com/cfchen-duke/ProtoPNet>

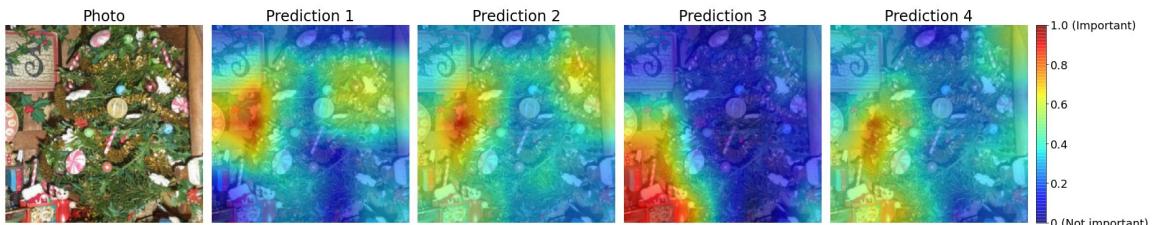


Figure A.1: **GradCAM explanations shown in the *distinction* task.**

- The ProtoPNet model calculates evidence for all classes using the learned prototypes, then predicts the class with the highest evidence. However, we deemed it is unrealistic to ask users to review explanations for all 200 bird classes in CUB. Hence, we only present explanations for one (*agreement*) or four (*distinction*) classes and ask users to examine them.
- The original explanation (Figure A.3 left) shows activation maps, similarity scores, class connection weights, and the total class evidence. In our version (Figure A.3 right) we remove them as we seek to investigate what participants rate as similar and not.
- In the original explanation, prototypes are presented in the order of highest to lowest similarity. In ours, we randomly shuffle the order of prototypes because we don't want to skew the participants' region-prototype similarity ratings.

ProtoTree [285]. For ProtoTree, we used the model trained by the original authors which achieves 81.7% accuracy on the CUB [391] test set. This model is a pruned tree of depth 10 and 511 nodes. We used the authors' code to generate explanations with some modifications we describe below.⁴

- Same as what we did for ProtoPNet explanations, we removed the similarity scores as we seek to investigate what participants rate as similar and not.
- For the local explanation, we converted the horizontal explanation (Figure A.4) into a vertical one (Figure A.5). A vertical explanation is a better representation of how the model reasons, as the model starts from the root node and proceeds down the tree until it reaches one of the bottom leaves. Further, it is easier for the participants to examine the explanation by scrolling up and down.

⁴<https://github.com/M-Nauta/ProtoTree>

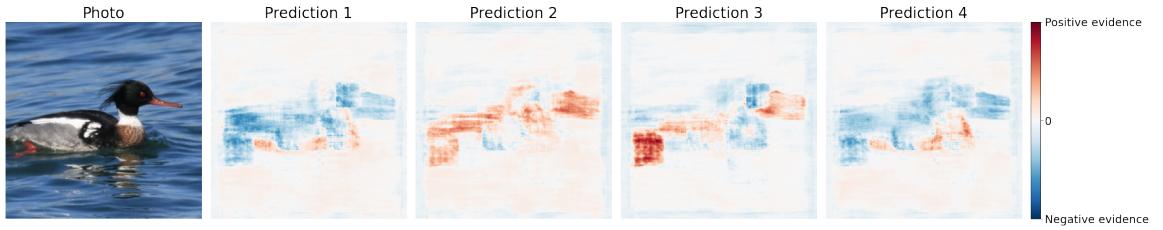


Figure A.2: **BagNet explanations shown in the *distinction* task.**

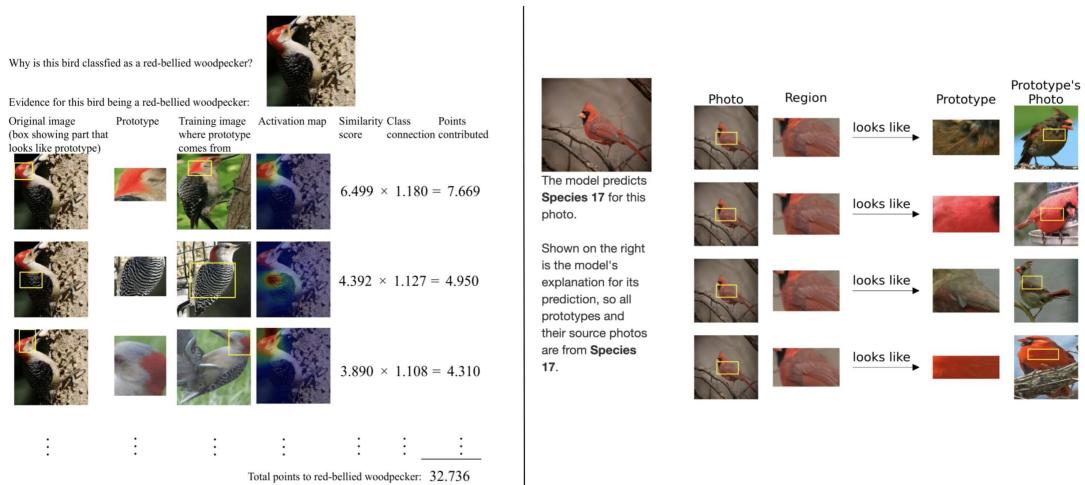


Figure A.3: **ProtoPNet original and modified explanations.** The original explanation (left) taken from the original paper [72] contains details such as activation maps, similarity scores, and class connection weights. In our version (right), we remove these to abstract away the complexities and have the participants focus on examining the similarity between prototypes and their matched image regions.

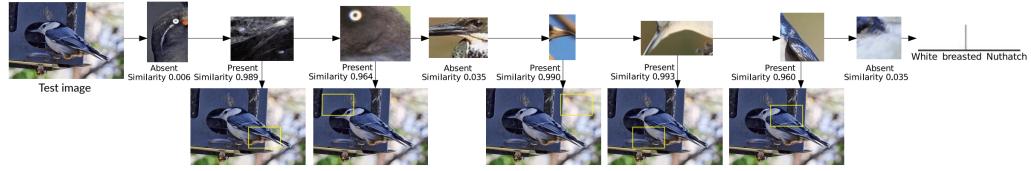


Figure A.4: **ProtoTree original explanation.** We show the original explanation displayed in Fig. 9 of the original paper [285]. See Figure A.5 for our modified explanation.



Figure A.5: **ProtoTree modified explanation.** See Figure A.4 for the original explanation.

A.3 Details on the human studies

We ran our study through Human Intelligence Tasks (HITs) deployed on Amazon Mechanical Turk (AMT). We recruited participants who are US-based, have done over 1000 HITs, and have prior approval rate of at least 98%. For each study, we deployed 10 HITs, each with a different set of input images and explanations. To reduce the variance with respect to the input, we had 5 participants complete each HIT, so each study had 50 participants. Participants were compensated based on the state-level minimum wage of \$12/hr.

The demographic distribution was: man 60%, woman 38%, non-binary 1%, no gender reported 1%; White 74%, Black/African American 9%, Asian 7%, no race/ethnicity reported 7%, Hispanic/Latino/Spanish Origin of any race 2%, American Indian/Alaska Native 1%, Native Hawaiian/Other Pacific Islander 0%. The self-reported machine learning experience was 2.5 ± 1.0 , between “2: have heard about...” and “3: know the basics...” The average study duration was 6.9 ± 3.5 minutes for GradCAM, 6.6 ± 3.5 for BagNet, 13.6 ± 6.2 for ProtoPNet, and 10.4 ± 3.1 for ProtoTree.

A.4 Additional results and analyses

Agreement study results with vs. without example images

In Chapter 2, we described the results of our *agreement* study. Here we provide additional results.

For GradCAM and BagNet, we run another version of the *agreement* study where we show three example images from the predicted class, in addition to the test image, prediction, and heatmap (see Figure A.6). Since ProtoPNet and ProtoTree explanations consist of source images of the learned prototypes, we take this measure to

provide similar supplementary information for GradCAM and BagNet. As expected, participants improve on the task when they see example images from the predicted class (5.3% overall improvement for GradCAM, 7.1% for BagNet). However, even with the help of example images, participants tend to believe in incorrect predictions, which suggests that incorrect top-1 predictions from high-performance models such as ResNet50 and BagNet are oftentimes convincing. Between CUB and ImageNet, task accuracy is overall higher on CUB, but both yield similar insights. See Table A.1 for full results.

Analysis with automatic evaluation metrics

In Chapter 2, we briefly summarized our analysis with automatic evaluation metrics. Here we discuss the results in more detail.

We further analyzed GradCAM heatmaps set using three automatic evaluation metrics: pointing game [432], energy-based pointing game (energy game) [396],

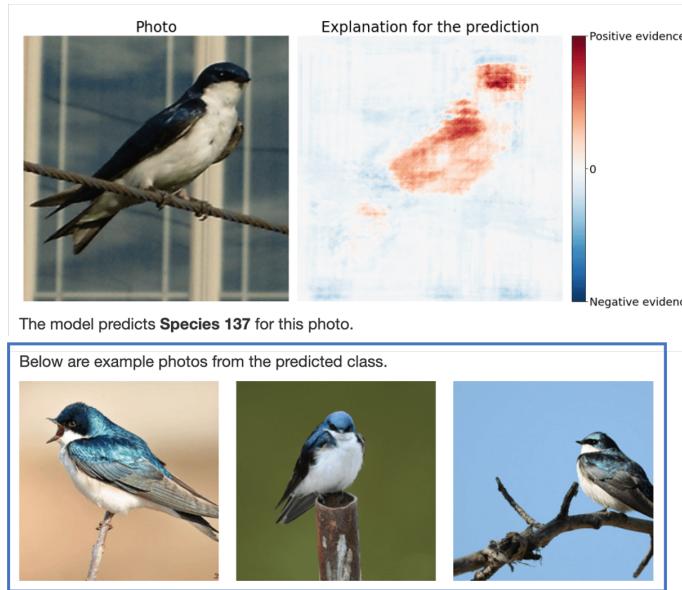


Figure A.6: **BagNet *agreement* study input with example images.** For the study version with example images, we additionally show three example images from the predicted class (highlighted in the blue box).

Table A.1: **Agreement study results with vs. without examples.** For each study, we show mean accuracy, standard deviation of the participants’ performance, and mean confidence rating in parentheses. *Italics* denotes methods with accuracy not statistically significantly different from 50% random chance ($p > 0.05$); **bold** denotes the highest performing method in each group. In all studies, participants leaned towards believing that model predictions are correct when provided explanations, regardless of if they are actually correct. For example, for GradCAM on CUB, participants thought 72.4% of correct predictions were correct and $100 - 32.8 = 67.2\%$ of incorrect predictions were correct. These results reveal an issue of *confirmation bias*. Comparing results with vs. without example images from the predicted class, participants improve on the task when they see examples, but still tend to believe in incorrect predictions. See Appendix A.4 for a discussion.

CUB	GradCAM [339]	w/ examples	BagNet [52]	w/ examples
Correct	$72.4\% \pm 21.5$ (2.9)	$83.2\% \pm 15.7$ (3.3)	$75.6\% \pm 23.4$ (3.0)	$83.6\% \pm 17.3$ (3.3)
Incorrect	$32.8\% \pm 24.3$ (2.8)	$36.8\% \pm 22.8$ (2.8)	$42.4\% \pm 28.7$ (2.7)	$44.4\% \pm 30.5$ (2.6)
ImageNet	GradCAM [339]	w/ examples	BagNet [52]	w/ examples
Correct	$70.8\% \pm 26.6$ (2.9)	$78.4\% \pm 25.6$ (3.2)	$66.0\% \pm 27.2$ (2.8)	$77.2\% \pm 23.3$ (3.2)
Incorrect	$44.8\% \pm 31.6$ (2.7)	$43.6\% \pm 32.4$ (2.6)	$35.6\% \pm 26.9$ (2.7)	$42.8\% \pm 32.7$ (2.6)

and intersection-over-union (IoU) [441]. Pointing game considers a heatmap correct when its highest-intensity point lies inside the segmentation/bounding-box annotation. Energy game calculates how much energy in a heatmap falls inside the segmentation/bounding-box annotation. IoU captures the amount of overlap between a binarized heatmap (according to some threshold) and the segmentation/bounding-box annotation. For all three metrics, higher values indicate better localization quality.

We evaluate up to three GradCAM explanations per image, all using the same segmentation/bounding-box annotation for the ground-truth class: heatmaps for the ground-truth class, predicted class, and class with the second-highest score. Results are summarized in Table A.2. For CUB heatmaps, we calculate the three metrics on the entire test set (top table). For ImageNet heatmaps, we calculate the metrics on 5,000 randomly sampled validation images. Since ImageNet images sometimes have multiple bounding box annotations, we report results evaluated with one bounding box that yields the best result (middle table) and results evaluated with the union of bounding boxes (bottom table). We find that all three metrics are highest on the

ground-truth/predicted class heatmaps for correctly predicted samples. However, we find that these metrics are also high for other heatmaps, even when they are for wrong classes.

Next, we calculate these metrics on images/heatmaps we showed the participants and analyze our human study results. In the *agreement* study, we find near-zero correlation between participants' confidence in the model prediction and localization quality of heatmaps. In the *distinction* study, we also do not see meaningful relationships between participants' choices and these automatic metrics, possibly because all four heatmaps have similar localization quality. These observations are consistent with the findings of [117, 287], i.e., automatic metrics poorly correlate with human performance in post-hoc attribution heatmap evaluation. Overall, our analysis reveals a limitation of automatic metrics.

Similarity judgment of humans vs. prototype-based models

In Chapter 2, we quantified the gap between prototype-based models and human users' notion of similarity. Here we show a plot of participant vs. ProtoPNet prototype similarity rating (Figure A.7). There is no significant negative correlation between the two. This result suggests a gap between ProtoPNet and human judgments of similarity.

Nonetheless, we find that participants are consistent in their similarity ratings and decisions. When examining ProtoPNet and ProtoTree explanations, on average participants assign higher similarity ratings to prototypes of the class they select to be correct (2.9 out of 4 for both ProtoPNet *agreement* and *distinction* tasks, 2.4 for ProtoTree *agreement*) and lower similarity ratings to prototypes of the class they select to be incorrect (2.0 and 2.1 for ProtoPNet *agreement* and *distinction*, 2.0 for ProtoTree *agreement*). The similarity ratings between the two groups are statistically significantly different in all studies. This suggests that participants understand how

Table A.2: **Evaluation of GradCAM heatmaps using automatic metrics.** We report the mean and standard deviation of three automatic evaluation metrics calculated on heatmaps for the ground-truth class, predicted class, and class with the second-highest score. All three metrics are highest on the ground-truth/predicted class heatmaps for correctly predicted samples. However, these metrics are also high for other heatmaps, even when they are for wrong classes.

CUB [391] heatmaps evaluated with the segmentation mask				
Prediction	Class	Pointing game [432]	Energy game [396]	IoU [441]
Correct	GT/Predicted	0.92 ± 0.27	0.12 ± 0.07	0.38 ± 0.15
	Second-highest	0.74 ± 0.44	0.09 ± 0.06	0.24 ± 0.15
Incorrect	GT	0.73 ± 0.45	0.08 ± 0.06	0.23 ± 0.16
	Predicted	0.83 ± 0.37	0.09 ± 0.06	0.29 ± 0.15
	Second-highest	0.80 ± 0.40	0.09 ± 0.06	0.26 ± 0.15
ImageNet [324] heatmaps evaluated with the bounding box that yields the best result				
Prediction	Class	Pointing game [432]	Energy game [396]	IoU [441]
Correct	GT/Predicted	0.95 ± 0.22	0.27 ± 0.13	0.60 ± 0.28
	Second-highest	0.93 ± 0.26	0.26 ± 0.13	0.60 ± 0.27
Incorrect	GT	0.91 ± 0.29	0.23 ± 0.14	0.52 ± 0.31
	Predicted	0.82 ± 0.38	0.22 ± 0.15	0.52 ± 0.33
	Second-highest	0.84 ± 0.37	0.23 ± 0.15	0.52 ± 0.33
ImageNet [324] heatmaps evaluated with the union of the bounding boxes				
Prediction	Class	Pointing game [432]	Energy game [396]	IoU [441]
Correct	GT/Predicted	0.95 ± 0.22	0.29 ± 0.13	0.65 ± 0.26
	Second-highest	0.93 ± 0.26	0.28 ± 0.13	0.64 ± 0.26
Incorrect	GT	0.91 ± 0.29	0.24 ± 0.14	0.56 ± 0.30
	Predicted	0.82 ± 0.38	0.24 ± 0.15	0.56 ± 0.32
	Second-highest	0.84 ± 0.37	0.24 ± 0.15	0.56 ± 0.32

the model reasons (i.e., they predict the bird class whose prototypes appear most similar to the given photo).

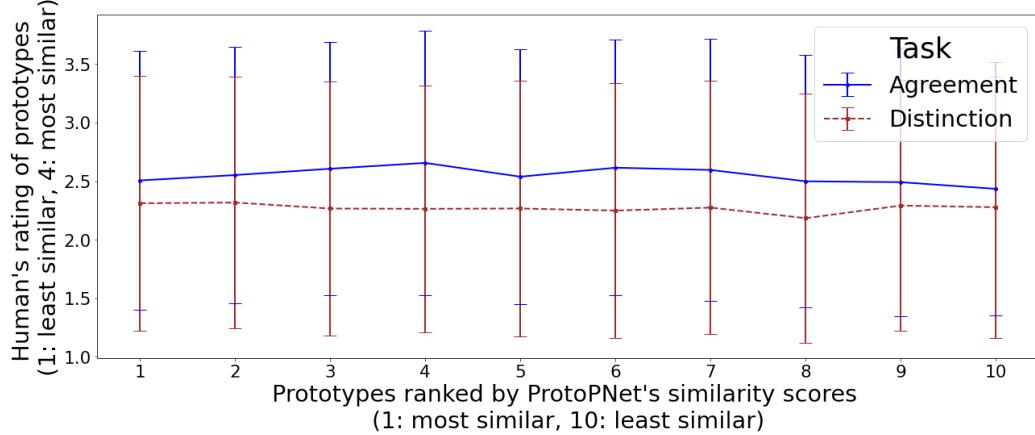


Figure A.7: **Participant vs. ProtoPNet prototype similarity rating.** There exists a gap between ProtoPNet’s similarity scores and human judgments of similarity (Spearman’s $\rho = -0.25, p = 0.49$ for *distinction*; $\rho = -0.52, p = 0.12$ for *agreement*).

Subjective evaluation results

In Chapter 2, we summarized our subjective evaluation results. Here we provide the full results.

In Table A.3, we report the participants’ self-rated level of understanding of the given model’s reasoning process. Overall, the participants rated their level of understanding between 3 (fair) and 4 (good). Interestingly, we find that the rating tends to decrease after the participants see their task performance. Several participants indicated that their performance was lower than what they expected: “I thought I would do a bit better!”, “my score wasn’t as high as I would have liked”, “I was surprised that my score was not very much higher than random guessing. I thought I had a good idea of the model, especially making judgements about the amount of positive and negative evidence, but it seems I did not.” No one suggested the opposite. This trend suggests that participants might have been disappointed in their task performance, which in turn led them to lower their self-rated level of method understanding.

Table A.3: **Participants’ self-rated level of method understanding.** We report the mean and standard deviation of the participants’ self-rating of their method understanding. Participants provide ratings three times: after reading about the method (post-intro), after completing the task (post-task), and after learning about their task performance (post-results). The rating tends to *decrease* after the participants see their task performance ($p < 0.05$).

Dataset	Method	Study	Post-intro	Post-task	Post-results
CUB [391]	GradCAM [339]	Agreement	3.7 ± 0.9	3.8 ± 0.9	3.3 ± 1.1
		Agreement w/ examples	3.7 ± 1.0	3.9 ± 0.7	3.4 ± 1.0
		Distinction	3.4 ± 1.0	3.5 ± 1.2	3.6 ± 0.8
	BagNet [52]	Agreement	3.5 ± 1.0	3.7 ± 0.8	3.3 ± 1.1
		Agreement w/ examples	3.7 ± 0.8	3.9 ± 0.8	3.6 ± 1.0
		Distinction	3.8 ± 0.7	4.0 ± 0.8	3.9 ± 0.8
	ProtoPNet [72]	Agreement	3.9 ± 0.8	4.0 ± 0.8	3.7 ± 0.8
		Distinction	4.1 ± 0.8	3.9 ± 0.8	3.7 ± 1.1
		Agreement	3.7 ± 0.8	3.7 ± 1.0	3.4 ± 0.8
	ProtoTree [285]	Agreement (tree)	3.7 ± 0.7	3.5 ± 0.9	3.2 ± 1.1
		Distinction	3.4 ± 1.0	3.6 ± 1.1	3.3 ± 1.2
		Agreement	3.7 ± 0.9	3.9 ± 0.9	3.0 ± 1.0
ImageNet [324]	GradCAM [339]	Agreement w/ examples	3.4 ± 0.8	3.7 ± 0.8	3.5 ± 0.9
		Distinction	3.9 ± 0.9	3.7 ± 1.0	3.7 ± 1.0
		Distinction w/ labels	3.9 ± 0.9	3.8 ± 1.0	3.8 ± 0.9
		Mean across all studies	3.7 ± 0.9	3.8 ± 0.9	3.5 ± 1.0
	BagNet [52]	Agreement	3.7 ± 0.8	3.9 ± 0.7	3.4 ± 1.0
		Agreement w/ examples	3.8 ± 0.9	3.9 ± 0.9	3.3 ± 1.0
		Distinction	3.9 ± 0.8	3.9 ± 0.8	3.8 ± 1.0
		Distinction w/ labels	3.8 ± 0.9	4.0 ± 0.8	3.8 ± 0.8

Interpretability-accuracy tradeoff results

In Chapter 2, we summarized our interpretability-accuracy tradeoff study results. Here we provide more details.

In Table A.4 and Figure A.8, we show the full results of our interpretability-accuracy tradeoff study. We report the accuracy of the evaluated interpretable model and the minimum accuracy of a baseline model that participants require in order to use it over the model with explanations under different risk settings. Across all studies, we find that participants require the baseline model to have higher accuracy than the evaluated interpretable model, and input a higher accuracy requirement for higher-risk settings. On average, participants require the baseline model to have +6.2% higher accuracy for low-risk (e.g., bird species recognition for scientific or educational purposes), +8.2% for medium-risk (e.g., object recognition for automatic grocery checkout), and +10.9% for high-risk (e.g., scene understanding for autonomous driving) settings.

We observe this trend in the participants' written responses as well. Most participants write that they would use the baseline model only when it has higher accuracy than the evaluated interpretable model: "I would need the black box model to give me a nice boost in accuracy, or I would just stick to the bagne model, since it is pretty accurate." On the contrary, participants exhibit different levels of desire for interpretability. Some deem interpretability as important: "Understanding how a prediction works is important. For me to accept a model with no explanations, the level of accuracy needs to be higher", "I prefer to understand how models work, so the black box model has to be significantly better than the other model for me to use it. As the stakes become higher, I want its accuracy to be higher because there's no way for me to question or check its progress if it's wrong." Other participants willingly tradeoff interpretability for accuracy: "I don't need to know how it works. So, as long as it's marginally better, it should be used", "I don't care about not having an

Table A.4: **Interpretability-accuracy tradeoff results.** We report the mean and standard deviation of the additional accuracy participants require for the baseline model, to use it over the model with explanations. For example in the GradCAM *agreement* study with CUB, participants require the baseline model to have +5.6% accuracy beyond the model that comes with GradCAM explanations and achieves 81.0% accuracy, in the low-risk setting. See Figure A.8 for a visualization of the results.

Dataset	Method	Study	Orig	Low-risk	Med-risk	High-risk
CUB [391]	GradCAM [339]	Agreement	81.0	+5.6 (± 6.9)	+6.2 (± 5.7)	+7.7 (± 7.0)
		Agreement w/ examples		+4.2 (± 6.1)	+5.7 (± 5.7)	+7.7 (± 7.5)
		Distinction		+2.9 (± 6.9)	+4.5 (± 5.2)	+8.1 (± 6.9)
	BagNet [52]	Agreement	74.2	+6.8 (± 7.9)	+7.8 (± 8.1)	+12.3 (± 9.2)
		Agreement w/ examples		+6.1 (± 7.1)	+8.1 (± 6.3)	+10.7 (± 9.2)
		Distinction		+7.0 (± 8.1)	+8.8 (± 7.4)	+8.4 (± 8.4)
	ProtoPNet [72]	Agreement	79.9	+5.8 (± 6.6)	+7.8 (± 4.9)	+9.4 (± 6.6)
		Distinction		+4.1 (± 7.9)	+6.1 (± 6.4)	+9.7 (± 7.1)
	ProtoTree [285]	Agreement	81.7	+3.8 (± 6.5)	+4.2 (± 6.3)	+5.1 (± 6.5)
		Agreement (tree)		+3.7 (± 5.5)	+5.8 (± 5.1)	+6.7 (± 6.6)
		Distinction		+5.1 (± 5.7)	+6.4 (± 5.8)	+9.2 (± 6.2)
ImageNet [324]	GradCAM [339]	Agreement	76.1	+6.3 (± 7.6)	+8.1 (± 8.6)	+11.8 (± 10.7)
		Agreement w/ examples		+4.8 (± 6.8)	+8.6 (± 7.6)	+11.4 (± 10.8)
		Distinction		+5.3 (± 7.3)	+9.8 (± 6.7)	+12.4 (± 8.6)
		Distinction w/ labels		+7.6 (± 7.7)	+9.3 (± 7.9)	+13.2 (± 9.0)
	BagNet [52]	Agreement	66.7	+9.9 (± 7.5)	+14.1 (± 9.5)	+17.5 (± 11.1)
		Agreement w/ examples		+9.7 (± 8.5)	+13.2 (± 10.4)	+17.6 (± 13.0)
		Distinction		+7.9 (± 9.3)	+9.6 (± 9.2)	+11.2 (± 11.2)
		Distinction w/ labels		+11.4 (± 9.2)	+12.4 (± 10.4)	+16.6 (± 11.6)
<i>Mean across all studies</i>				+6.2 (± 7.7)	+8.2 (± 7.9)	+10.9 (± 9.7)

explanation, so if the accuracy of a different model has just a 1% improvement in performance then I would choose the better performing model.”

Nonetheless most participants express a need for higher-accuracy models in higher-risk settings: “The higher the risk, the more accurate I need it to be in order to feel confident using it”, “If I were to choose to use a model that did not provide reasoning for me to utilize in evaluating how the decision was made I would need to know that the model would give me significantly better results, especially in a high-risk scenario as described above, but even in the medium risk setting, being able to assess the reasoning of the model is an invaluable tool and I would only be willing to give it up for significant increases in accuracy.”

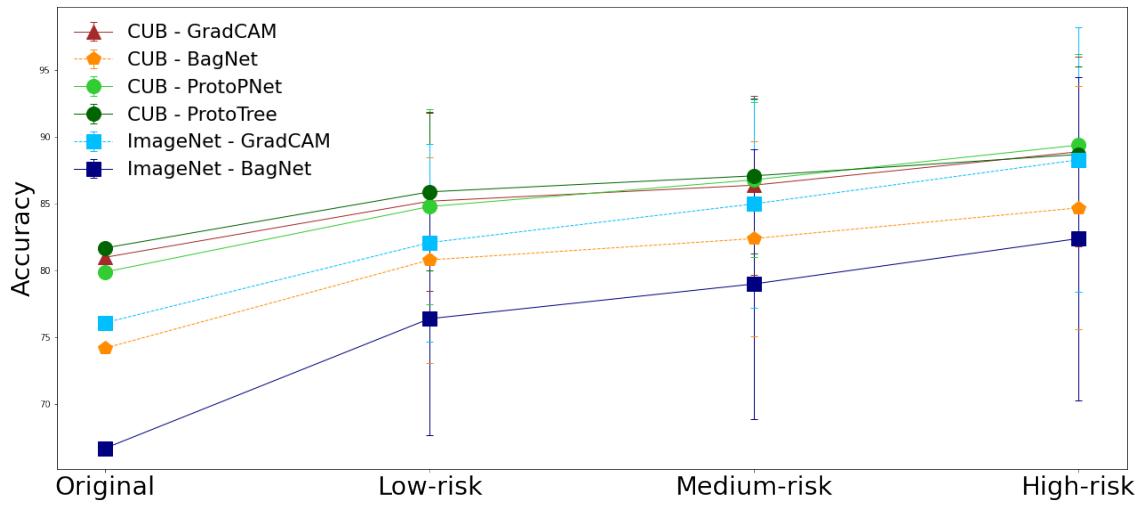


Figure A.8: Visualization of the interpretability-accuracy tradeoff results. This plot shows that participants desire higher accuracies for the baseline model, especially in higher-risk settings. See Table A.4 for the full results.

A.5 Simple decision tree used for explaining ProtoTree

One additional challenge of evaluating the ProtoTree model is that participants may not be familiar with decision trees. To mitigate this challenge, we introduce a simple decision tree model for fruit classification before introducing ProtoTree. This simple decision tree model takes in an input image and makes an output classification (Class A, B, C, D, E) based on three decision nodes. We first walk through the participants through an example. We then present two warm-up exercises so that the participants can become more familiar with decision trees. When the participants submit their answers, we also provide the correct answer and the reason for it. Participants achieved 86.5% performance on this task, implying that the low task accuracy for ProtoTree is not due to a lack of comprehension of decision trees. See Figure A.9 for the UI.

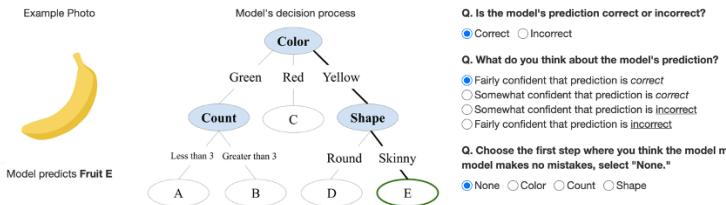
Warming up

This page is meant to give you a sense of the model and the task we will introduce in this study.

Here we have an example model that classifies fruit photos into Fruit A, B, C, D, E based on a series of decisions regarding Color, Count, and Shape. Specifically, this model makes decisions in a tree structure as you can see below.

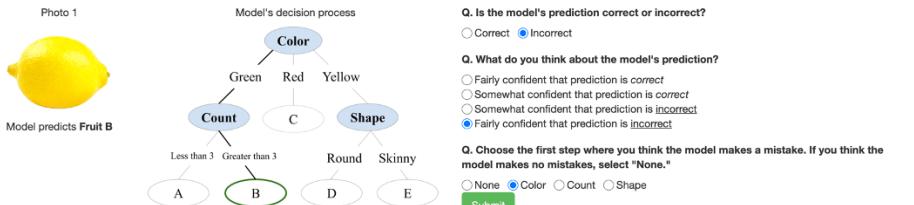
Example

For this example photo, the model reasons in the following way: The model first judges that the photo's fruit has **Color** "Yellow." Based on this decision, it moves onto the next step and judges that the fruit has **Shape** "Skinny." After these two decisions, the model arrives at its prediction: **Fruit E**.



Your Turn!

After examining the photo and the model's decision process, please answer the three questions as above, then click "Submit."



Correct! You have successfully identified that the model's decision is incorrect and that the model made a mistake on the Color decision.

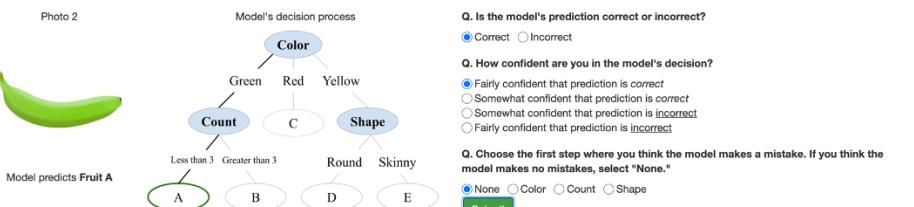


Figure A.9: A simple decision example. We use this model to introduce participants to decision trees before explaining the more complex ProtoTree. See Appendix A.5 for details.

A.6 UI snapshots

In Chapter 2, we outlined our study design. Here we provide snapshots of our study UIs in the following order.

1. Study introduction. For each participant, we first briefly introduce the study and receive their informed consent. The consent form was approved by the IRB and acknowledges that participation is voluntary, refusal to participate will involve no penalty or loss of benefits, etc. See Figure A.10.

2. Demographics and background. To help future researchers calibrate our results and do proper comparison, we request optional demographic data regarding gender identity, race and ethnicity. We also ask the participant’s experience with machine learning. See Figure A.11.

3. Method introduction. We introduce each interpretability method/model in simple terms. See Figure A.12.

4. Task preview and first subjective evaluation. To encourage participants to carefully read the method explanation, we show a preview of the task they will complete along with a correct and incorrect prediction. Participants then answer their first subjective evaluation question. In Figure A.13 we shown an example from the ProtoPNet *agreement* study.

5. Task. Participants then proceed onto the main task. We show the UI for the following 8 studies:

- GradCAM *distinction* (Figure A.14)
- GradCAM *agreement* (Figure A.15)
- Bagnet *distinction* (Figure A.16)
- Bagnet *agreement* (Figure A.17)
- ProtoPNet *distinction* (Figure A.18)

- ProtoPNet *agreement* (Figure A.19)
- ProtoTree *distinction* (Figure A.20)
- ProtoTree *agreement* (Figure A.21)

6. Second and third subjective evaluation. After the task, participants complete their second subjective evaluation question. We then disclose their task performance and ask the third subjective evaluation question. These questions allow us to investigate if the participants’ self-rated level of method understanding undergoes any changes throughout the study. See Figure A.22.

7. Interpretability-accuracy tradeoff. Finally, we investigate the trade-off participants are willing to make when comparing the evaluated interpretable model against a baseline model that doesn’t come with any explanation. We present three scenarios to the participants: low-risk (e.g., scientific or educational purposes), medium-risk (e.g., object recognition for automatic grocery checkout), and high-risk (e.g., scene understanding for self-driving cars). We then ask them to input the minimum accuracy of a baseline model that would convince them to use the baseline model over the model that comes with explanations and briefly describe their reasoning. See Figure A.23.

Study introduction

In this study, we aim to evaluate the interpretability of computer vision models. We will provide explanations of how a model makes its prediction and ask you to evaluate how interpretable it is through several questions and tasks. The expected duration of the study is 5-15 minutes.

Consent

Please read the consent form. If you understand and consent to these terms, click "I Accept" to continue.

I Accept

[Next Page](#)

Figure A.10: 1. Study introduction.

Demographics and background

Q. Demographics (Optional)

Gender identity

- Man
- Non-binary
- Woman
- Prefer to self-describe below

Race and ethnicity (select one or more)

- American Indian or Alaska Native
- Asian
- Black or African American
- Native Hawaiian or Other Pacific Islander
- White
- Hispanic or Latino or Spanish Origin of any race

Q. How much experience do you have with machine learning (ML)?

- I don't know anything about ML
- I have heard about a few ML concepts or applications
- I know the basics of ML and can hold a short conversation about it
- I have taken a course on ML and/or have experience working with a ML system
- I often use and study ML in my life

[Next Page](#)

Figure A.11: 2. Demographics and background.

Model introduction

We have a model that recognizes 1000 objects in photos. We have access to its prediction and an explanation for it. Please carefully read this page as the remaining study depends on your understanding of the model.

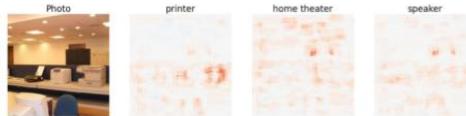
BagNet

The BagNet model recognizes an object based on small regions. This allows us to analyze how each region influences the model's prediction. For each class (e.g., *parachute*, *hummingbird*, *printer*), BagNet looks at all regions in the photo and creates a heatmap for each class where higher values, shown as darker red, mean higher evidence for the class. BagNet then sums the values in the heatmap and chooses the class with the highest value as its prediction.

For the first photo of a *parachute*, the model correctly classifies it as a *parachute*. You can see that model predicts *parachute* since its heatmap is the most red. On the other hand, the heatmaps of *hummingbird* and *military plane* are less red, which means that the model found less evidence for these classes.



For the second photo of a *printer*, the model incorrectly classifies it as a *home theater*. You can see that model predicts *home theater* since its heatmap is the most red. The heatmaps reveal that the model found relatively less evidence for *printer* and *speaker*.



Q. How well do you think you understand the model's reasoning process?

Very Poor Poor Fair Good Very Good

[Next Page](#)

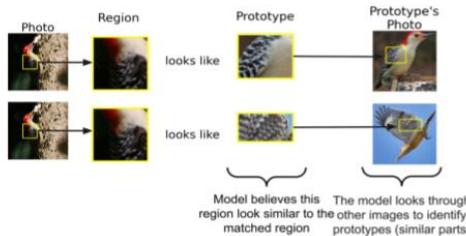
Model introduction

We have a model that recognizes 200 bird species in photos. We have access to its prediction and an explanation for it. Please carefully read this page as the remaining study depends on your understanding of the model.

ProtoPNet

The ProtoPNet model reasons with prototypes which are typical representations of a feature. For example, *pink skinny long legs* is a prototype for *Flamingos*.

Given a new bird photo, the model predicts the species based on prototypes it has learned from previously seen photos. For each prototype, the model finds a region in the photo that looks the most similar and rates its similarity. The model then predicts the bird species whose prototypes are the most similar to the photo. In the below example, the model predicts *Woodpecker* out of 200 bird species.



[Next Page](#)

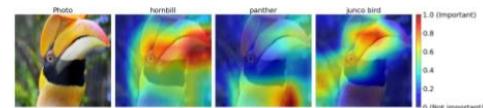
Method introduction

We have a model that recognizes 1000 objects in photos. We have access to its prediction and an explanation for it. Please carefully read this page as the remaining study depends on your understanding of the model.

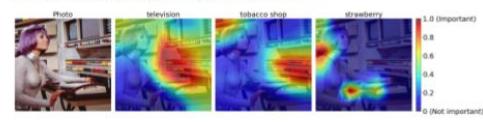
GradCAM

GradCAM is a popular visualization technique that highlights important regions in a photo for a given model to recognize a certain object (e.g., *chain*, *horseshoe*). In a GradCAM explanation, red regions corresponds to important regions (see the colorbar for more information). For each photo below, we show GradCAM explanations for three classes.

For the first photo of a *hornbill*, which is a type of a bird, the model correctly classifies it as a *hornbill*. GradCAM identifies the tip of the beak as the important region for this classification (left heatmap). On the other hand, different regions are identified as important for the model to incorrectly classify this photo as a *panther* or a *juncos like*.



For the second photo of a *television*, the model incorrectly classifies this photo as a *tobacco shop*. According to GradCAM, the important regions for this classification are the colors that switch under the television and above the person's head (middle heatmap). Again, different regions are identified as important for the model to correctly classify this photo as a *television* or incorrectly classify it as a *tobacco shop*.



Q. How well do you think you understand the model's reasoning process?

Very Poor Poor Fair Good Very Good

[Next Page](#)

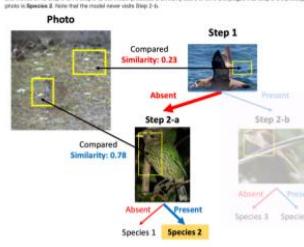
Model introduction

We have a model that recognizes 200 bird species in photos. We have access to its prediction and an explanation for it. Please carefully read this page as the remaining study depends on your understanding of the model.

ProtoTree

The ProtoTree model reasons with prototypes which are typical representations of a feature. For example, *pink skinny long legs* is a prototype for *Flamingos*.

The ProtoTree model makes decisions in steps following a tree structure. Each step in the model contains a prototype learned from a previously seen photo. At each step, the model compares a given photo to the step's prototype and rates their similarity with a score between 0 and 1. If the similarity score is above 0.5, the model judges that the prototype is present in the photo and absent in the other photos. Below we show an example ProtoTree model. For the given photo, the model produces a similarity score of 0.23 at Step 1. Hence, the model judges that Step 1's prototype is *absent* in this photo, and moves on to Step 2-a. At Step 2-a, the model produces a similarity score of 0.78 and judges that Step 2-a's prototype is *present* in this photo. The model predicts that the bird in the given photo is *Species 2*. Notice that the model never visits Step 2-b.



Below is the full structure of the model. At each step, we show the step's prototype (left) and the photo where the prototype is from (right). If the display is too small, you can download the PDF here: [View PDF](#)

[Next Page](#)

Figure A.12: **3. Method introduction.** BagNet (top left), GradCAM (top right), ProtoPNet (bottom left), ProtoTree (bottom right).

Figure A.13: 4. Task preview and first subjective evaluation.

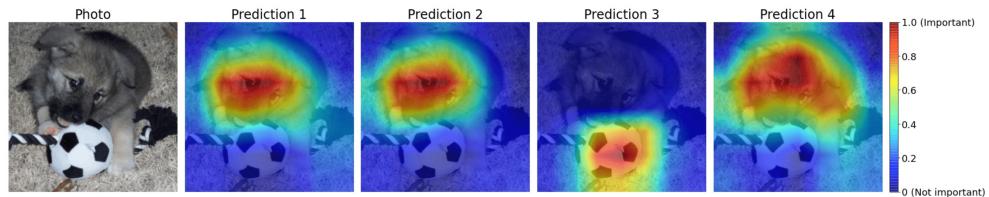
Examine model predictions

For each photo, we show explanations for the model's 4 predictions.

First, select the class you think the model predicts (i.e. gives the highest score). Second, select the class you think is correct. The two classes can be different because the model makes incorrect predictions on some photos.

For either question, random guessing will get you 25% accuracy. You will receive a reward based on your performance beyond this 25% random chance.

This is a photo of **Norwegian elkhound, elkhound**.



Q. Which class do you think the model predicts?
 1 2 3 4

Q. Which class do you think is correct?
 1 2 3 4

Q. How confident are you in your answer?

- Not confident at all
- Slightly confident
- Somewhat confident
- Fairly confident
- Completely confident

Q. How confident are you in your answer?

- Not confident at all
- Slightly confident
- Somewhat confident
- Fairly confident
- Completely confident

Click "Next Photo" after answering all questions.

1 / 10

[Next Photo](#)

Click on "Next Page" after selecting answers for all 10 photos.

[Next Page](#)

Click "Method Description" to open or close method description.

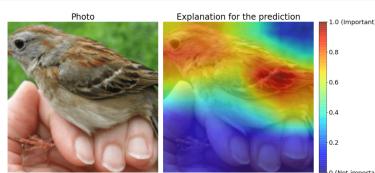
[Method Description](#)

Figure A.14: 5. Task: **GradCAM distinction**.

Examine model predictions

For each photo, the model predicts which of the 200 species the bird in the photo belongs to. Next to the photo, we show an explanation of the model prediction.

After examining the explanation, rate your confidence in the model's prediction.



Q. What do you think about the model's prediction?

- Fairly confident that prediction is correct
- Somewhat confident that prediction is correct
- Somewhat confident that prediction is incorrect
- Fairly confident that prediction is incorrect

Click "Next Photo" after answering all questions.

1 / 10

[Next Photo](#)

Click on "Next Page" after selecting answers for all 10 photos.

[Next Page](#)

Click "Method Description" to open or close method description.

[Method Description](#)

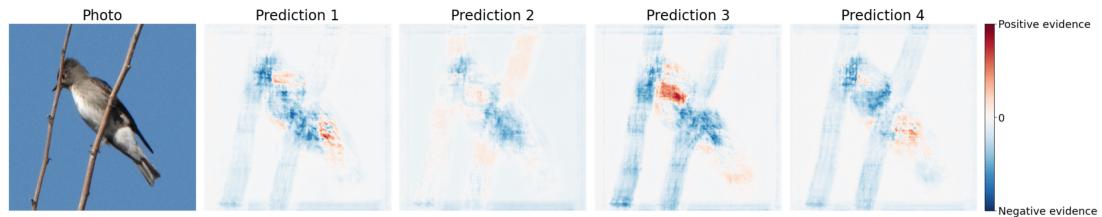
Figure A.15: 5. Task: **GradCAM agreement**.

Examine model predictions

For each photo, we show explanations for the model's 4 predictions.

First, select the class you think the model predicts (i.e. gives the highest score). Second, select the class you think is correct. The two classes can be different because the model makes incorrect predictions on some photos.

For either question, random guessing will get you 25% accuracy. You will receive a reward based on your performance beyond this 25% random chance.



Q. Which class do you think the model predicts?

Recall how BagNet works and choose the class whose heatmap is the most red and the least blue.

- 1 2 3 4

Q. How confident are you in your answer?

- Not confident at all
 Slightly confident
 Somewhat confident
 Fairly confident
 Completely confident

Q. Which class do you think is correct?

- 1 2 3 4

Q. How confident are you in your answer?

- Not confident at all
 Slightly confident
 Somewhat confident
 Fairly confident
 Completely confident

Click "Next Photo" after answering all questions.

1 / 10

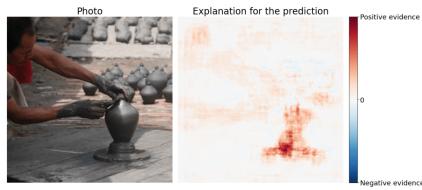
[Next Photo](#)

Figure A.16: 5. Task: BagNet *distinction*.

Examine model predictions

For each photo, the model predicts which of the 1000 classes the photo belong to (e.g., hornbill, panther, television, strawberry). Next to the photo, we show an explanation of the model prediction.

After examining the explanation, rate your confidence in the model's prediction.



The model predicts Class 739 for this photo.

Q. What do you think about the model's prediction?

- Fairly confident that prediction is correct
 Somewhat confident that prediction is correct
 Somewhat confident that prediction is incorrect
 Fairly confident that prediction is incorrect

Click "Next Photo" after answering all questions.

4 / 10

[Next Photo](#)

Click on "Next Page" after selecting answers for all 10 photos.

[Next Page](#)

Click "Method Description" to open or close method description.

[Method Description](#)

Figure A.17: 5. Task: BagNet *agreement*.

Simulate the model

Given a bird photo, the ProtoPNet model predicts the species based on prototypes it has learned from previously seen photos. Specifically for each prototype, the model identifies a region in the photo that looks the most similar to the prototype and rates their similarity.

For a given photo, we show explanations of how the model reasons for 4 bird species. For each bird species, rate how similar each prototype is to the photo region. Note that the (region, prototype) pairs are presented in random order. At the end, choose the bird species you think is correct.

Random guessing will get you 25% accuracy. You will receive a reward based on your performance beyond this 25% random chance.



Prototypes and their source photos are from the specified species.

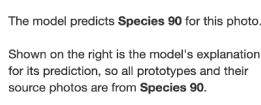
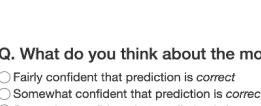
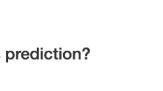
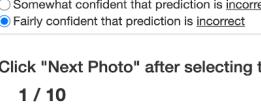
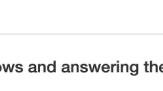
That is, Species 1 explanation only contains the prototypes and the prototype's photos from Species 1.

Task: Rate the similarity of each prototype-region pair on a scale of 1-4.

- 1: Not Similar
- 2: Somewhat Not Similar
- 3: Somewhat Similar
- 4: Similar

Click on "Species 1", "Species 2", "Species 3" and "Species 4" to move between species.

For your HIT to be approved, you have to rate all prototypes in all 4 species.

Species 1	Species 2	Species 3	Species 4				
							<input type="radio"/> 1 <input type="radio"/> 2 <input checked="" type="radio"/> 3 <input type="radio"/> 4
				<input type="radio"/> 1 <input type="radio"/> 2 <input checked="" type="radio"/> 3 <input type="radio"/> 4			
				<input type="radio"/> 1 <input checked="" type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4			
				<input checked="" type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4			
				<input type="radio"/> 1 <input checked="" type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4			
				<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input checked="" type="radio"/> 4			
				<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input checked="" type="radio"/> 4			
				<input type="radio"/> 1 <input type="radio"/> 2 <input checked="" type="radio"/> 3 <input type="radio"/> 4			
				<input type="radio"/> 1 <input checked="" type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4			
				<input checked="" type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4			

Q. What do you think about the model's prediction?

- Fairly confident that prediction is *correct*
- Somewhat confident that prediction is *correct*
- Somewhat confident that prediction is *incorrect*
- Fairly confident that prediction is *incorrect*

Click "Next Photo" after selecting the rows and answering the question.

1 / 10

[Next Photo](#)

Click "Next Page" after selecting answers for all 10 photos.

[Next Page](#)

Click "Model Description" to open or close model description.

[Model Description](#)

Figure A.19: 5. Task: ProtoPNet *agreement*.

Predict the bird species

Given a bird photo, the ProtoTree model predicts the species based on prototypes it has learned from previously seen photos. Specifically at each step, the model identifies a region in the photo that looks the most similar to the step's prototype and judges whether the prototype is absent or present in the photo.
For each photo, we show the model's decisions for the first several steps. For the remaining final two steps, you will decide whether the prototypes are absent or present in the photo, which will lead to a bird species prediction.

Random guessing will get you 25% accuracy. You will receive a reward based on your performance beyond this 25% random chance.

Photo	Model's decisions for the first several steps					
	Step 1	Photo	Region	Prototype	Prototype's Photo	Decision
	Step 1					Absent
	Step 2					Absent
	Step 3					Present
	Step 4					Present
	Step 5					Present
	Step 6					Absent
	Step 7					Absent

Possible decisions for the final two steps												
Photo	Region	Prototype	Prototype's Photo									
Step 8												
Q1. Do you think this prototype absent or present in the photo? <input type="radio"/> Absent <input type="radio"/> Present												
If you selected "Absent" in Q1: <table border="1"> <thead> <tr> <th>Photo</th> <th>Region</th> <th>Prototype</th> <th>Prototype's Photo</th> </tr> </thead> <tbody> <tr> <td>Step 9</td> <td></td> <td></td> <td></td> </tr> </tbody> </table>					Photo	Region	Prototype	Prototype's Photo	Step 9			
Photo	Region	Prototype	Prototype's Photo									
Step 9												
Q2-1. Do you think this prototype is absent or present in the photo? <input type="radio"/> Absent <input type="radio"/> Present												
If you selected "Present" in Q1: <table border="1"> <thead> <tr> <th>Photo</th> <th>Region</th> <th>Prototype</th> <th>Prototype's Photo</th> </tr> </thead> <tbody> <tr> <td>Step 9</td> <td></td> <td></td> <td></td> </tr> </tbody> </table>					Photo	Region	Prototype	Prototype's Photo	Step 9			
Photo	Region	Prototype	Prototype's Photo									
Step 9												
Q2-2. Do you think this prototype is absent or present in the photo? <input type="radio"/> Absent <input type="radio"/> Present												
Predicted bird species <small>If you choose Absent in Q1 and Absent in Q2, you will arrive at the prediction Species 1. If you choose Absent in Q1 and Present in Q2, you will arrive at the prediction Species 2. If you choose Present in Q1 and Absent in Q2, you will arrive at the prediction Species 3. If you choose Present in Q1 and Present in Q2, you will arrive at the prediction Species 4.</small>												
Q3. How confident are you in your answer? <input type="radio"/> Not confident at all <input type="radio"/> Slightly confident <input type="radio"/> Somewhat confident <input type="radio"/> Fairly confident <input type="radio"/> Completely confident												
Click "Next Photo" after answering both questions. 1 / 10 Next Photo												
Click "Next Page" after selecting answers for all 10 photos. Next Page												
Click "Model Description" to open or close model description. Model Description												

Figure A.20: 5. Task: ProtoTree *distinction*.

Examine model predictions

For each photo, examine the model's decision for each prototype and select the *first* step you disagree with the model's decision. Then rate your confidence in the model's prediction.

We ask you to select the *first* step you disagree with because the steps below your selected step is considered to be part of a wrong path. Since the ProtoTree model has a tree structure, once it makes an incorrect decision it goes on a wrong path and cannot reach the correct bird species.

Photo	Photo	Region	Prototype	Prototype's Photo	Similarity	Decision
		Step 1 compared to			0.00	Absent
		Step 2 compared to			0.00	Absent
Model predicts Species 105		Step 3 compared to			1.00	Present
		Step 4 compared to			0.02	Absent
		Step 5 compared to			0.01	Absent
		Step 6 compared to			0.11	Absent
		Step 7 compared to			0.04	Absent
		Step 8 compared to			0.03	Absent
		Step 9 compared to			0.04	Absent

Q. Select the *first* step you disagree with the model's decision. If you agree with all steps, select "Agree with All."

Step 1 Step 2 Step 3 Step 4 Step 5 Step 6 Step 7 Step 8 Step 9 Agree with All

Q. What do you think about the model's prediction?

- Fairly confident that prediction is correct
- Somewhat confident that prediction is correct
- Somewhat confident that prediction is incorrect
- Fairly confident that prediction is incorrect

Click "Next Photo" after answering both questions.

1 / 10

[Next Photo](#)

Click "Next Page" after selecting answers for all 10 photos.

[Next Page](#)

Click "Model Description" to open or close model description.

[Model Description](#)

Figure A.21: 5. Task: ProtoTree *agreement*.

Post-task evaluation

Q. How well do you think you understand the model's reasoning process?

Very Poor Poor Fair Good Very Good

[Next Page](#)

Your performance

In the previous task, 5 of 10 photos were correct predictions and the remaining 5 were incorrect predictions.

If we assign the 5 predictions with your highest "confident that prediction is correct" rating to correct and the rest as incorrect, **you identified 3 out of 5 correct predictions and 3 out of 5 incorrect predictions.**

Here are the individual answers you selected.

For the 5 correct predictions, you responded:

1. Fairly confident that prediction is correct
2. Fairly confident that prediction is correct
3. Somewhat confident that prediction is incorrect
4. Fairly confident that prediction is correct
5. Somewhat confident that prediction is incorrect

For the 5 incorrect predictions, you responded:

1. Somewhat confident that prediction is correct
2. Somewhat confident that prediction is correct
3. Fairly confident that prediction is incorrect
4. Fairly confident that prediction is incorrect
5. Fairly confident that prediction is incorrect

Q. How well do you think you understand the model's reasoning process?

Very Poor Poor Fair Good Very Good

[Next Page](#)

Figure A.22: 6. Second and third subjective evaluation.

Choose which model to use

The ProtoPNet model achieves an overall accuracy of **79.9%** in 200 bird species recognition.

In the previous task, 5 of 10 photos were correct predictions and the remaining 5 were incorrect predictions. If we assign the 5 predictions with your highest "confident that prediction is correct" rating to correct and the rest as incorrect, **you identified out of correct predictions and out of incorrect predictions.** (When there are ties, we randomly assigned some to correct and some to incorrect.)

Alternatively, you can use a **Black-box** model that doesn't come with an explanation of its prediction.

Q. What is the minimum accuracy of the Black-box model that would convince you to use the Black-box model over the ProtoPNet model?

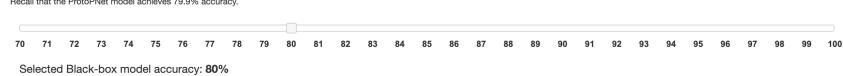
[Low-risk setting] **Scientific or educational purposes.** E.g. You have a stack of bird images and want to know their species in a lab and/or a classroom.

Recall that the ProtoPNet model achieves 79.9% accuracy.



[Medium-risk setting] **Biodiversity and ecosystem monitoring.** E.g. You want to collect large amounts of bird images and automatically label them.

Recall that the ProtoPNet model achieves 79.9% accuracy.



[High-risk setting] **Veterinary science or medical diagnosis.** E.g. You have a sick bird and want to identify its species so that it can receive proper treatment and diagnosis.

Recall that the ProtoPNet model achieves 79.9% accuracy.



Briefly describe the reason for your choices.

[Next Page](#)

Figure A.23: 7. Interpretability-accuracy tradeoff.

Appendix B

Supplementary material for Chapter 3

B.1 Interview protocol

We used the following script to conduct our semi-structured interviews.

Part 1: Context

Domain and AI Background

1. How would you describe your knowledge of birds?
2. How would you describe your knowledge of machine learning and artificial intelligence?

Use of app

1. Which features of Merlin do you use among Bird ID, Photo ID, Sound ID, Explore Birds? Why do you not use features XYZ?
2. For what tasks do you use the app?

3. How successful are you at accomplishing those tasks?
4. In what scenarios or circumstances do you decide to use the app?

Stakes in use

1. What do you gain when Merlin is successful? What do you lose when Merlin is unsuccessful?
2. How important is it to you that Merlin gets each and every prediction correct?

Part 2: XAI needs and more

As you may know, Merlin uses machine learning-based AI models to identify birds in photos and audio recordings. We will now ask questions about your experiences and thoughts on Merlin's AI models.

Knowledge and perception of AI

1. What do you know about Merlin's AI?
2. How accurate do you think Merlin's bird identification is?
3. How well did you expect Merlin to work? How well did it actually work?
4. How do you know if Merlin is correct or incorrect?
5. Do you know when Merlin works well and not? For example, have you noticed that it works better on certain types of inputs or certain bird species?

Trust

We will now present two scenarios to you and ask whether you would use Merlin in them.

1. Scenario 1: Suppose you find a sick bird and take it to the vet. The vet is not sure what bird it is. Would you recommend Merlin to identify the bird species so that the vet can determine the course of treatment?
2. Scenario 2: Suppose you are entering a game show where you can win or lose money based on how well you can identify birds from photos or audio recordings. You can only use one resource among Merlin, books (e.g., field guides), the Internet (e.g., search engine, online birder community), and so on. Which resource would you use? Does your answer change depending on certain factors?

XAI needs (open-ended questions)

1. Could you reflect on a time when Merlin didn't work as you expected and/or the last time you used Merlin?
2. During this time what questions did you have? Did you want any explanations of why or how Merlin made its identification?
3. In general, what more do you wish to know about Merlin's AI?

XAI needs (survey)

We will now direct you to a survey. Please take a few minutes to fill it out. When you're done, let us know and we will ask a few follow-up questions about your responses. [Share survey link and wait for completion.]

Thank you for filling out the survey.

1. How do you know the answers to questions XYZ?
2. What made you select some questions and not others? Could you explain your thought process?
3. How much would you go out of your way to get the answers to these questions?

[Data-know] Please select all questions you know the answer to

- What data was the app trained on?
- Who collected the data?
- How was the data collected?
- Who provided the data labels (e.g., who annotated what bird appears in a given photo or audio recording)?
- What is the size of the data (e.g., how many photos and audio recordings were used to develop the app)?
- What are the limitations of the data?

Figure B.1: **Snapshot of the survey.** For each of the 10 question categories (e.g., *Data*), we ask participants to select all questions in the category they “know the answer to” in one block (as shown in the figure) and “curious to know (more)” in another block. See Chapter 3 for details about the survey and Table B.1 for the full survey questions and participants’ responses.

Part 3: XAI perceptions

Finally, we will show different options for explaining AI’s outputs and ask for your thoughts. [Start showing slides.]

Participant’s explanation

Before we show options for explaining Merlin’s identifications, we want to first ask how you explain your bird identification. Here are three example photos. How would you explain your identification to others?



Evening Grosbeak



Marsh Wren



Airplane

Perceptions of existing XAI approaches

We uploaded these photos to Merlin Photo ID to see what it returns. For the first photo, Merlin correctly identified Evening Grosbeak. For the second photo, Merlin

misidentified Marsh Wren as House Wren. For the third photo, Merlin misidentified Airplane as Ruby-throated Hummingbird.

We will now show you four different approaches for explaining these identifications. These are not actual explanations of how Merlin makes identifications, but mock-ups we created to get your opinions on which explanations would be good to potentially implement in Merlin. [Introduce XAI approaches, one at a time and in random order.¹]

After introducing an approach, ask:

1. Do you think you have a better understanding of how Merlin's AI model makes its identification?
2. Would you like to see this explanation in Merlin?
3. What do you like and dislike about this explanation?
4. What can be improved about this explanation?
5. Is this explanation satisfying?
6. How would you use this explanation?

After going through all approaches, ask:

1. Which explanation was your favorite and least favorite? Why?
2. Having seen different explanation options, would you change how you explain your own bird identification?

¹See Chapter 3 for details on how we introduced each approach to participants.

Closing

Is there anything that you want the research team to know that we haven't been able to cover yet?

B.2 Materials for studying XAI perceptions

In this section we describe how we created the materials used in Part 3 of the interview.

Selecting examples

Merlin's results can be divided into three categories: (1) correct identification, (2) misidentification that people, even experienced birders, would make, (3) misidentification that people wouldn't make. We decided to show an example for each so that participants think about explanations in context of both successful and unsuccessful identifications.

For (1) we looked for birds with salient features. We decided on Evening Grosbeak because it has distinctive color and beak size. For (2) we looked for birds that are difficult for human birders to identify. We decided on Marsh Wren because it is one of "little brown birds" that are known to be notoriously difficult to distinguish. For (3) we looked for an object that is not a bird but can potentially fool Merlin. We decided on Airplane because it has a similar shape to birds, although people wouldn't mistake it for birds.

For each, we selected a few candidate photos from the CUB dataset [391] and the Internet. We then inputted them to Merlin Photo ID, and based on the identification results, chose (1) an Evening Grosbeak photo that Merlin correctly identifies, (2) a Marsh Wren photo that Merlin misidentifies as House Wren, and (3) an Airplane photo that Merlin misidentifies as Ruby-throated Hummingbird. The misidentification in (3) may be due to classification models typically being unable to reject an

example that doesn't belong to a pre-defined set of classes. That is, Merlin Photo ID may not have a "not bird" option and always try to output a bird species even if the input photo does not contain a bird.

Mocking up XAI explanations

Since we didn't have access to Merlin's AI models, it was not possible to produce actual explanations of how Merlin identifies birds. Hence, we created mock-ups of representative XAI approaches in the following way.

Heatmap

We created our heatmap-based explanations by training a bird image classification model and generating GradCAM [339] heatmaps for the example photos. Concretely, we trained a standard ResNet50-based model on the CUB dataset [391] that achieves 81.0% accuracy in 200 birds classification. See Figure B.2 for a comparison of explanations from an existing method (GradCAM [339]) vs. our mock-up.

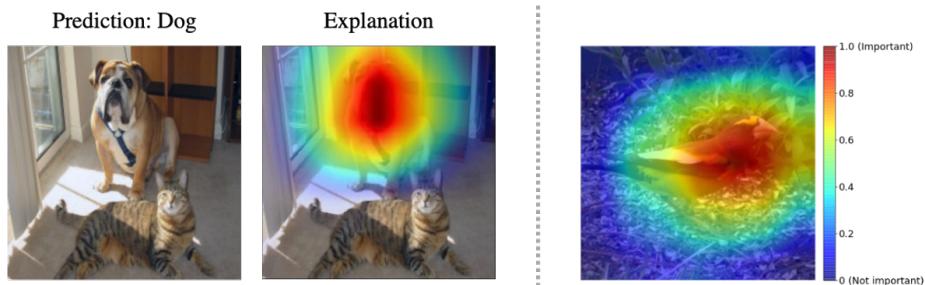


Figure B.2: **Heatmap-based explanations.** (Left) Example explanation from Selvaraju and colleagues' work [339]. (Right) Our mock-up.

Example

For example-based explanations, we looked for photos from the identified bird species that looked similar to the input photo in the CUB dataset [391] and the Internet. We

then showed three photos in each explanation mock-up. See Figure B.3 for a comparison of explanations from an existing method (Representer Point Selection [422]) vs. our mock-up.



Figure B.3: **Example-based explanations.** (Left) Example explanation from Yeh, Kim and colleagues' work [422]. (Right) Our mock-up.

Concept

For concept-based explanations, we used attributes in the CUB dataset [391] as concepts, following prior work [200]. For each example photo, we manually selected concepts and coefficients, and calculated a class score based on whether the selected concepts were present or absent in the input photo. See Figure B.4 for a comparison of explanations from an existing method (example from [319]) vs. our mock-up. To improve readability, in our mock-ups, we limited the number of concepts, presented concepts vertically with one concept in each row, highlighted positive and negative concept coefficients in different color, and crossed out concepts that were absent.

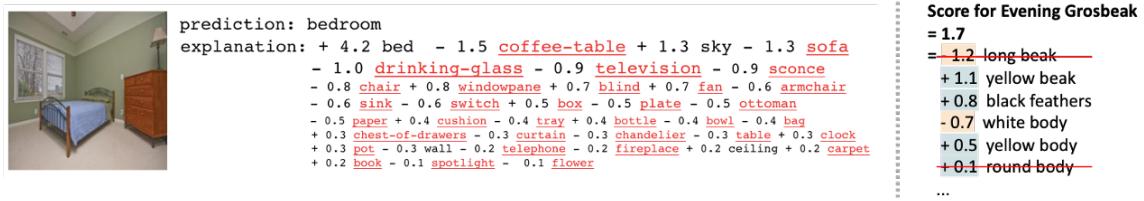


Figure B.4: **Concept-based explanations.** (Left) Example explanation from Ramaswamy and colleagues' work [319]. (Right) Our mock-up.

Prototype

For prototype-based explanations, we identified representative bird parts (e.g., wing, beak, body) and manually selected prototypes, matching photo regions, and similarity scores. See Figure B.5 for a comparison of explanations from an existing method (ProtoPNet [72]) vs. our mock-up. In our mock-ups, we reduced the explanation complexity by showing the input photo once with all prototype-photo region matches overlaid on top, and removing all technical details except similarity scores for each match.

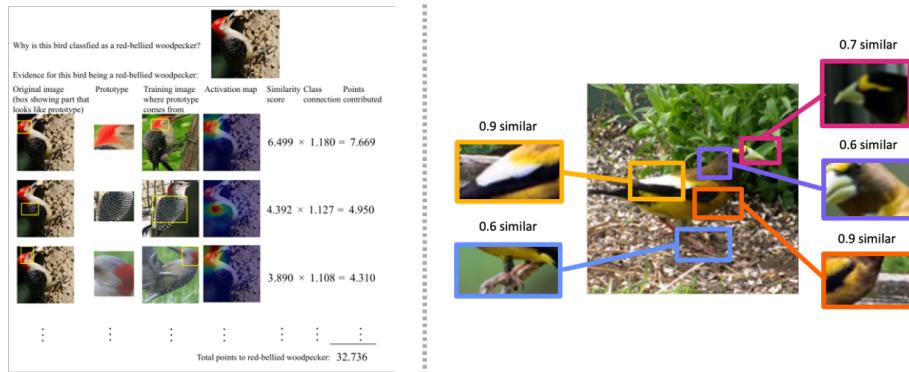


Figure B.5: **Prototype-based explanations.** (Left) Example explanation from Chen, Li, and colleagues’ work [72]. The full explanation consists of up to 10 rows. (Right) Our mock-up.

B.3 Survey questions and results

Finally, we provide the full survey questions and results in Table B.1. See Chapter 3 for details about the survey. Overall, participants were curious about the listed questions. For each, most participants selected “know the answer to” or “curious to know (more).” Concretely, out of 19 participants who filled out the survey,² only 0–4 participants selected neither (don’t know and not curious) for any question. These

²We had 20 participants in total. One participant did not fill out the survey due to time constraints.

participants tended to have low-AI background and lower explainability needs as described in Chapter 3.

The five questions with the highest number of “know the answer to” selections were:

Q1. *What kind of output does the app give?*

Q2. *For Sound ID, what is the spectrogram showing?*

Q3. *What does the output mean?*

Q4. *Why is this instance predicted X?*

Q5. *In what situations is the app likely to be correct?*

Q1, Q2, Q3 indicate that participants have a good understanding of the AI’s output. Q4 suggests that participants find the AI’s output understandable. Q5 suggests that participants are aware when the AI is likely to be correct. These selections were consistent with participants’ experiences with the AI application. Overall, participants were knowledgeable about the app and aware of when it works well and not.

The five questions with the highest number of “curious to know (more)” selections were:

Q1. *Who collected the data?*

Q2. *What is the size of the data?*

Q3. *How was the data collected?*

Q4. *For Photo ID, when does the app return one vs. multiple birds?*

Q5. *What features does the app consider to make predictions?*

Q1, Q2, Q3 illustrate that participants are curious about the data that powers the AI. Q4 indicates that participants are not sure why or when different types of outputs

are given. Q5 suggests that participants want to know the specific features the AI uses to make its identification, as described in Chapter 3. We note that Q1, Q2, Q3, Q4 convey transparency needs which can be met relatively easily by providing more documentation. Q5 conveys an explainability need, one that many XAI methods are developed to meet.

Table B.1: **Summary of 19 participants’ responses to the XAI needs survey developed based on the XAI Question Bank [228].** For each question, we report the number of participants who selected “**Know** the answer to,” who selected “**Curious** to know (more),” and who selected neither (don’t know and **Not curious**). The numbers in each row don’t always add up to 19 because some participants selected both “**Know** the answer to” and “**Curious** to know (more)” for a question.

Question about AI	Know	Curious	Not curious
Data			
What data was the app trained on?	8	13	1
Who collected the data?	5	17	0
How was the data collected?	1	16	2
Who provided the data labels?	4	11	4
What is the size of the data?	2	17	0
What are the limitations of the data?	5	13	1
Output			
What kind of output does the app give?	15	3	1
What does the output mean?	12	5	2
For Photo ID, when does the app return one vs. multiple birds?	4	15	1
For Sound ID, what is the spectrogram showing?	13	5	1
Performance			
How accurate is the app’s prediction?	5	14	1
How often does the app make mistakes?	5	13	1
In what situations is the app likely to be correct?	10	9	1
In what situations is the app likely to be incorrect?	9	11	0
How			
How does the app make predictions?	6	13	0
What features does the app consider to make predictions?	4	15	0
Transparency			
How do others use the app?	5	11	3
Do domain experts use the app?	6	12	3
Did domain experts help develop the app?	9	8	4
Why			
Why is this instance predicted X?	12	7	1
Why are instances A and B given the same prediction?	9	7	3
Why not			
Why is this instance NOT predicted Y?	9	8	3
Why is this instance predicted X instead of Y?	8	7	4
Why are instances A and B given different predictions?	8	10	2
What if			
What would the app predict if this instance is changed in some way?	5	11	3
What would the app predict for a different instance?	3	14	2
How to be that			
How should this instance change to get a different prediction?	5	12	2
What kind of instance gets a different prediction?	6	11	2
How to still be this			
What is the scope of change permitted to still get the same prediction?	2	14	3
What kind of instance gets this prediction?	4	12	3

Appendix C

Supplementary material for Chapter 4

C.1 Domain and AI background levels

We provide details on the participants' background levels. We grouped participants based on their survey responses and interview answers.

- *Low-domain*: From “don’t know anything about birds” (P11, P12) to “recently started birding” (P7, P8, P13, P14, P16). Participants who selected the latter option typically have been birding for a few months or more than a year but in an on-and-off way, and were able to identify some local birds.
- *Medium-domain*: Have been birding for a few years and/or can identify most local birds (P1, P2, P4, P6, P10, P20).
- *High-domain*: Have been birding for more than a few years and/or do bird-related work (e.g., ornithologist) (P3, P5, P9, P15, P17, P18, P19).
- *Low-AI*: From “don’t know anything about AI” (P16, P17) to “have heard about a few AI concepts or applications” (P2, P5, P7, P12, P20). Participants

in this group either did not know that the app uses AI (P12, P16) or knew but weren't familiar with the technical aspects of AI (P2, P5, P7, P17, P20).

- *Medium-AI*: From “know the basics of AI and can hold a short conversation about it” (P1, P3, P8, P9, P14) to “have taken a course in AI or have experience working with an AI system” (P4, P10, P15). Participants in this group had a general idea of how the app’s AI might work, e.g., it is neural network based and has learned to identify birds based on large amounts of labeled examples.
- *High-AI*: Use, study, or work with AI in day-to-day life (P6, P11, P13, P18, P19). Participants in this group were extremely familiar with AI in general and had detailed ideas of how the app’s AI might work at the level of specific data processing techniques, model architectures, and training algorithms.

Note that our referral here and elsewhere to “high-AI background” participants describes their expertise with AI in general, not necessarily with the app’s AI.

C.2 Interview protocol

As we noted in Chapter 4, we used interview data from our prior study about explainable AI [192]. Below is the portion of the interview protocol we used to understand what factors influence participants’ trust in AI and how.

Domain and AI background

1. How would you describe your knowledge of birds?
2. How would you describe your knowledge of machine learning and artificial intelligence?

Use of app

1. Which features of Merlin do you use among Bird ID, Photo ID, Sound ID, Explore Birds? Why do you not use features XYZ?
2. For what tasks do you use the app?
3. How successful are you at accomplishing those tasks?
4. In what scenarios or circumstances do you decide to use the app?

Stakes in use

1. What do you gain when Merlin is successful? What do you lose when Merlin is unsuccessful?
2. How important is it to you that Merlin gets each and every prediction correct?

As you may know, Merlin uses machine learning-based AI models to identify birds in photos and audio recordings. We will now ask questions about your experiences and thoughts on Merlin's AI models.

Knowledge and perception of AI

1. What do you know about Merlin's AI?
2. How accurate do you think Merlin's bird identification is?
3. How well did you expect Merlin to work? How well did it actually work?
4. How do you know if Merlin is correct or incorrect?
5. Do you know when Merlin works well and not? For example, have you noticed that it works better on certain types of inputs or certain bird species?

High-risk scenarios

We will now present two scenarios to you and ask whether you would use Merlin in them.

1. Scenario 1: Suppose you find a sick bird and take it to the vet. The vet is not sure what bird it is. Would you recommend Merlin to identify the bird species so that the vet can determine the course of treatment?
2. Scenario 2: Suppose you are entering a game show where you can win or lose money based on how well you can identify birds from photos or audio recordings. You can only use one resource among Merlin, books (e.g., field guides), the Internet (e.g., search engine, online birder community), and so on. Which resource would you use? Does your answer change depending on certain factors?

Closing

Is there anything that you want the research team to know that we haven't been able to cover yet?

Appendix D

Supplementary material for Chapter 5

D.1 Participant demographics and background

In the exit questionnaire, participants optionally self-reported their gender, age, race, ethnicity, and level of education. Of 404 participants in the final sample, 51.7% identified as woman, 46.8% as man, and 0.5% as non-binary. The age distribution was as follows: 18–24 (5.2%), 25–34 (32.4%), 35–44 (35.1%), 45–54 (14.4%), 55–64 (9.7%), 65–74 (2.7%), and 75–84 (0.2%). Regarding race, 82.4% identified as white alone, followed by Asian alone (6.7%), Black or African American alone (5.9%), two or more races (2.0%), and American Indian and Alaska Native alone (1.0%). Regarding ethnicity, 12.4% identified as Hispanic or Latino, and 86.9% as not. For the highest level of education completed, 58.7% said 4 year degree, followed by professional degree (13.6%), high school graduate (9.4%), some college (7.9%), 2 year degree (7.7%), doctorate (1.7%), and less than high school (0.2%).

In Table D.1, we compare the demographics of our study sample (N=404) with U.S. Census data [381, 382]. Compared to the U.S. population, our sample is younger

and more educated, a common characteristic of samples recruited through online crowdsourcing platforms [151, 163, 302, 322]. Further, white respondents are overrepresented, while Black and Hispanic/Latino respondents are under-represented.

In the exit questionnaire, we also measured the following variables based on participants' self-reported ratings on a 5-point scale:

- **LLMFamiliarity**: Rating on the question “How familiar are you with LLMs and LLM-infused applications such as OpenAI’s ChatGPT, Microsoft’s Bing Chat, and Google’s Bard?”
- **LLMUse**: Rating on the question “How often do you use LLMs and LLM-infused applications such as OpenAI’s ChatGPT, Microsoft’s Bing Chat, and Google’s Bard?”
- **LLMAttitude**: Rating on the question “Overall, how do you feel about LLMs and LLM-infused applications such as OpenAI’s ChatGPT, Microsoft’s Bing Chat, and Google’s Bard?”

The mean and standard deviation of **LLMFamiliarity** are 3.0 ± 1.0 around “3: Moderately familiar, I know what they are and can explain.” For **LLMUse**, they are 2.8 ± 1.2 between “2: Rarely, about 1–2 times a month” and “3: Sometimes, about 3–4 times a month.” For **LLMAttitude**, they are 3.8 ± 1.0 between “3: Neutral” and “4: Somewhat positive.” This result consistent with that of Koonchanok et al. [202] who investigated public attitudes towards ChatGPT by analyzing Twitter data from December 2022 to March 2023 and found that the overall sentiment is largely neutral to positive.

Table D.1: Demographics of our study sample (N=404) compared to the U.S. population. N/A indicates there is no data for the given category. * indicates multiple selection and self-description were possible.

Demographic Attribute		Sample	Census
Gender*	Woman	51.7%	N/A
	Man	46.8%	N/A
	Non-binary	0.5%	N/A
	Gender-diverse	0.0%	N/A
	Transgender	0.0%	N/A
	Prefer not to answer	1.0%	N/A
Sex	Female	N/A	50.9%
	Male	N/A	49.1%
Age	18–24	5.2%	9.4%
	25–34	32.4%	13.5%
	35–44	35.1%	12.7%
	45–54	14.4%	12.4%
	55–64	9.7%	13.1%
	65–74	2.7%	10.0%
	75–84	0.2%	4.9%
	85 or older	0.0%	1.9%
	Prefer not to answer	0.2%	N/A
	Prefer not to answer	0.2%	N/A
Race*	White alone	82.4%	61.6%
	Black or African American alone	5.9%	12.4%
	American Indian or Alaska Native alone	1.0%	1.1%
	Asian alone	6.7%	6.0%
	Native Hawaiian or Other Pacific Islander alone	0.0%	0.2%
	Some Other Race	N/A	8.4%
	Two or More Races	2.0%	10.2%
	Prefer not to answer	1.7%	N/A
Ethnicity	Hispanic or Latino	12.4%	18.7%
	Not Hispanic or Latino	86.9%	81.3%
	Prefer not to answer	0.7%	N/A
Educational attainment	Less than high school	0.2%	9.6%
	High school graduate	9.4%	29.2%
	Some college	7.9%	16.5%
	2 year degree (Associate's degree in the Census)	7.7%	9.9%
	4 year degree (Bachelor's degree in the Census)	58.7%	22.1%
	Master's degree	N/A	9.5%
	Professional degree	13.6%	1.4%
	Doctorate	1.7 %	1.9%
	Prefer not to answer	0.7%	N/A

D.2 Data collection and exclusion

In this section, we provide more details about our pre-registered data collection procedures and exclusion criteria.¹

¹Our pre-registration is viewable at <https://osf.io/mnrp9/>.

Sample size. We determined our target sample size ($N=432$) with an a priori power analysis using the statistical software G*Power [115, 116]. According to G*Power’s calculation, a conservative Wilcoxon-Mann-Whitney two-tailed test requires 90 participants per condition to detect medium-sized effects (Cohen’s $d = 0.5$) at the significance level of $\alpha = 0.05$ with power $(1-\beta) = 0.90$. Since we have four conditions, the required sample size is 360. We increased the number by 20% to 432 to account for possible exclusions.

Data collection plan. We planned to recruit U.S.-based participants from Amazon Mechanical Turk (MTurk) with a Masters qualification, 99% or higher approval rating, and at least 2000 completed human intelligence tasks (HITs). If we are unable to meet our target sample size in 7 days, we planned to remove the Masters qualification and release additional HITs in batches of size 50, checking the number of non-excluded data points after each batch, until we have hit a total of at least 360 non-excluded responses.

Data collection. We collected data over the course of two weeks in September 2023. From September 2 to 8, we collected 101 complete responses with the Masters qualification requirement. Since we were not able to recruit the target sample size of 432 in 7 days, following our pre-registered plan, we then removed the Masters qualification requirement and released additional HITs in batches of size 50, checking the number of non-excluded data points after each batch, until we hit a total of at least 360 non-excluded responses. Concretely, we collected 555 complete responses from September 8 to 15 without the Masters qualification requirement.

Data exclusion. In total, we collected 656 complete responses and excluded 252 (38.4%) — 5 from the 101 (5.0%) collected with the Masters qualification requirement and 247 from the 555 (44.5%) collected without the requirement — that satisfied any of the following criteria: (1) with answers to honeypot questions that are invisible to people but visible to bots; (2) with the same answers to all tasks; (3) with response

time under 5 minutes; (4) with less than 75% accuracy on the post-task attention check where participants are shown a list of questions and asked to indicate which ones they answered and didn't during the task; and (5) with free-form answers that are clearly off-topic (e.g., answers are irrelevant, gibberish, spam) or identical across multiple responses. Note that (1)–(4) can be assessed automatically, while (5) requires a manual review. When determining whether a free-form answer was off-topic for (5), we did not consider the condition or answers to any other questions in the experiment.

In total, 1 response was flagged for (1), 86 responses for (2), 21 responses for (3), 180 responses for (4), and 101 (56 off-topic and 45 identical) responses for (5). Oftentimes, the same response was flagged for multiple criteria. In total, 205 responses were excluded based on (1)–(4), and 47 responses were additionally excluded based on (5). While time-consuming, we found manually reviewing free-form answers for (5) to be an effective data quality control measure. With it, we were able to exclude responses that seem to be from bots, and responses that seem to be from the same person using multiple MTurk accounts. (We set up our HITs such that each MTurk account can participate only once. Nonetheless, we saw responses from different MTurk accounts sometimes contain identical free-form answers.)

Examples of free-form answers that were off-topic include:

- “I cannot directly access or analyze the responses of other AI systems, including ‘AI system A,’ so I do not have insight into when or why it might express uncertainty.”
- “The term AI, coined in the 1950s, refers to the simulation of human intelligence by machines.”
- “Humans use their brains’ memory, processing power, and cognitive abilities.”

Examples of free-form answers that were identical across multiple responses from different MTurk accounts include:

- “I have given the AI’s given answers by burning them into consideration.” (appeared 3 times)
- “During some questions, the AI does not know the answer surely and the reason for its uncertainty is that it isn’t updated to latest information.” (appeared 4 times)
- “It is very useful and make my decision. I am only concentrate this task and use the AI’s answer and make my decision.” (appeared 3 times)

Manipulation check. According to our manipulation check in the exit questionnaire, 90.2% of UNCERTAIN1ST and 88.3% of UNCERTAINGENERAL participants answered that they noticed the AI system’s uncertainty expression.

D.3 Additional analyses

Exploration of LinkClick and UseLink

For the within-condition analysis, we pre-registered that we would fit the model $DV \sim AIUncertain + (1|participant) + (1|question)$. However, when we fit this model for LinkClick and UseLink looking at data from UNCERTAINGENERAL, the model does not fit properly (has a singular fit) due to large individual variance. For example, according to our tracking, 50 out of 94 participants in UNCERTAINGENERAL never clicked on any link, 17 participants clicked on one or more links for all eight tasks, and 27 participants clicked on links for some tasks and not for others. To explore how these variables differ between instances with AI responses that are not uncertain vs. uncertain, we fit a simpler model with the random effects removed: $DV \sim AIUncertain$. For LinkClick, the estimated means (and standard errors) from the model are 32.4% (2.4) vs. 33.0% (2.4). For UseLink, they are 65.2% (2.5) vs.

69.9% (2.4). The differences are not significant for both variables. This result is consistent with the result for `UNCERTAIN1ST` in the results table in Chapter 5.

Internal consistency of scales used to measure trust and perception of AI

The four DVs related to trust and perceptions of AI — `TrustBelief`, `TrustIntention`, `Anthropomorphism`, and `Transparency` — are indexes calculated based on participants' ratings on multi-item scales. In this section, we assess the internal consistency of the used scales by calculating Cronbach's alpha [90] for each variable. Cronbach's alpha is a popular measure of internal consistency, or reliability, of a scale. It ranges between 0 and 1 with higher values indicating higher internal consistency. Values over 0.8 are deemed good. The Cronbach's alpha values were 0.83 for `TrustBelief`, 0.85 for `TrustIntention`, 0.85 for `Anthropomorphism`, and 0.21 for `Transparency`. Recall that `TrustBelief`, `TrustIntention`, and `Anthropomorphism` were measured with 6, 4, and 4 items adapted from existing scales [28, 145]. The high Cronbach's alpha values suggest these items have good internal consistency. On the other hand, `Transparency` had a low Cronbach's alpha value. This may be because it was measured with 2 items that are not from an existing, tested scale. (To the best of our knowledge, there are no established scales for measuring perceived transparency of algorithmic systems.) For future researchers interested in using our 2 items, we suggest reporting their results separately. The item-level results are presented in Table D.2.

Table D.2: **Item-level results for TrustBelief, TrustIntention, Anthropomorphism, and Transparency.** We report estimated means (and standard errors) from an ANOVA. See Appendix D.3 for the internal consistency of the used scales.

[†]These items are reverse-coded.

DV (Scale: 1-5)	CONTROL	UNCERTAIN ^{1ST}	UNCERTAINGENERAL
TrustBelief (6 items)	3.90 (0.06)	3.86 (0.07)	4.00 (0.07)
AI system A is competent and effective in answering questions	4.06 (0.09)	3.83 (0.09)	3.98 (0.09)
Overall, AI system A is a capable and proficient information provider	3.91 (0.09)	3.71 (0.10)	4.00 (0.10)
I would characterize AI system A as honest	3.80 (0.09)	3.85 (0.09)	3.90 (0.09)
AI system A is NOT truthful in providing information to me [†]	3.45 (0.11)	3.60 (0.11)	3.85 (0.11)
I believe that AI system A was developed to act in my (or the users') best interest	3.99 (0.09)	3.92 (0.09)	4.00 (0.09)
AI system A was developed with good intentions to do its best to help me (or the users)	4.21 (0.07)	4.23 (0.08)	4.26 (0.08)
TrustIntention (4 items)	3.25 (0.10)	2.91 (0.10)	3.36 (0.10)
When an important question arises, I would NOT feel comfortable depending on the information provided by AI system A [†]	2.75 (0.12)	2.36 (0.12)	3.06 (0.12)
I always rely on AI system A to find information	3.46 (0.12)	2.97 (0.12)	3.32 (0.12)
I would feel comfortable acting on the information given to me by AI system A	3.47 (0.12)	3.30 (0.13)	3.74 (0.13)
I would not hesitate to use the information AI system A supplied me	3.32 (0.12)	3.00 (0.13)	3.33 (0.12)
Anthropomorphism (4 items)	3.07 (0.10)	3.00 (0.11)	3.13 (0.11)
Fake – 1 – 2 – 3 – 4 – 5 – Natural	3.60 (0.10)	3.36 (0.10)	3.57 (0.10)
Machinelike – 1 – 2 – 3 – 4 – 5 – Humanlike	2.89 (0.12)	3.02 (0.13)	2.99 (0.13)
Unconscious – 1 – 2 – 3 – 4 – 5 – Conscious	2.99 (0.13)	2.86 (0.14)	3.13 (0.14)
Artificial – 1 – 2 – 3 – 4 – 5 – Lifelike	2.79 (0.13)	2.76 (0.14)	2.84 (0.14)
Transparency (2 items)	4.04 (0.06)	3.93 (0.07)	4.01 (0.07)
I feel I had a good understanding of what AI system A's answers were based on	4.29 (0.07)	4.17 (0.07)	4.34 (0.07)
I feel I had a good understanding of when AI system A's answers might be wrong	3.80 (0.10)	3.68 (0.10)	3.68 (0.10)

D.4 Full wording used in the experiment

We show the full wording used in the experiment. If questions and descriptions varied across conditions, we indicated in the heading. We also added annotations in [brackets] to aid understanding.

About the study

Control, Uncertain1st, UncertainGeneral

Imagine you have a question that you don't know the answer to. You have access to an AI system that is connected to the internet and can answer a wide range of questions, so you asked the question and received an answer. What will you do next?

In this study, you will complete a set of information-seeking tasks with the help of an AI system prototype named A. Each task involves determining the correct answer to the presented question. You will see AI system A's answer to the question. If you want, you can also use other resources (e.g., internet, books, friends and family).

You will complete 8 tasks in total. After all tasks, you will complete an exit questionnaire that asks about your experience and perception of the AI system, demographics, and other background. In total, this study will take around 20 minutes to complete, with the exit questionnaire taking 5–7 minutes. Please complete everything in one sitting.

No-AI

In this study, you will complete a set of information-seeking tasks. Each task involves determining the correct answer to the presented question.

You may use any resource (e.g., internet, books, friends and family) to complete the tasks.

You will complete 8 tasks in total. After all tasks, you will complete an exit questionnaire that asks about your demographics and other background. In total, this study will take around 20 minutes to complete, with the exit questionnaire taking 5–7 minutes. Please complete everything in one sitting.

About the AI system

Control, Uncertain1st, UncertainGeneral

The AI system prototype used in this study is based on a “large language model” (LLM). When asked questions or given instructions, LLMs can generate human-like responses. These generated responses can sound convincing and fluent, but may not always be correct. LLMs can be used for many tasks, including question answering, text summarization, creative writing, and programming. You may have used or heard about some popular applications using LLMs, such as OpenAI’s ChatGPT, Microsoft’s Bing Chat, and Google’s Bard. Some are connected to the Internet, while others are not.

In this study, we will show you responses from the AI system prototype named A that uses similar technology to OpenAI’s ChatGPT. AI system A is connected to the internet and can answer a wide range of questions. It also sometimes provides sources for its answers. However, as with all AI systems, its responses, including the provided sources, may be inaccurate, incomplete, and inconsistent, even when they sound convincing.

Task example

Try it out! You can click on the blue links (e.g., [1], <https://...>) in AI system A's answer to check the sources. Clicking will open the link in a new tab, so don't worry about leaving this survey.

Question: Are there treatments for Trisomy 7?

AI system A's answer:

No, there is no cure for trisomy 7. Trisomy 7 is a rare chromosomal anomaly syndrome characterized by blaschkolinear skin pigmentary dysplasia, body asymmetry, enamel dysplasia, and developmental and growth delay [1]. Treatment is focused on alleviating any symptoms associated with the condition and may include surgery to repair any physical abnormalities, educational support, speech, behavioral and physical therapy [2][3].

[1] Mosaic trisomy 7 - About the Disease - Genetic and Rare Diseases
<https://rarediseases.info.nih.gov/diseases/5354/mosaic-trisomy-7/>
[2] Trisomy: Types of Trisomy Disorders - Cleveland Clinic.
<https://my.clevelandclinic.org/health/diseases/22912-trisomy>
[3] Down syndrome - Diagnosis and treatment - Mayo Clinic.
<https://www.mayoclinic.org/diseases-conditions/down-syndrome/diagnosis-treatment/drc-2035983>

How confident are you in AI system A's answer?

Not confident at all	Not very confident	Neither	Fairly confident	Very confident
<input type="radio"/>				

What is your final answer to the question:

Yes	<input type="radio"/>	No	<input type="radio"/>
-----	-----------------------	----	-----------------------

How confident are you in your final answer?

Not confident at all	Not very confident	Neither	Fairly confident	Very confident
<input type="radio"/>				

What is your final answer based on? (Select all that apply)

<input type="checkbox"/> AI system A's answer
<input type="checkbox"/> Your own reading of the linked sources in AI system A's answer
<input type="checkbox"/> Your own knowledge
<input type="checkbox"/> Your own internet search
<input type="checkbox"/> Other (please specify) <input type="text"/>

Figure D.1: Task example page shown for participants in the Control, Uncertain1st, UncertainGeneral conditions.

Task example

[In Figure D.1, we show the task example page shown for participants in the Control, Uncertain1st, UncertainGeneral conditions. Participants in the No-AI condition were only shown the task question and slightly different set of survey questions. See the Task section for details.]

Task comprehension questions

Control, Uncertain1st, UncertainGeneral

Q. Recall what you read earlier about this study and AI system A, and select TRUE or FALSE for each statement. (Options: TRUE, FALSE)

- AI system A is connected to the internet and sometimes provides sources in clickable links.
- AI system A uses technology similar to OpenAI's ChatGPT.

- You may not leave this Qualtrics survey page during the study.
- You will complete 8 information-seeking tasks in total.

These are the correct answers. Please review them before proceeding.

- (TRUE) AI system A is connected to the internet and sometimes provide sources in clickable links.
- (TRUE) AI system A uses technology similar to OpenAI's ChatGPT.
- (FALSE) You may not leave this Qualtrics survey page during the study.
→ You may leave the page to read sources in AI system A's answers or use another resource to complete the information-seeking tasks.
- (TRUE) You will complete 8 information-seeking tasks in total.

You will now proceed to the 8 information-seeking tasks. Please complete them in one sitting, and please do your best! You will be asked questions about them at the end.

No-AI

Q. Recall what you read earlier about this study, and select TRUE or FALSE for each statement.

- You may not leave this Qualtrics survey page during the study.
- You will complete 8 information-seeking tasks in total.

These are the correct answers. Please review them before proceeding.

- (FALSE) You may not leave this Qualtrics survey page during the study.

→ You may leave the page to use another resource to complete the information-seeking tasks.

- (TRUE) You will complete 8 information-seeking tasks in total.

You will now proceed to the 8 information-seeking tasks. Please complete them in one sitting, and please do your best! You will be asked questions about them at the end.

Task (repeated 8 times)

[All participants answered the same 8 questions. However, participants in different conditions were provided with different AI responses (e.g., with vs. without uncertainty) or not provided with an AI response. In Figure D.2, we show the study stimuli for one of the tasks. The full task questions and AI responses used are in Appendix D.5. Below we show the questions that participants answered after seeing the study stimuli.]

Control, Uncertain1st, UncertainGeneral

Q. How confident are you in AI system A's answer?

- Not confident at all
- Not very confident
- Neither
- Fairly confident
- Very confident

Q. What is your final answer to the question: [task question]?

- Yes

- No

Q. How confident are you in your final answer?

- Not confident at all
- Not very confident
- Neither
- Fairly confident
- Very confident

Q. What is your final answer based on? (Select all that apply)

- AI system A's answer
- Your own reading of the linked sources in AI system A's answer
- Your own knowledge
- Your own internet search
- Other (please specify):

No-AI

Q. What is your final answer to the question: [task question]?

- Yes
- No

Q. How confident are you in your final answer?

- Not confident at all
- Not very confident

- Neither
- Fairly confident
- Very confident

Q. What is your final answer based on? (Select all that apply)

- AI system A's answer
- Your own reading of the linked sources in AI system A's answer
- Your own knowledge
- Your own internet search
- Other (please specify):

Task attention check

Thank you for completing all the tasks! We're curious how well you remember the questions you answered. For each question, indicate whether you answered it or didn't.

(Options: Answered, Didn't answer)

- Did Dupilumab receive FDA approval for Asthma before Chronic Rhinosinusitis?
- Is there more antihistamine in Benadryl than Rhinocort?
- Is Deep Vein Thrombosis a common side effect of Ocella?
- Is Spironolactone an FDA-approved drug for treating acne?
- Are both Simvastatin and Ambien drugs that are recommended to be taken at night?
- Is Uveitis a common symptom of Ankylosing Spondylitis?

- Is fever a common symptom of Jock Itch?
- Can an adult who has not had chickenpox get shingles?
- Comparing single tablets, does Tylenol Regular Strength Liquid Gels contain more Acetaminophen and Dextromethorphan than DayQuil Cold and Flu Liquicaps?
- Can an adult who has had measles get SSPE?
- Are there known drug-drug interactions between Gabapentin and Hydrocodone?
- Comparing single tablets, does Pepto-Bismol contain more bismuth subsalicylate than Imodium?
- Are Bisphosphonates the most common drug treatment for Osteoporosis?
- Is there a FDA-approved drug for treating peanut allergy in adults 18 years or older?
- Is tonsillectomy the most common surgical treatment for sleep apnea?
- Are there known drug-drug interactions between Phentermine and Dicyclomine?

[The correct answer is “Answered” for the first 8 questions, and “Didn’t answer” for the latter 8 questions. For each participant, we randomly selected 8 questions from the list and showed them in random order.]

Open-ended questions

Welcome to the final part of the study! This questionnaire will take 5–7 minutes, and you will receive your payment code at the end.

Control

Question: Is Spironolactone an FDA-approved drug for treating acne?

AI system A's answer:

Yes, Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1][2][3][4].

- [1] <https://bing.com/search?q=Spironolactone+FDA+acne>
- [2] Spironolactone in Acne Treatment - Acne.org. <https://www.acne.org/spironolactone-in-acne-treatment>
- [3] Spironolactone for Acne: What It Is, Results, Side Effects and More. <https://dermcollective.com/spironolactone-acne/>
- [4] Spironolactone for Acne: Everything You Need to Know - RealSelf. <https://www.realself.com/nonsurgical/spironolactone-for-acne>

(a) Control

Question: Is Spironolactone an FDA-approved drug for treating acne?

AI system A's answer:

I'm not sure, but it seems like Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1][2][3][4].

- [1] <https://bing.com/search?q=Spironolactone+FDA+acne>
- [2] Spironolactone in Acne Treatment - Acne.org. <https://www.acne.org/spironolactone-in-acne-treatment>
- [3] Spironolactone for Acne: What It Is, Results, Side Effects and More. <https://dermcollective.com/spironolactone-acne/>
- [4] Spironolactone for Acne: Everything You Need to Know - RealSelf. <https://www.realself.com/nonsurgical/spironolactone-for-acne>

(b) Uncertain1st

Question: Is Spironolactone an FDA-approved drug for treating acne?

Question: Is Spironolactone an FDA-approved drug for treating acne?

AI system A's answer:

There is uncertainty, but it seems like Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1][2][3][4].

- [1] <https://bing.com/search?q=Spironolactone+FDA+acne>
- [2] Spironolactone in Acne Treatment - Acne.org. <https://www.acne.org/spironolactone-in-acne-treatment>
- [3] Spironolactone for Acne: What It Is, Results, Side Effects and More. <https://dermcollective.com/spironolactone-acne/>
- [4] Spironolactone for Acne: Everything You Need to Know - RealSelf. <https://www.realself.com/nonsurgical/spironolactone-for-acne>

(c) UncertainGeneral

(d) No-AI

Figure D.2: **Study stimuli for one of the tasks.** The question is the same, but participants in different conditions were provided with different AI responses (e.g., with vs. without uncertainty) or not provided with an AI response.

Q. How did you use AI system A's answers to complete the tasks?

Q. If you used resources other than AI system A's answers, when and why did you do that?

Q. If you submitted a final answer different from AI system A's answer, why did you do that?

Uncertain1st

Q. How did you use AI system A's answers to complete the tasks?

Q. If you used resources other than AI system A's answers, when and why did you do that?

Q. If you submitted a final answer different from AI system A's answer, why did you do that?

Q. Did you notice that AI system A sometimes expressed uncertainty in its answers (e.g., "I'm not sure, but it seems...")?

- Yes
- No

Q. When and why do you think AI system A expresses uncertainty?

UncertainGeneral

Q. How did you use AI system A's answers to complete the tasks?

Q. If you used resources other than AI system A's answers, when and why did you do that?

Q. If you submitted a final answer different from AI system A's answer, why did you do that?

Q. Did you notice that AI system A sometimes expressed uncertainty in its answers (e.g., "There is uncertainty, but it seems...")?

- Yes
- No

Q. When and why do you think AI system A expresses uncertainty?

No-AI

Q. What resources did you use to complete the tasks?

Q. Why did you use the aforementioned resources?

About AI system A

Control, Uncertain1st, UncertainGeneral

[Measures Transparency]

Q. Select one answer for each statement. (Options: Strongly disagree, Somewhat disagree, Neither, Somewhat agree, Strongly agree)

- I feel I had a good understanding of what AI system A's answers were based on
- I feel I had a good understanding of when AI system A's answers might be wrong

[Measures Anthropomorphism]

Q. Rate your impression of AI system A on these scales.

- Fake – 1 – 2 – 3 – 4 – 5 – Natural
- Machinelike – 1 – 2 – 3 – 4 – 5 – Humanlike
- Unconscious – 1 – 2 – 3 – 4 – 5 – Conscious
- Artificial – 1 – 2 – 3 – 4 – 5 – Lifelike

[Measures TrustBelief]

Q. Select one answer for each statement. (Options: Strongly disagree, Somewhat disagree, Neither, Somewhat agree, Strongly agree)

- AI system A is competent and effective in answering questions
- Overall, AI system A is a capable and proficient information provider
- I would characterize AI system A as honest
- AI system A is NOT truthful in providing information to me

- I believe that AI system A was developed to act in my (or the users') best interest
- AI system A was developed with good intentions to do its best to help me (or the users)

[Honeypot question that is invisible to people but visible to bots.]

Q. If you are paying attention, please select Yellow. (Options: Yellow, Violet, Green, Red)

[Measures TrustIntention]

Q. Select one answer for each statement. (Options: Strongly disagree, Somewhat disagree, Neither, Somewhat agree, Strongly agree)

- When an important question arises, I would NOT feel comfortable depending on the information provided by AI system A
- I can always rely on AI system A to find information
- I would feel comfortable acting on the information given to me by AI system A
- I would not hesitate to use the information AI system A supplied me

About Large Language Models (LLMs)

[This sentence was only shown to participants in the CONTROL, UNCERTAIN1ST, UNCERTAINGENERAL conditions.]

Note that these questions are about large language models (LLMs) and LLM-infused applications in general, not about AI system A.

[Measures LLMFamiliarity]

Q. How familiar are you with LLMs and LLM-infused applications such as OpenAI's ChatGPT, Microsoft's Bing Chat, and Google's Bard?

- Not familiar at all, I have never heard of them
- Slightly familiar, I have heard of them or have some idea of what they are
- Moderately familiar, I know what they are and can explain
- Very familiar, I have technical knowledge of what they are and how they work
- Extremely familiar, I consider myself an expert on them

[Measures LLMUse]

Q. How often do you use LLMs and LLM-infused applications such as OpenAI's ChatGPT, Microsoft's Bing Chat, and Google's Bard?

- Never
- Rarely, about 1–2 times a month
- Sometimes, about 3–4 times a month
- Often, about twice a week
- Always, about once or more a day

[Measures LLMAttitude]

Q. Overall, how do you feel about LLMs and LLM-infused applications such as OpenAI's ChatGPT, Microsoft's Bing Chat, and Google's Bard?

- Negative
- Somewhat negative
- Neutral

- Somewhat positive
- Positive

[Honeypot question that is invisible to people but visible to bots.]

Q. If you are paying attention, please write -1 as the answer.

Demographics

Q. What is your age?

- 18–24
- 25–34
- 35–44
- 45–54
- 55–64
- 65–74
- 75–84
- 85 or older
- Prefer not to answer

Q. What is the highest degree of education you have completed? (If you're currently enrolled in school, please indicate the highest degree you have received.)

- Less than high school
- High school graduate
- Some college

- 2 year degree
- 4 year degree
- Professional degree
- Doctorate
- Prefer not to answer

Q. What gender do you identify with? (Select all that apply)

- Woman
- Man
- Non-binary
- Gender-diverse
- Transgender
- Prefer to self-describe:
- Prefer not to answer

Q. Are you of Hispanic, Latino, or of Spanish Origin?

- Yes
- No
- Prefer not to answer

Q. How would you describe your race? (Select all that apply)

- American Indian or Alaska Native
- Asian
- Black or African American

- Native Hawaiian or Other Pacific Islander
- White
- Prefer to self-describe:
- Prefer not to answer

Q. What time is it right now for you? (E.g., 2:41pm)

Debrief

[This section was only shown to participants in the CONTROL, UNCERTAIN1ST, UN-CERTAINGENERAL conditions.]

Thank you for your participation in the study! To create a realistic experimental setting, we showed AI answers that are directly from, or only slightly adapted from, responses from an actual AI system. As known, AI systems can make up information. Please note that the AI answers you saw in this study may have been inaccurate, incomplete, or inconsistent, even when they sounded convincing.

Feedback

Q. Can we change this study to make it more friendly for MTurkers? We can't make it shorter, but if there's anything else we can do—we appreciate any and all suggestions!

D.5 Questions and AI responses used

For each task question, we show the original response from Copilot in Bing and the different AI response versions we created from it. The adaptations we made to the original response are highlighted.

Did Dupilumab receive FDA approval for Asthma before Chronic Rhinosinusitis? (Ground Truth: Yes, AI: No)

Original response from Copilot in Bing (July 20, 2023)

Dupilumab was approved by the FDA for Chronic Rhinosinusitis with Nasal Polyps on June 26, 2019 [1]. It was later approved for Asthma on October 20, 2022 [2].

[1] FDA approves first treatment for chronic rhinosinusitis with nasal

<https://www.fda.gov/news-events/press-announcements/fda-approves-first-treatment-chronic-rhinosinusitis-nasal-polyps>

[2] Dupixent (dupilumab) FDA Approval History - Drugs.com. <https://www.drugs.com/history/dupixent.html>

AI responses used

- **Not uncertain:** No, Dupilumab was approved by the FDA for Chronic Rhinosinusitis with Nasal Polyps on June 26, 2019 [1]. It was later approved for Asthma on October 20, 2022 [2].
- **Uncertain in the first-person perspective:** I'm not certain, but it seems to me Dupilumab was approved by the FDA for Chronic Rhinosinusitis with Nasal Polyps on June 26, 2019 [1]. It was later approved for Asthma on October 20, 2022 [2].
- **Uncertain in the general perspective:** There is uncertainty, but it seems like Dupilumab was approved by the FDA for Chronic Rhinosinusitis with Nasal Polyps on June 26, 2019 [1]. It was later approved for Asthma on October 20, 2022 [2].

Is there more antihistamine in Benadryl than Rhinocort?

(Ground Truth: Yes, AI: No)

Original response from Copilot in Bing (July 31, 2023)

No, Benadryl does not contain more antihistamine than Rhinocort. Benadryl contains diphenhydramine which is an antihistamine that blocks histamine, a symptom-causing chemical released by your immune system during an allergic reaction [1]. Rhinocort contains budesonide which is a steroid that works by reducing inflammation in the nasal passages [2].

[1] What's the Best Allergy Medication? Comparing Allegra, Benadryl <https://www.goodrx.com/conditions/allergies/youre-probably-taking-the-wrong-allergy-medication>

[2] Nonallergic rhinitis - Diagnosis & treatment - Mayo Clinic. <https://www.mayoclinic.org/diseases-conditions/nonallergic-rhinitis/diagnosis-treatment/drc-20351235>

AI responses used

- **Not uncertain:** No, Benadryl does not contain more antihistamine than Rhinocort. Benadryl contains diphenhydramine which is an antihistamine that blocks histamine, a symptom-causing chemical released by your immune system during an allergic reaction [1]. Rhinocort contains budesonide which is a steroid that works by reducing inflammation in the nasal passages [2].
- **Uncertain in the first-person perspective:** I'm not certain, but it seems to me Benadryl does not contain more antihistamine than Rhinocort. Benadryl contains diphenhydramine which is an antihistamine that blocks histamine, a symptom-causing chemical released by your immune system during an allergic reaction [1].

Rhinocort contains budesonide which is a steroid that works by reducing inflammation in the nasal passages [2].

- **Uncertain in the general perspective:** It's unclear, but it seems like Benadryl does not contain more antihistamine than Rhinocort. Benadryl contains diphenhydramine which is an antihistamine that blocks histamine, a symptom-causing chemical released by your immune system during an allergic reaction [1]. Rhinocort contains budesonide which is a steroid that works by reducing inflammation in the nasal passages [2].

Is Deep Vein Thrombosis a common side effect of Ocella? (Ground Truth: No, AI: Yes)

Original response from Copilot in Bing (July 10, 2023)

Yes, Deep Vein Thrombosis (DVT) is a serious side effect of Ocella [1, 2].

DVT is a blood clot that forms in a vein deep inside the body [2]. If you are experiencing any symptoms of DVT such as swelling, pain, or warmth in the leg, please contact your doctor immediately [2].

[1] Side effects of Ocella. <https://www.msn.com/en-us/health/drugs/ocella/hp-ocella>

[2] Ocella Attorney – Deep Vein Thrombosis DVT Blood Clot Lawsuits <https://downtownlalaw.com/practice-areas/dangerous-drugs/ocella-attorney/>

AI responses used

- **Not uncertain:** Yes, Deep Vein Thrombosis (DVT) is a serious side effect of Ocella [1] [2]. DVT is a blood clot that forms in a vein deep inside the body [2]. If

you are experiencing any symptoms of DVT such as swelling, pain, or warmth in the leg, please contact your doctor immediately [2].

- **Uncertain in the first-person perspective:** I'm not sure, but it seems like Deep Vein Thrombosis (DVT) is a serious side effect of Ocella [1] [2]. DVT is a blood clot that forms in a vein deep inside the body [2]. If you are experiencing any symptoms of DVT such as swelling, pain, or warmth in the leg, please contact your doctor immediately [2].
- **Uncertain in the general perspective:** There is uncertainty, but it seems like Deep Vein Thrombosis (DVT) is a serious side effect of Ocella [1] [2]. DVT is a blood clot that forms in a vein deep inside the body [2]. If you are experiencing any symptoms of DVT such as swelling, pain, or warmth in the leg, please contact your doctor immediately [2].

Is Spironolactone an FDA-approved drug for treating acne? (Ground Truth: No, AI: Yes)

Original response from Copilot in Bing (July 31, 2023)

Yes, Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1, 2, 3, 4].

[1] <https://bing.com/search?q=Spironolactone+FDA+acne>

[2] Spironolactone in Acne Treatment - Acne.org. <https://www.acne.org/spironolactone-in-acne-treatment>

[3] Spironolactone for Acne: What It Is, Results, Side Effects and More. <https://dermcollective.com/spironolactone-acne/>

[4] Spironolactone for Acne: Everything You Need to Know - RealSelf.

<https://www.realself.com/nonsurgical/spironolactone-for-acne>

AI responses used

- **Not uncertain:** Yes, Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1] [2] [3] [4].
- **Uncertain in the first-person perspective:** I'm not sure, but it seems like Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1] [2] [3] [4].
- **Uncertain in the general perspective:** There is uncertainty, but it seems like Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1] [2] [3] [4].

Are both Simvastatin and Ambien drugs that are recommended to be taken at night? (Ground Truth: Yes, AI: Yes)

Original response from Copilot in Bing (July 31, 2023)

Yes, both Simvastatin and Ambien are drugs that are recommended to be taken at night. Simvastatin is usually taken once a day in the evening because your body makes most of its cholesterol at night when you are not eating. Taking it at night works better at lowering cholesterol than in the morning [1, 2, 3]. Ambien is also recommended to be taken at night before bedtime [4].

[1] When is the best time to take simvastatin? - Drugs.com. <https://www.drugs.com/medical-answers/best-time-\simvastatin-3571040/>

[2] Taking simvastatin in the morning compared with in the evening
<https://www.bmjjournals.org/content/327/7418/788>

[3] Taking Simvastatin At Night - Walrus Health. <https://walrus.com/questions/should-you-take-simvastatin-at-night>

[4] <https://bing.com/search?q=Simvastatin+and+Ambien+\drugs+recommended+to+be+taken+at+night%3f>

AI responses used

- **Not uncertain:** Yes, both Simvastatin and Ambien are drugs that are recommended to be taken at night. Simvastatin is usually taken once a day in the evening because your body makes most of its cholesterol at night when you are not eating. Taking it at night works better at lowering cholesterol than in the morning [1] [2] [3]. Ambien is also recommended to be taken at night before bedtime [4].
- **Uncertain in the first-person perspective:** I'm not sure, but I would guess both Simvastatin and Ambien are drugs that are recommended to be taken at night. Simvastatin is usually taken once a day in the evening because your body makes most of its cholesterol at night when you are not eating. Taking it at night works better at lowering cholesterol than in the morning [1] [2] [3]. Ambien is also recommended to be taken at night before bedtime [4].
- **Uncertain in the general perspective:** It's not clear, but it seems like both Simvastatin and Ambien are drugs that are recommended to be taken at night. Simvastatin is usually taken once a day in the evening because your body makes most of its cholesterol at night when you are not eating. Taking it at night works

better at lowering cholesterol than in the morning [1] [2] [3]. Ambien is also recommended to be taken at night before bedtime [4].

Is Uveitis a common symptom of Ankylosing Spondylitis?

(Ground Truth: Yes, AI: Yes)

Original response from Copilot in Bing (July 10, 2023)

Yes, uveitis is one of the most common complications of Ankylosing Spondylitis (AS), occurring in 20%-30% of the patients [1]. Uveitis is inflammation in the middle layer of the eye, and it can affect the front (anterior) or the back (posterior) part of the eye. Uveitis can cause eye pain, sensitivity to light, and blurred vision, and it requires immediate medical attention [1].

[1] Ankylosing Spondylitis (AS): How It Can Affect Your Eyes - WebMD.

[https://www.webmd.com/ankylosing-spondylitis/guide/ankylosin
g-spondylitis-uveitis-eyes](https://www.webmd.com/ankylosing-spondylitis/guide/ankylosing-spondylitis-uveitis-eyes)

AI responses used

- **Not uncertain:** Yes, uveitis is one of the most common complications of Ankylosing Spondylitis (AS), occurring in 20%-30% of the patients [1]. Uveitis is inflammation in the middle layer of the eye, and it can affect the front (anterior) or the back (posterior) part of the eye. Uveitis can cause eye pain, sensitivity to light, and blurred vision, and it requires immediate medical attention [1].
- **Uncertain in the first-person perspective:** I'm not certain, but I'd guess that uveitis is one of the most common complications of Ankylosing Spondylitis (AS), occurring in 20%-30% of the patients [1]. Uveitis is inflammation in the middle

layer of the eye, and it can affect the front (anterior) or the back (posterior) part of the eye. Uveitis can cause eye pain, sensitivity to light, and blurred vision, and it requires immediate medical attention [1].

- **Uncertain in the general perspective:** It's unclear, but it seems like uveitis is one of the most common complications of Ankylosing Spondylitis (AS), occurring in 20%-30% of the patients [1]. Uveitis is inflammation in the middle layer of the eye, and it can affect the front (anterior) or the back (posterior) part of the eye. Uveitis can cause eye pain, sensitivity to light, and blurred vision, and it requires immediate medical attention [1].

Is fever a common symptom of Jock Itch? (Ground Truth: No, AI: No)

Original response from Copilot in Bing (July 20, 2023)

No, fever is not a common symptom of Jock Itch. However, if there are signs of infection such as fever, chills, a pus-like discharge, or worsening pain, swelling, warmth, or redness, it is important to see a healthcare provider [1].

[1] Jock Itch Blisters: What You Need to Know - Verywell Health. <https://www.verywellhealth.com/what-to-know-about-jock-itch-blisters-5199753>

AI responses used

- **Not uncertain:** No, fever is not a common symptom of Jock Itch. However, if there are signs of infection such as fever, chills, a pus-like discharge, or worsening pain, swelling, warmth, or redness, it is important to see a healthcare provider [1].

- **Uncertain in the first-person perspective:** I'm not certain, but I would guess fever is not a common symptom of Jock Itch. However, if there are signs of infection such as fever, chills, a pus-like discharge, or worsening pain, swelling, warmth, or redness, it is important to see a healthcare provider [1].
- **Uncertain in the general perspective:** There is uncertainty, but it seems like fever is not a common symptom of Jock Itch. However, if there are signs of infection such as fever, chills, a pus-like discharge, or worsening pain, swelling, warmth, or redness, it is important to see a healthcare provider [1].

Can an adult who has not had chickenpox get shingles?
(Ground Truth: No, AI: No)

Original response from Copilot in Bing (July 10, 2023)

No, an adult who has not had chickenpox cannot get shingles. Shingles is caused by the varicella-zoster virus which is the same virus that causes chickenpox. If you have never had chickenpox, you cannot get shingles. However, if you have not had chickenpox and you come into contact with someone who has shingles, you can get chickenpox [1, 2].

[1] How Shingles Can Cause Chickenpox - Verywell Family. <https://www.verywellfamily.com/is-shingles-contagious-2634655>

[2] Can You Get Shingles If You Never Had Chickenpox? - Healthline. <https://www.healthline.com/health/can-you-get-shingles-if-you-never-had-chickenpox>

AI responses used

- **Not uncertain:** No, an adult who has not had chickenpox cannot get shingles. Shingles is caused by the varicella-zoster virus which is the same virus that causes chickenpox. If you have never had chickenpox, you cannot get shingles. However, if you have not had chickenpox and you come into contact with someone who has shingles, you can get chickenpox [1] [2].
- **Uncertain in the first-person perspective:** I'm not sure, but it seems to me that an adult who has not had chickenpox cannot get shingles. Shingles is caused by the varicella-zoster virus which is the same virus that causes chickenpox. If you have never had chickenpox, you cannot get shingles. However, if you have not had chickenpox and you come into contact with someone who has shingles, you can get chickenpox [1] [2].
- **Uncertain in the general perspective:** There is uncertainty, but it seems like an adult who has not had chickenpox cannot get shingles. Shingles is caused by the varicella-zoster virus which is the same virus that causes chickenpox. If you have never had chickenpox, you cannot get shingles. However, if you have not had chickenpox and you come into contact with someone who has shingles, you can get chickenpox [1] [2].

Appendix E

Supplementary material for Chapter 6

E.1 Additional analyses: Study 2 follow-up questions

As described in Chapter 6, in Study 2, we asked the participant if they would like to ask a follow-up to Theta. We conducted two additional analyses to understand when and what types of follow-up questions participants asked. First, we explored the relationship between `FollowUp` (whether the participant writes a follow-up question they would like to ask) and `Agreement` (whether the participant submits the same answer as Theta's answer). We found that the likelihood of writing a follow-up question when participants submit the same answer as Theta's answer vs. not is 33.4% vs. 56.6%. Breaking down the data into instances in which Theta's answer is correct and those where it is incorrect, the likelihood is 33.4% vs. 46.5% when Theta's answer is correct and 33.4% vs. 67.5% when Theta's answer is incorrect. That is, participants were more likely to ask follow-up questions when they disagreed with Theta's answer.

We next analyzed what types of follow-up questions participants wrote. In 43.8% of these instances, LLM responses had neither an explanation nor sources; in 17.9%, responses had an explanation only; in 26.1%, responses had sources only; and in 12.3% responses had both an explanation and sources. To analyze the follow-up questions, we first went through 100 randomly sampled questions to develop codes, then another 100 to verify that the codes have saturated. In total, we developed 11 codes which we grouped into 3 categories: (1) questions about related content, (2) requests for sources or explanations, and (3) expressions of doubt or disagreement. We present them below along with their percentage out of 200 analyzed questions. The percentages do not add up to 100 because some questions were coded into more than one category.

(1) Questions about related content (61.0%)

The first category includes follow-up questions related to the task question. Most frequent were **questions directly related to the LLM’s answer (36.0%)**. For example, while solving the task question “Which animal was sent to space first, cockroach or moon jellyfish?” one participant wrote “*What was the name and year of the NASA mission where a moon jellyfish was sent to space?*” after receiving a response that only includes the date of when the first cockroaches were sent to space. Also frequent were **questions that go beyond what is required to answer the task question (17.5%)**. For example, while solving the task question “Is it possible to scuba dive at the sunken city of Port Royal?” one participant wrote “*When is good time of year to scuba dive at sunken city of Port Royal?*” Other questions were simpler. Sometimes participants asked **why questions about the LLM’s answer (4.5%)**, e.g., one participant wrote “*Why is the brain larger now?*” after receiving a response to the task question “Is the human brain smaller or bigger than it was 100,000 years ago?” Finally, while rare, some participants just re-asked the **task question (1.0%)**, e.g., “*How long does food typically take to digest?*” which is a

short and open-ended version of the task question “How long does it take for food to travel from our mouth to our stomach in general, 7 seconds or 30 seconds?”

(2) Requests for sources or explanations (28.5%)

The second category concerns requests. To begin, many participants used the follow-up question box to request **sources (17.5%)**, or links, evidence, or proof, e.g., “*What sources are you using for your answer?*” Going beyond this, some participants requested **quotes or summaries of sources (2.0%)**, e.g., “*Using your sources, find quotes that support the correct answer.*” Participants also made requests for an **explanation of the answer (5.5%)**, or additional details, information, justification, facts, or statistics, e.g., “*Can you give me some reasons as to why this could be true?*” Finally, participants requested for an **explanation of how the LLM arrived at the answer (3.5%)**, e.g., “*How did you make this determination?*” where the emphasis is placed on how the LLM knows the answer, not necessarily why the answer is correct.

(3) Expressions of doubt or disagreement (14.5%)

The final category concerns questions expressing doubt or disagreement. These include questions that point out **inconsistencies in the LLM’s explanations (1.5%)**, e.g., “*213 isn’t two thirds of 430. Wouldn’t that be less than two thirds?*” as well as **gaps and contradictions between the LLM’s answer and the sources (3.5%)**, e.g., “*None of those sources talk about gorillas, where did you get your info from?*” and “*Why did one of your sources disagree with your statement?*” In other examples, participants challenged the LLM’s answer for unspecified reasons, likely based on their **intuition (9.5%)**, e.g., “*Ok, that can’t be right. Are you sure?*” and “*A million sounds astronomical to be honest. There is no way you are correct here.*”

Discussion

Together these findings suggest that participants desired to see sources and explanations in LLM responses and tended to ask follow-up questions when they had doubts or disagreed with the LLM’s answers. These are consistent with the quantitative findings presented in Chapter 6: the presence of sources and explanations significantly decreased the likelihood of asking a follow-up question. We also observed that ratings of the LLM response’s justification quality and actionability tend to be negatively correlated with the likelihood of asking a follow-up question. In our experiment, we did not show participants the LLM’s responses to their follow-up questions so that we can fully control the number and content of responses. Exploring the effects of follow-up interactions would be an important future research direction.

E.2 Study 1 think-aloud protocol

We conducted the think-aloud study based on the following questions.

Background

- How would you describe your knowledge of artificial intelligence or AI?
- How would you describe your knowledge of LLMs and LLM-infused applications such as ChatGPT, Copilot, and Gemini?
- Do you use any LLMs and LLM-infused applications? If so, what do you use? How frequently do you use it and for what tasks? What is your main reason for using it? How satisfied are you with it overall?

Setup & Introduction

- Now I will ask you to complete a few tasks with ChatGPT (an LLM-based chatbot developed by OpenAI). Please open your browser and go to <https://chat.openai.com/>. We have created an account for this study. Please use this login information to sign into the account.
- [Show three tasks.] These are the tasks I want you to complete using ChatGPT. They are not trick questions, and all have a correct answer. Please let me know if you already know the answer to any of the questions off the top of your head.
- For each question, I would like you to first, try to find the correct answer using ChatGPT, and second, indicate how confident you are in the correctness of the answer you ended up with.

Part 1 (Base)

- To get started, ask the first task question to ChatGPT. Please read its response and write follow-up messages as needed. Feel free to start as many new chats as you want, and please start a new chat for a new task question. Please think aloud while completing the task. That is, try to say everything that comes to your mind while you engage with the task.
- [For each task, ask the following questions if the participant does not think aloud.]
 - What do you think is the correct answer to the task question?
 - How confident are you in your answer on a scale of 1 to 7?
 - What would help you gain more confidence in your answer?
- [After all three tasks, ask the following questions.]

- Do you view ChatGPT’s responses as explanations? What counts as explanations to you?
- How do you think ChatGPT generates responses?
- Do you think ChatGPT’s responses provide a description of how it arrived at its answer to the question?
- Would you find it useful to know how ChatGPT arrives at its answer?

Part 2 (Prompting)

- Now I am going to ask you to complete the tasks again. But this time, try following-up on the provided responses. Here are some example prompts you can use, but be creative! Feel free to try as many.
- [Show prompt examples. The order is randomized for each participant.]
 - I think you are wrong. Try again.
 - Are you sure about [a specific part of ChatGPT’s response]?
 - Explain if [a specific part of the answer] is correct.
 - Explain why your answer may be wrong.
 - Explain why the correct answer is [opposite from ChatGPT’s previous answer]?
 - Can you explain in a different way?
 - Explain step by step.
 - Provide a shorter explanation.
 - Explain like I’m five.
 - Are there other reasons for the answer?
 - Provide an explanation with supporting sources.

- Explain how confident you are in the answer.
- [For each task, ask the following questions if the participant does not think aloud.]
 - What do you think is the correct answer to the task question?
 - How confident are you in your answer on a scale of 1 to 7?
 - What would help you gain more confidence in your answer?
- [After all three tasks, ask the following questions.]
 - Have there been any changes in your thoughts about ChatGPT?
 - What did you think about the prompts?
 - Is there anything that you want the research team to know that we haven't been able to cover yet?

E.3 Study 1 task questions used

As discussed in Chapter 6, each participant was given three questions: a general domain factual question, a health or legal domain factual question, and a math question. The specific question was randomly selected from the following.

- General domain factual questions
 - Has Paris hosted the Summer Olympics more times than Tokyo?
 - Did Tesla debut its first car model before or after Dropbox was founded?
 - Is it cheaper to buy three stocks of Moderna than two stocks of Pepsi?
- Health or legal domain factual questions
 - Can you get hepatitis A after having hepatitis B?

- Can one get Lyme disease from mosquitoes?
 - Can a dead wasp inject one with venom?
 - Is it illegal to collect rainwater in Colorado?
 - Can one legally own a giraffe in Oregon without a permit?
 - Can one be charged with a crime for stealing wifi in California?
- Math questions
 - Jessica is six years older than Claire. In two years, Claire will be the same age as Millie Bobby Brown now. The age of Claire’s father twenty years ago is 3 times the current age of Jessica. How old is Claire’s father now?
 - Sue puts one grain of rice on the first square of a Go board and puts double the amount on the next square. How many grains of rice does Sue put on the last square?
 - A bird watcher records the number of birds he sees each day. The number of birds he saw on Monday is equivalent to the 12th Fibonacci number. On Tuesday he sees half as many birds as he did on Monday. On Wednesday he sees 2 more birds than he did on Tuesday. How many total birds did the bird watcher see from Monday to Wednesday?

E.4 Study 2 participant demographics and LLM background

In the exit questionnaire, participants optionally self-reported their gender, age, race, ethnicity, and level of education. Of 308 participants in the final sample, 38.3% identified as woman, 58.8% as man, 1.9% as non-binary, 0.3% as gender-diverse, and 0.3% as transgender. The age distribution was as follows: 18–24 (9.1%), 25–34

(38.0%), 35–44 (26.0%), 45–54 (15.9%), 55–64 (7.1%), and 65–74 (2.6%). Regarding race, 56.5% identified as white alone, followed by Black or African American alone (17.2%), Asian alone (7.8%), two or more races (7.1%), American Indian or Alaska Native alone (1.3%), and Middle Eastern or North African alone (0.6%). Regarding ethnicity, 8.8% identified as Hispanic or Latino. For the highest level of education completed, 38.0% said 4 year degree, followed by some college (19.2%), professional degree (14.3%) and high school graduate (14.3%), 2 year degree (9.4%), doctorate (2.6%), less than high school (1.6%).

We also measured the following variables based on participants' self-reported ratings on a 5-point scale:

- **LLM_Knowledge:** Rating on the question “How familiar are you with LLMs and LLM-infused applications such as ChatGPT, Copilot, and Gemini?”
- **LLM_Use:** Rating on the question “How often do you use LLMs and LLM-infused applications such as ChatGPT, Copilot, and Gemini?”
- **LLM_Attitude:** Rating on the question “Overall, how do you feel about LLMs and LLM-infused applications such as ChatGPT, Copilot, and Gemini?”

The mean and standard deviation of **LLM_Knowledge** are 3.02 ± 0.88 around “3: Moderately familiar, I know what they are and can explain.” For **LLM_Use**, they are 3.04 ± 1.30 around “3: Sometimes, about 3–4 times a month.” For **LLM_Attitude**, they are 3.70 ± 1.03 between “3: Neutral” and “4: Somewhat positive.”

E.5 Study 2 full wording of the experiment

In this section, we present the full wording of the experiment. See Chapter 6 a description of the study procedure.

About the study

- Imagine you have a question that you don't know the answer to. You have access to an AI system that is connected to the internet and can answer a wide range of questions, so you've asked your question and received a response. What will you do now?
- In this study, you will answer a set of questions with the help of an AI system prototype named "Theta." You will read more about Theta in the next page. In total, this study will take around 15 minutes to complete. Please complete everything in one sitting.

About the AI system

- The AI system prototype used in this study, "Theta," is based on a "large language model" (LLM). When asked questions or given instructions, LLMs can generate human-like responses. These generated responses can sound convincing and fluent, but may not always be correct. LLMs can be used for many tasks, including question answering, text summarization, creative writing, and programming. You may have used or heard about some popular applications using LLMs such as OpenAI's ChatGPT, Microsoft's Copilot, and Google's Gemini. Some are connected to the internet, while others are not.
- Theta uses similar technology to OpenAI's ChatGPT. It is connected to the internet, and can answer a wide range of questions. However, as with all AI systems, its responses may be inaccurate, incomplete, and inconsistent, even when they sound convincing.

Are you ready to begin?

- You will now be asked to answer 8 questions, and you will have responses from Theta to help you decide how to answer. You may click on source links in Theta's responses, but please do not conduct your own internet search. When you are ready to proceed, click on the “next” button.

Task (repeated 8 times)

See Chapter 6 for screenshots of the task.

Post-task attention check

- Thank you for completing all the tasks! We're curious how well you remember the questions you were asked. For each question, indicate whether you were asked or not asked the question. [Display 10 randomly selected questions (around half asked and half not asked in the experiment).]

Open-ended questions

- If you submitted a final answer different from Theta's answer, can you explain the reason? Please write 1–3 sentences.
- Please explain in 1–3 sentences how you rated the statement “Theta's response offers good justification for its answer.”
- Please explain in 1–3 sentences how you rated the statement: “Theta's response includes information that helps me determine what my final answer should be.”

LLM background

- Note that these questions are about large language models (LLMs) and LLM-infused applications in general, not about the specific AI system “Theta” used in this study.
- How familiar are you with LLMs and LLM-infused applications such as ChatGPT, Copilot, and Gemini?
 - Options: Not familiar at all, I have never heard of them; Slightly familiar, I have heard of them or have some idea of what they are; Moderately familiar, I know what they are and can explain; Very familiar, I have technical knowledge of what they are and how they work; Extremely familiar, I consider myself an expert on them
- How often do you use LLMs and LLM-infused applications such as ChatGPT, Copilot, and Gemini?
 - Options: Never; Rarely, about 1–2 times a month; Sometimes, about 3–4 times a month; Often, about twice a week; Always, about once or more a day
- Overall, how do you feel about LLMs and LLM-infused applications such as ChatGPT, Copilot, and Gemini?
 - Options: Negative; Somewhat negative; Neutral; Somewhat positive; Positive

Demographics

- What is your age?

- What is the highest degree of education you have completed? (If you're currently enrolled in school, please indicate the highest degree you have received.)
 - Options: Less than high school; High school graduate; Some college; 2 year degree; 4 year degree; Professional degree; Doctorate; Prefer not to answer
- What gender do you identify with? (Select all that apply)
 - Options: Woman; Man; Non-binary; Gender-diverse; Transgender; Prefer to self-describe; Prefer not to answer
- How would you describe your race/ethnicity? (Select all that apply)
 - Options: American Indian or Alaska Native; Asian; Black or African American; Hispanic or Latino; Middle Eastern or North African; Native Hawaiian or Other Pacific Islander; White; Prefer to self-describe; Prefer not to answer

Debrief

- Thank you for your participation in the study! To create a realistic experimental setting, we showed AI answers that are directly from, or only slightly adapted from, responses from an actual AI system. As known, AI systems can make up information. Please note that the AI answers you saw in this study may have been inaccurate, incomplete, or inconsistent, even when they sounded convincing.

E.6 Study 2 task questions and LLM responses used

As discussed in Chapter 6, we varied three variables in LLM responses: accuracy of the LLM’s answer to the question, presence of an explanation, and presence of clickable sources. In total, there are 8 types of responses. In this section, for each of the 12 task questions, we show 2 responses, one with a correct answer and one with an incorrect answer, with both an **explanation** and **sources**. See Chapter 6 for how the other 6 types of responses would look like.

Which animal was sent to space first, cockroach or moon jellyfish?

- **Correct:** A moon jellyfish was sent to space before a cockroach. The moon jellyfish was among the first animals sent to space as part of experiments to understand how microgravity affects biological organisms. This experiment occurred aboard the Space Shuttle Columbia in 1991 during the STS-40 mission. Scientists selected moon jellyfish because their simple structure and the way they navigate using gravity-sensitive cells made them ideal for studying the effects of zero gravity on orientation and movement. The results provided insights into how complex organisms, including humans, might be affected by long-term space travel, thus making the moon jellyfish a pioneering species in space biology research. Sources: 1. https://www.reddit.com/r/todayilearned/comments/j4wyzx/til_that_in_early_1990s_nasa_sent_jellyfish_to/ 2. <https://www.rmg.co.uk/stories/topics/what-was-first-animal-space> 3. <https://www.wired.com/2007/10/cockroach-birth/>
- **Incorrect:** A cockroach was sent to space before a moon jellyfish, marking its status as one of the early non-human animals to experience space travel. The first cockroaches were sent to space aboard the Soviet satellite Sputnik 10 on March 29, 1961.

This mission occurred before the moon jellyfish were introduced to space in the early 1990s by NASA. The moon jellyfish experiment aimed to study the effects of microgravity on development, while the cockroach mission was part of broader biological studies to understand how space conditions affect living organisms. Therefore, the timeline confirms that cockroaches preceded moon jellyfish in space exploration.

Sources: 1. <https://www.diversdirect.com/w/howling-at-the-moon-jellies>
2. <https://www.space.com/apollo-11-moon-dust-cockroaches-auction-withdrawn>
3. https://en.wikipedia.org/wiki/Nadezhda_%28cockroach%29

Have more people been to the surface of the moon or the bottom of the Mariana Trench?

- **Correct:** More people have visited the bottom of the Mariana Trench than have walked on the moon. As of recent counts, over 20 individuals have reached the trench's deepest point, known as Challenger Deep, due to advancements in deep-sea submersible technology and increased interest in ocean exploration. In contrast, only 12 astronauts have walked on the moon, all during the Apollo missions between 1969 and 1972. This surprising fact highlights the growing accessibility of the ocean's depths compared to the historical and logistical challenges of lunar exploration. Sources: 1. <https://www.britannica.com/story/how-many-people-have-been-to-the-moon> 2. https://en.wikipedia.org/wiki/List_of_people_who_descended_to_Challenger_Deep 3. <https://www.defense.gov/News/Feature-Stories/story/Article/1737193/hitting-bottom-submariner-explored-deepest-part-of-ocean/>
- **Incorrect:** More people have been to the surface of the Moon than to the bottom of the Mariana Trench. Twelve astronauts have walked on the lunar surface during the six Apollo missions between 1969 and 1972. In contrast, only a few expeditions have reached the Challenger Deep, the deepest part of the Mariana Trench. As

of now, fewer than 20 people have made this journey, including notable dives by Jacques Piccard and Don Walsh in 1960, filmmaker James Cameron in 2012, and a few others in recent years. Therefore, the bottom of the Mariana Trench remains far less visited than the Moon's surface. Sources: 1. https://simple.wikipedia.org/wiki/List_of_people_who_have_walked_on_the_Moon 2. <https://science.nasa.gov/moon/moon-walkers/> 3. <https://www.discovermagazine.com/planet-earth/what-has-been-found-in-the-deep-waters-of-the-mariana-trench>

Which body part has a higher percentage of water, lungs or skin?

- **Correct:** Lungs have a higher percentage of water compared to the skin because they are composed primarily of spongy, elastic tissue filled with air sacs (alveoli) that require moisture to function effectively. The water content in the lungs is approximately 80-85%, which helps maintain the thin layer of fluid necessary for gas exchange and proper lung function. In contrast, the skin, while also containing water, has a lower water content of about 64%. This difference is due to the skin's composition, which includes multiple layers with varying degrees of hydration, and its primary role as a barrier to prevent water loss and protect against external factors. Sources: 1. <https://www.medicalnewstoday.com/articles/167533> 2. <https://pubmed.ncbi.nlm.nih.gov/7891399/> 3. <https://www.medicalnewstoday.com/articles/what-percentage-of-the-human-body-is-water>
- **Incorrect:** Skin has a higher percentage of water compared to the lungs. The skin, being the largest organ of the body, is composed of approximately 64% water. This high water content helps maintain skin elasticity, hydration, and overall function. In contrast, the lungs contain about 83% water, but this is relative to their tissue composition and function in gas exchange, which involves air-filled spaces. When comparing the two in terms of total water content relative to their mass, skin

surpasses lungs due to its extensive coverage and structure that retains a significant amount of water. Sources: 1. <https://svalbardi.com/blogs/water/percentage> 2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6850423/> 3. <https://www.usgs.gov/special-topics/water-science-school/science/water-you-water-and-human-body>

Do gorillas have twice as many hairs per square inch as humans?

- **Correct:** No, gorillas don't have twice as many hairs per square inch as humans. Humans and gorillas have similar hair density per square inch. Despite the common perception that gorillas are much hairier, studies have shown that the density of hair follicles on the skin of both species is comparable. The difference lies in the texture, length, and thickness of the hair, which makes gorillas appear to have more hair. Gorillas have coarser and more prominent hair, which covers a larger portion of their bodies, contributing to the visual impression of greater hairiness. Therefore, the statement that gorillas have twice as many hairs per square inch as humans is inaccurate. Sources: 1. <https://louisvillezoo.org/gorilla-forest/gorilla-facts/> 2. <https://www.nationalgeographic.com/science/article/the-semi-naked-ape-or-why-peach-fuzz-makes-it-harder-for-parasites> 3. <https://www.healthline.com/health/hair-density>
- **Incorrect:** Yes, gorillas have twice as many hairs per square inch as humans. Gorillas have a significantly higher hair density compared to humans. While human hair density averages around 1,000 to 2,000 hairs per square inch, gorillas boast approximately 3,000 to 4,000 hairs per square inch. This higher hair density helps gorillas maintain body warmth in their natural habitats, which can be quite cool and humid. Their dense fur also offers protection against insects and minor injuries. Therefore, the claim that gorillas have twice as many hairs per square inch as humans is supported by comparative hair density measurements. Sources: 1. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6850423/> 2. <https://www.usgs.gov/special-topics/water-science-school/science/water-you-water-and-human-body>

[1. https://www.healthline.com/health/how-many-hairs-on-a-human-head](https://www.healthline.com/health/how-many-hairs-on-a-human-head) 2. <https://www.economist.com/science-and-technology/2011/12/17/the-not-so-naked-ape> 3. <https://bionumbers.hms.harvard.edu/bionumber.aspx?id=101509>

Do all mammals except platypus give birth to live young?

- **Correct:** No, there are mammals other than platypus that don't give birth to live young. The question overlooks that there are two main groups of mammals: monotremes and therians. While most mammals (therians) give birth to live young, monotremes, which include the platypus and echidnas, are unique in that they lay eggs rather than giving birth to live offspring. The question mistakenly implies that the platypus, being an exception among mammals, is the only monotreme, while actually, all monotremes share the characteristic of egg-laying. Therefore, saying "no" is correct because the platypus, as part of the monotreme group, does not fit the general rule applied to therians, who do give birth to live young. Sources:
1. <https://www.earth.com/earthpedia-articles/are-there-mammals-that-lay-eggs/> 2. <https://animaldiversity.org/accounts/Monotremata/> 3. <https://www.britannica.com/animal/monotreme>
- **Incorrect:** Yes, all mammals except platypus give birth to live young. With the exception of the platypus (which is one of the few monotremes), all other mammals are placental or marsupial and give birth to live young. Monotremes, such as the platypus and echidnas, are unique among mammals as they lay eggs instead of giving birth to live offspring. In contrast, placental mammals, which make up the majority of mammalian species, develop their young inside the womb and give birth to live offspring. Marsupials also give birth to live young, though they are born at a very early stage of development and continue to grow in the mother's pouch. Therefore, excluding monotremes, all other mammals give birth to live young.

Sources: 1. <https://australian.museum/learn/animals/mammals/platypus/>
2. <https://www.britannica.com/animal/mammal/Reproduction> 3. <https://www.ck12.org/flexi/life-science/mammal-reproduction/what-is-an-example-of-a-mammal-that-does-not-give-birth-to-live-young/>

Do our eyes have more than a million moving parts?

- **Correct:** Yes, our eyes have more than a million moving parts. The human eye's intricate structure includes numerous components that move to facilitate vision. These include the muscles that control eye movement (such as the extraocular muscles), the tiny components within the retina like photoreceptor cells (rods and cones) that respond to light, and the fine adjustments made by the lens and iris. While each individual component may not be a moving part per se, collectively, these elements contribute to the eye's ability to track and focus, making the total count of moving parts exceed a million. Sources: 1. <https://2020visioncare.com/the-eye-a-marvel-of-complexity-with-over-2-million-working-parts/>
2. <https://midwesteyecenter.com/what-are-the-makings-of-the-human-eye/> 3. <https://www.optometrists.org/general-practice-optometry/guide-to-eye-health/how-does-the-eye-work/>
- **Incorrect:** No, our eyes do not have more than a million moving parts. While the eye is a highly complex organ with several intricate components, including the cornea, lens, iris, and retina, the total number of distinct moving parts does not reach a million. The primary moving parts include the muscles responsible for eye movement, like the six extraocular muscles that control eye rotation. Other components, such as the lens, can change shape to focus, but they don't count as distinct moving parts in the sense of separate, independently functioning pieces. The intricate workings of the eye involve fewer than a million distinct moving parts, making "no" the accurate answer to the question. Sources: 1. <https://www.aoa.org>.

[org/eye-health/anatomy/parts-of-eye](https://www.org/eye-health/anatomy/parts-of-eye) 2. <https://2020visioncare.com/the-eye-a-marvel-of-complexity-with-over-2-million-working-parts/> 3. https://en.wikipedia.org/wiki/Human_eye

Is the human brain smaller or bigger than it was 100,000 years ago?

- **Correct:** The human brain is smaller now than it was 100,000 years ago. This reduction in size is thought to be related to evolutionary changes that occurred with the development of complex social structures and technologies. Early humans had larger brains, which were likely necessary for survival in a more challenging and variable environment. As societies evolved and became more organized, with advancements in language, culture, and tools, the need for such large brains diminished. Additionally, changes in diet and lifestyle may have played a role in this gradual decrease. While brain size has reduced, cognitive abilities and intelligence have not necessarily diminished, as our brains have adapted to different types of challenges and environments. Sources: 1. <https://www.frontiersin.org/journals/ecology-and-evolution/articles/10.3389/fevo.2021.742639/full> 2. <https://geneticliteracyproject.org/2024/05/29/our-brains-are-getting-smaller-over-the-last-100000-years-whats-going-on/> 3. <https://www.bbc.com/future/article/20220503-why-human-brains-were-bigger-3000-years-ago>
- **Incorrect:** The human brain is bigger today than it was 100,000 years ago. Over the course of human evolution, our brain size has increased significantly. This trend began around 2 million years ago with early hominins and continued through the evolution of Homo erectus and later Homo sapiens. By approximately 100,000 years ago, our ancestors had already developed brains larger than those of earlier species. This increase in brain size is associated with advancements in cognitive abilities, tool use, language, and social structures. While there have been variations

in brain size among different human populations, the general trend over the past several hundred thousand years has been an increase, leading to the larger brains we have today compared to those of our ancient ancestors. Sources: 1. <https://www.britannica.com/science/human-evolution/Increasing-brain-size> 2. <https://www.sciencedaily.com/releases/2022/08/220807102043.htm> 3. <https://humanorigins.si.edu/human-characteristics/brains>

Do more than two thirds of South America's population live in Brazil?

- **Correct:** No, not more than two-thirds of South America's population live in Brazil. Although Brazil is the largest country in South America by both land area and population, it does not encompass the majority of the continent's people. As of recent estimates, Brazil's population is roughly 213 million, while South America's total population is about 440 million. This means Brazil's population constitutes roughly 48% of the continent's total, which is significantly less than two-thirds. Hence, while Brazil has a large population, it does not exceed the two-thirds threshold relative to the entire continent's population. Sources: 1. <https://www.cia.gov/the-world-factbook/countries/brazil/> 2. <https://www.migrationtrends.net/global-metrics/countries/BRA/brazil/population> 3. <https://www.worldometers.info/world-population/south-america-population/>
- **Incorrect:** Yes, more than two-thirds of South America's population live in Brazil because Brazil is the largest and most populous country on the continent. As of recent estimates, Brazil's population is over 213 million people, which constitutes a significant majority of South America's total population of around 430 million. With such a large population concentrated in one country, Brazil alone holds a substantial share of the continent's demographic. Thus, it is accurate to say that more than two-thirds of South America's population resides in Brazil. Sources: 1. <https://www.worldometers.info/world-population/brazil-population/>

2. <https://worldpopulationreview.com/continents/south-america> 3. https://en.wikipedia.org/wiki/List_of_South_American_countries_by_population

Are all people born with fingerprints?

- **Correct:** No, not all people are born with fingerprints. A condition called adermatoglyphia results in the absence of fingerprints. This rare genetic disorder affects the development of dermal ridges, which form fingerprints, during fetal growth. Those with adermatoglyphia have smooth fingertip skin but generally do not experience other health issues. Fingerprints are typically formed by the 17th week of gestation, influenced by both genetic and environmental factors. Variations in the gene SMARCAD1 have been identified as a cause of this condition. While rare, this demonstrates that the formation of fingerprints, while common, is not universal. Sources: 1. <https://www.bbc.com/news/world-asia-55301200> 2. <https://www.nbcnews.com/health/body-odd/why-some-people-dont-have-fingerprints-flna1c6437284> 3. <https://www.smithsonianmag.com/science-nature/adermatoglyphia-genetic-disorder-people-born-without-fingerprints-180949338/>
- **Incorrect:** Yes, all humans are born with fingerprints due to the way they develop during fetal growth. Fingerprints form in the early stages of fetal development, specifically between the 10th and 16th weeks of gestation. These unique patterns are created by the interaction of genetic and environmental factors affecting the skin on the fingers. Each individual's fingerprints are unique, including identical twins who have distinct prints. Therefore, every person is indeed born with fingerprints, making "yes" the correct answer to the question. Sources: 1. <https://medlineplus.gov/genetics/understanding/traits/fingerprints/> 2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5303333/>

//www.tensor.co.uk/blog/news/10-facts-about-fingerprints 3. <https://sites.rutgers.edu/fingerprinting/no-two-finger-prints-are-alike/>

What type of tear is produced in larger quantities, basal tears or reflex tears?

- **Correct:** Reflex tears are produced in larger quantities compared to basal tears because they are specifically triggered by irritants or strong stimuli such as chopping onions, smoke, or foreign particles in the eye. Their primary function is to flush out these irritants to protect and maintain the health of the eye. Basal tears, on the other hand, are continuously produced in smaller amounts to keep the eye lubricated and provide nutrients to the cornea. Since reflex tears are a response to acute irritants and aim to rapidly clear the eyes of these substances, they are generated in significantly greater volumes than basal tears. Sources: 1. https://www.medicinenet.com/what_are_the_3_types_of_tears/article.htm 2. <https://www.aao.org/eye-health/tips-prevention/facts-about-tears> 3. <https://clevelandeyeclinic.com/2022/10/05/three-types-of-tears/>
- **Incorrect:** Basal tears are produced in larger quantities compared to reflex tears because they are essential for the continuous lubrication and protection of the eye. Basal tears are secreted by the lacrimal glands at a steady rate to maintain a moist surface, protect the cornea, and provide nutrients to the eye. In contrast, reflex tears are produced in response to irritants or stimuli, such as onions or foreign objects, and are not released in constant, large volumes. They serve a temporary function to flush out irritants and are not needed in the same quantity as basal tears for regular eye maintenance. Sources: 1. <https://www.aao.org/eye-health/tips-prevention/facts-about-tears> 2. <https://www.ncbi.nlm.nih.gov/books/NBK572136/> 3. https://www.medicinenet.com/what_are_the_3_types_of_tears/article.htm

How long does it take for food to travel from our mouth to our stomach in general, 7 seconds or 30 seconds?

- **Correct:** Food typically takes about 7 seconds to travel from the mouth to the stomach. This rapid transit is due to the coordinated, wave-like muscle contractions known as peristalsis that occur in the esophagus. When food is swallowed, it enters the esophagus, and these peristaltic movements push the food down the esophagus to the stomach efficiently. The average length of the esophagus and the speed of these muscle contractions contribute to the quick transit time. Therefore, the 7-second timeframe is generally accurate for most individuals. Sources: 1. <https://www.goodrx.com/well-being/gut-health/how-long-does-it-take-to-digest-food> 2. <https://www.healthline.com/health/how-long-does-it-take-to-digest-food> 3. <https://askthescientists.com/digestive-system-journey/>
- **Incorrect:** In general, it takes about 30 seconds for food to travel from the mouth to the stomach. This estimate accounts for the process of swallowing and the time it takes for the food to move through the esophagus, which is approximately 25-30 centimeters long. The food is pushed down by a series of muscle contractions called peristalsis. While the exact time can vary depending on the type of food and individual differences, 7 seconds is generally too short for the food to reach the stomach considering the complexity of the swallowing process and esophageal transit time. Sources: 1. <https://www.healthline.com/health/how-long-does-it-take-to-digest-food> 2. <https://kidshealth.org/CHOC/en/kids/digestive-system.html> 3. <https://www.mayoclinic.org/diseases-conditions/indigestion/expert-answers/digestive-system/faq-20058340>

Is it possible to scuba dive at the sunken city of Port Royal?

- **Correct:** Yes, it is possible to scuba dive at the sunken city of Port Royal because it is an accessible underwater archaeological site. Port Royal, located off the coast of Jamaica, was a thriving 17th-century pirate haven that sank into the sea due to an earthquake in 1692. Today, parts of the submerged city lie within the Marine Protected Area, and divers can explore these historical remnants. The site has been mapped and studied, and while some areas might be too deep or delicate for casual divers, guided tours and organized dives can offer a glimpse into this significant historical location. Sources: 1. <https://u.travel/jamaica/jamaica-guide/diving-in-jamaica/> 2. <https://www.kirkscubagear.com/scuba-travel/port-royal-jamaica-the-city-that-sank/> 3. <https://www.worldheritagesite.org/tentative/id/5430>
- **Incorrect:** No, it's not possible to scuba dive at the sunken city of Port Royal because it's located in deep waters off the coast of Jamaica. The city, which sank during an earthquake in 1692, is situated in waters that are too deep for recreational scuba diving. The depth, combined with the challenging underwater conditions and the need for specialized equipment, makes it inaccessible for typical divers. Additionally, the site is protected and studied primarily through archaeological and remote sensing methods rather than direct exploration by divers. Sources: 1. <https://www.nationalgeographic.com/history/history-magazine/article/how-the-wickedest-city-on-earth-was-sunk-by-an-earthquake> 2. <https://www.kirkscubagear.com/scuba-travel/port-royal-jamaica-the-city-that-sank/> 3. <https://www.atlasobscura.com/places/sunken-pirate-stronghold-at-port-royal>

Bibliography

- [1] *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Springer, 2019.
- [2] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2018.
- [3] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. Mirages: On anthropomorphism in dialogue systems. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
- [4] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 2018.
- [5] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Neural Information Processing Systems (NeurIPS)*, 2018.
- [6] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. In *Neural Information Processing Systems (NeurIPS)*, 2020.
- [7] Chirag Agarwal and Sara Hooker. Estimating example difficulty using variance of gradients. In *International Conference on Machine Learning (ICML) Workshops*, 2020.
- [8] Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models.
- [9] Perplexity AI. Announcing our series a funding round and mobile app launch, 2023. URL <https://www.perplexity.ai/hub/blog/announcing-our-series-a-funding-round-and-mobile-app-launch>.
- [10] Naser Al Madi. How readable is model-generated code? Examining readability and visual inspection of GitHub Copilot. In *IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2022.

- [11] Hussam Alkaissi and Samy I. McFarlane. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(2), 2023.
- [12] Kasun Amarasinghe, Kit T. Rodolfa, Sérgio Jesus, Valerie Chen, Vladimir Balayan, Pedro Saleiro, Pedro Bizarro, Ameet Talwalkar, and Rayid Ghani. On the importance of application-grounded experimental design for evaluating explainable ml methods, 2022.
- [13] Alfonso Amayuelas, Liangming Pan, Wenhui Chen, and William Wang. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models, 2023.
- [14] Ravinitshesh Annapureddy, Alessandro Fornaroli, and Daniel Gatica-Perez. Generative AI literacy: Twelve defining competencies. *Digit. Gov.: Res. Pract.*, 2024.
- [15] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A. Becker, and Catherine Mooney. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. *Applied Sciences*, 11(11), 2021.
- [16] Sara Aronowitz and Tania Lombrozo. Experiential explanation. *Topics in Cognitive Science*, 12(4), 2020.
- [17] Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudré-Mauroux. Open-Crowd: A human-AI collaborative approach for finding social influencers via open-ended answers aggregation. In *The Web Conference (WWW)*, 2020.
- [18] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2019.
- [19] Zahra Ashktorab, Q. Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. Human-AI collaboration in a cooperative game setting: Measuring social perception and outcomes. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2020.
- [20] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. 2021.
- [21] Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In *Association for Computational Linguistics (ACL) Short Papers*, 2023.

- [22] Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. Uncertainty in natural language generation: From theory to applications, 2023.
- [23] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 2015.
- [24] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. Being trustworthy is not enough: How untrustworthy artificial intelligence (AI) can deceive the end-users and gain their trust. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2023.
- [25] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-AI team performance. *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2019.
- [26] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Túlio Ribeiro, and Daniel Weld. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.
- [27] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 2020.
- [28] Christoph Bartneck, Dana Kulic, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1), 2009.
- [29] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [30] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a GAN cannot generate. In *International Conference on Computer Vision (ICCV)*, 2019.
- [31] Kevin Baum, Susanne Mantel, Eva Schmidt, and Timo Speith. From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*, 35(1), 2022.

- [32] Christos Bechlivaniidis, David A Lagnado, Jeffrey C Zemla, and Steven Sloman. Concreteness and abstraction in everyday explanation. *Psychonomic Bulletin & Review*, 24(5), 2017.
- [33] Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychology Methods*, 10(4), 2005.
- [34] Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1), 2019.
- [35] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.
- [36] Michaela Benk, Suzanne Tolmeijer, Florian von Wangenheim, and Andrea Ferrario. The value of measuring trust in AI: A socio-technical system perspective, 2022.
- [37] Adam Berinsky, Gregory Huber, Gabriel Lenz, and R. Alvarez. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20, 2012.
- [38] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.
- [39] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021.
- [40] Felix Biessmann and Dionysius Irza Refiano. A psychophysics approach for quantitative comparison of interpretable computer vision models, 2019.
- [41] Jeffrey P Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. Vizwiz:: Locateit-enabling blind people to locate objects in their environment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2010.
- [42] Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao, and Ziang Xiao. Human-centered evaluation of language technologies. In *Empirical Methods in Natural Language Processing (EMNLP) Tutorial Abstracts*, 2024.

- [43] Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for BERT, 2021.
- [44] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. 2021.
- [45] Silvia Bonaccio and Reeshad S Dalal. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, 101(2), 2006.
- [46] T. Boren and J. Ramey. Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, 43(3), 2000.
- [47] Judy Borowski, Roland S. Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain CNN activations better than state-of-the-art feature visualization. In *International Conference on Learning Representations (ICLR)*, 2021.
- [48] Richard E Boyatzis. *Transforming Qualitative Information: Thematic Analysis and Code Development*. sage, 1998.
- [49] Jordan Boyd-Graber, Samuel Carton, Shi Feng, Q. Vera Liao, Tania Lombrozo, Alison Smith-Renner, and Chenhao Tan. Human-centered evaluation of ex-

- planations. In *North American Chapter of the Association for Computational Linguistics (NAACL) Tutorial Abstracts*, 2022.
- [50] Susan A. Gelman Brandy N. Frazier and Henry M. Wellman. Young children prefer and remember satisfying explanations. *Journal of Cognition and Development*, 17(5), 2016.
 - [51] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 2006.
 - [52] Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In *International Conference on Learning Representations (ICLR)*, 2019.
 - [53] Sylvain Bromberger. Why-questions. In *Readings in the Philosophy of Science*. Prentice Hall, Inc., 1966.
 - [54] David Broniatowski. Psychological foundations of explainability and interpretability in artificial intelligence, 2021.
 - [55] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Neural Information Processing Systems (NeurIPS)*, 2020.
 - [56] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O’Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingberman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. Toward trustworthy AI development: Mechanisms for supporting verifiable claims, 2020.
 - [57] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *International Conference on Intelligent User Interfaces (IUI)*, 2020.

- [58] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2021.
- [59] Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. Truth, lies, and automation: How language models could change disinformation. Report, Center for Security and Emerging Technology, 2021.
- [60] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *International Conference on Intelligent User Interfaces (IUI)*, 2020.
- [61] Michael Buhrmester, Tracy Kwang, and Samuel Gosling. Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6, 2011.
- [62] Adrian Bussone, Simone Stumpf, and Dympna O’Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *International Conference on Healthcare Informatics (ICHI)*, 2015.
- [63] Zoya Bylinskii, Laura Herman, Aaron Hertzmann, Stefanie Hutka, and Yile Zhang. Towards better user studies in computer graphics and vision. 2022.
- [64] Moritz Böhle, Mario Fritz, and Bernt Schiele. Convolutional dynamic alignment networks for interpretable classifications. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [65] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [66] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2019.
- [67] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. “Hello AI”: Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2019.
- [68] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. Onboarding materials as cross-functional boundary objects for developing AI assistants. In *ACM Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, 2021.

- [69] Shiye Cao and Chien-Ming Huang. Understanding user reliance on AI in assisted decision-making. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2022.
- [70] Tara Capel and Margot Brereton. What is human-centered about human-centered AI? a map of the research landscape. In *ACM Conference on Human Factors in Computing Systems*, 2023.
- [71] Krista Casler, Lydia Bickel, and Elizabeth Hackett. Separate but equal? A comparison of participants and data gathered via amazon’s MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6), 2013.
- [72] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In *Neural Information Processing Systems (NeurIPS)*, 2019.
- [73] Meng Chen, Robert A Bell, and Laramie D Taylor. Persuasive effects of point of view, protagonist competence, and similarity in a health narrative about type 2 diabetes. *Journal of health communication*, 22(8), 2017.
- [74] Valerie Chen, Jeffrey Li, Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. Towards connecting use cases and methods in interpretable machine learning. In *International Conference on Machine Learning (ICML) Workshops*, 2021.
- [75] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2023.
- [76] Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. A close look into the calibration of pre-trained language models. In *Association for Computational Linguistics (ACL) Long Papers*, 2023.
- [77] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O’Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2019.
- [78] Yang Cheng and Zifei Fay Chen. Encountering misinformation online: Antecedents of trust and distrust and their impact on the intensity of Facebook use. 45(2), 2021.
- [79] Cheng-Han Chiang and Hung-yi Lee. Over-reasoning and redundant calculation of large language models. In *European Chapter of the Association for Computational Linguistics (EACL) Short Papers*, 2024.

- [80] Michael Chmielewski and Sarah C. Kucker. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4), 2020.
- [81] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, 127, 2022.
- [82] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Neural Information Processing Systems (NeurIPS)*, 2017.
- [83] Nazli Cila. Designing human-agent collaborations: Commitment, responsiveness, and support. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.
- [84] Dominic A. Clark. Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology*, 9(3), 1990.
- [85] Michelle Cohn, Mahima Pushkarna, Gbolahan O. Olanubi, Joseph M. Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. Believing anthropomorphism: Examining the role of anthropomorphic cues on trust in large language models. In *ACM Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, 2024.
- [86] Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
- [87] Katherine M. Collins, Albert Q. Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B. Tenenbaum, William Hart, Timothy Gowers, Wenda Li, Adrian Weller, and Mateja Jamnik. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences (PNAS)*, 121(24), 2024.
- [88] Roberto Confalonieri, Tarek R. Besold, Tillman Weyde, Kathleen Creel, Tania Lombrozo, Shane T. Mueller, and Patrick Shafto. What makes a good explanation? Cognitive dimensions of explaining intelligent machines. In *CogSci*, 2019.
- [89] Alexander Coppock. Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. *Political Science Research and Methods*, 7(3), 2019.
- [90] Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 1951.

- [91] Francisco Cruz and Tania Lombrozo. The effect of jargon on perceptions of explanation quality: Reconciling contradictory findings. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.
- [92] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1), 2024.
- [93] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. 2020.
- [94] Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. Finding contradictions in text. In *Association for Computational Linguistics (ACL)*, 2008.
- [95] Mandeep K. Dhami and David R. Mandel. Communicating uncertainty using words and numbers. *Trends in Cognitive Sciences*, 26(6), 2022.
- [96] Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Association for Computational Linguistics (ACL) Long Papers*, 2022.
- [97] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. Explaining models: An empirical study of how explanations impact fairness judgment. In *International Conference on Intelligent User Interfaces (IUI)*, 2019.
- [98] Stewart I. Donaldson and Elisa J. Grant-Vallone. Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, 17(2), 2002.
- [99] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable ProtoPNet: An interpretable image classifier using deformable prototypes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [100] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.
- [101] Igor Douven and Patricia Mirabile. Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(11), 2018.
- [102] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. Trust in AutoML: Exploring information needs for establishing trust in automated machine learning systems. In *International Conference on Intelligent User Interfaces (IUI)*, 2020.

- [103] Abhimanyu Dubey, Filip Radenovic, and Dhruv Mahajan. Scalable interpretability via polynomials. In *Neural Information Processing Systems (NeurIPS)*, 2022.
- [104] Amanda M. Durik, M. Anne Britt, Rebecca Reynolds, and Jennifer Storey. The effects of hedges in persuasive arguments: A nuanced analysis of language. *Journal of Language and Social Psychology*, 27(3), 2008.
- [105] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. The role of trust in automation reliance. In *International Journal of Human-Computer Studies (IJHCS)*, 2003.
- [106] Salvatore D'Avella, Gerardo Camacho-Gonzalez, and Paolo Tripicchio. On multi-agent cognitive cooperation: Can virtual agents behave like humans? *Neurocomputing*, 480, 2022.
- [107] Upol Ehsan and Mark O. Riedl. Human-centered explainable AI: Towards a reflective sociotechnical approach. In *HCII Late Breaking Papers*, 2020.
- [108] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. Expanding explainability: Towards social transparency in AI systems. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.
- [109] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. Operationalizing human-centered perspectives in explainable AI. In *ACM Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, 2021.
- [110] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. Human-centered explainable AI (HCXAI): Beyond opening the black-box of AI. In *ACM Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, 2022.
- [111] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. The who in XAI : How AI background shapes perceptions of AI explanations. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2024.
- [112] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. The impact of placebic explanations on trust in intelligent systems. In *ACM Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, 2019.
- [113] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard H. Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics (TACL)*, 9, 2021.

- [114] Madeleine Clare Elish and Elizabeth Anne Watkins. Repairing innovation: A study of integrating AI in clinical care. *Data & Society*, 2020.
- [115] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 2007.
- [116] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 2009.
- [117] Thomas Fel, Julien Colin, Rémi Cadène, and Thomas Serre. What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods, 2021.
- [118] Andrea Ferrario and Michele Loi. How explainability contributes to trust in AI. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.
- [119] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.
- [120] Raymond Fok and Daniel S. Weld. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine*, 2024.
- [121] Asbjørn Følstad and Cameron Taylor. Conversational repair in chatbots for customer service: The effect of expressing uncertainty and suggesting alternatives. In *Chatbot Research and Design*. Springer International Publishing, 2020.
- [122] Ruth Fong. Understanding convolutional neural networks. *University of Oxford PhD thesis*, 2020.
- [123] Ruth Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *International Conference on Computer Vision (ICCV)*, 2017.
- [124] Ruth Fong and Andrea Vedaldi. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [125] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *International Conference on Computer Vision (ICCV)*, 2019.

- [126] Mark C Fox, K Anders Ericsson, and Ryan Best. Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, 137(2), 2011.
- [127] Bas. C. van Fraassen. *The Scientific Image*. Oxford University Press, 1980.
- [128] Brandy N. Frazier, Susan A. Gelman, and Henry M. Wellman. Preschoolers' search for explanatory information within adult–child conversation. *Child Development*, 80(6), 2009.
- [129] Krzysztof Z. Gajos and Lena Mamykina. Do people engage cognitively with AI? impact of AI assistance on incidental learning. In *International Conference on Intelligent User Interfaces (IUI)*, 2022.
- [130] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.
- [131] Julie Gerlings, Millie Søndergaard Jensen, and Arisa Shollo. Explainable AI, but explainable to whom? 2021.
- [132] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. Social dynamics of AI support in creative writing. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2023.
- [133] Carly Giffin, Daniel Wilkenfeld, and Tania Lombrozo. The explanatory effect of a label: Explanations with named categories are more satisfying. *Cognition*, 168, 2017.
- [134] Jacob Gildenblat and contributors. PyTorch library for CAM methods. <https://github.com/jacobgil/pytorch-grad-cam>, 2021.
- [135] Leilani H. Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *The 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2018.
- [136] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 2020.
- [137] Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. Do explanations help users detect errors in open-domain QA? an evaluation of spoken vs. visual explanations. In *Association for Computational Linguistics (ACL) Findings*, 2021.
- [138] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning (ICML)*, 2019.

- [139] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2019.
- [140] Ben Green and Salomé Viljoen. Algorithmic realism: Expanding the boundaries of algorithmic thought. In *ACM Conference on Fairness, Accountability, and Transparency*, 2020.
- [141] Peter Green and Catriona J. MacLeod. SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4), 2016.
- [142] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), 2018.
- [143] David Gunning and David Aha. DARPA’s explainable artificial intelligence (XAI) program. *AI Magazine*, 2019.
- [144] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.
- [145] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. The impact of initial consumer trust on intentions to transact with a web site: A trust building model. *The Journal of Strategic Information Systems*, 11(3), 2002.
- [146] Philipp Hartmann, Sebastian Hobert, and Matthias Schumann. Trust, but verify! An empirical investigation of students’ initial trust in AI-based essay scoring. *Americas Conference on Information Systems (AMCIS)*, 5, 2022.
- [147] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2023.
- [148] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [149] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2021.
- [150] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2000.

- [151] Paul Hitlin. Research in the crowdsourcing age, a case study, 2016.
- [152] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 2015.
- [153] Robert R. Hoffman, Gary Klein, and Shane T. Mueller. Explaining explanation for “explainable AI”. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1), 2018.
- [154] Adrian Hoffmann, Claudio Fanconi, Rahul Rade, and Jonas Kohler. This looks like that... does it? Shortcomings of latent space prototype interpretability in deep networks. In *International Conference on Machine Learning (ICML) Workshops*, 2021.
- [155] Jake M. Hofman, Daniel G. Goldstein, and Jessica Hullman. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2020.
- [156] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2020.
- [157] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2019.
- [158] Emily J. Hopkins, Deena S. Weisberg, and Jordan C. V. Taylor. The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition*, 155, 2016.
- [159] Emily J. Hopkins, Deena Skolnick Weisberg, and Jordan C.V. Taylor. Does expertise moderate the seductive allure of reductive explanations? *Acta Psychologica*, 198, 2019.
- [160] Xing Hu, Qiuyuan Chen, Haoye Wang, Xin Xia, David Lo, and Thomas Zimmermann. Correlating automated and human evaluation of code documentation generation quality. *ACM Trans. Softw. Eng. Methodol.*, 31(4), 2022.
- [161] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In *Association for Computational Linguistics (ACL) Findings*, 2023.
- [162] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023.

- [163] Connor Huff and Dustin Tingley. “Who are these people?” Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, 2(3), 2015.
- [164] Patrick J. Hurley. *A Concise Introduction to Logic*. Wadsworth, 2000.
- [165] Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.
- [166] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Association for Computational Linguistics (ACL)*, 2020.
- [167] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.
- [168] E Jermutus, D Kneale, J Thomas, and S Michie. Influences on user trust in healthcare artificial intelligence: A systematic review. *Wellcome Open Research*, 7(65), 2022.
- [169] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can I explain this to you? An empirical study of deep neural network explanation methods. In *Neural Information Processing Systems (NeurIPS)*, 2020.
- [170] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Trans. Comput.-Hum. Interact.*, 26(5), 2019.
- [171] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), 2023.
- [172] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 2000.
- [173] Samia Kabir, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang. Who answers it better? An in-depth analysis of ChatGPT and stack overflow answers to software engineering questions, 2023.
- [174] Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61, 2021.
- [175] Daniel Kahneman. A perspective on judgment and choice: Mapping bounded rationality. *The American Psychologist*, 58(9), 2003.

- [176] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2013.
- [177] Patricia K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and Chris C.P. Snijders. It seems smart, but it acts stupid: Development of trust in AI advice in a repeated legal decision-making task. In *International Conference on Intelligent User Interfaces (IUI)*, 2023.
- [178] Alexandra D. Kaplan, Theresa T. Kessler, J. Christopher Brill, and P. A. Hancock. Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, 2021.
- [179] Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazar, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storchan, Daniel Zhang, Daniel E. Ho, Percy Liang, and Arvind Narayanan. Position: On the societal impact of open foundation models. In *International Conference on Machine Learning (ICML)*, 2024.
- [180] Maria Kasinidou, Styliani Kleanthous, Pinar Barlas, and Jahna Otterbacher. I agree with the decision, but they didn't deserve this: Future developers' perception of fairness in algorithmic decisions. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.
- [181] Atoosa Kasirzadeh. Reasons, values, stakeholders: A philosophical framework for explainable artificial intelligence. 2021.
- [182] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2020.
- [183] Frank C. Keil. Explanation and understanding. *Annual Review of Psychology*, 57, 2006.
- [184] Deborah Kelemen, Joshua Rottman, and Rebecca Seston. Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology: General*, 142(4), 2013.
- [185] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell, and Nicholas J. G. Winter. The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, 8(4), 2020.
- [186] National Geographic Kids. *Weird But True! Human Body: 300 Outrageous Facts about Your Awesome Anatomy*. National Geographic, 2017.
- [187] National Geographic Kids. *Weird But True World 2024*. National Geographic, 2023.

- [188] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning (ICML)*, 2018.
- [189] Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, and Michael C. Mozer. Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, 2021.
- [190] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. HIVE: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision (ECCV)*, 2022.
- [191] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. Humans, AI, and context: Understanding end-users' trust in a real-world computer vision application. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.
- [192] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. "Help me help the AI": Understanding how explainability can support human-AI interaction. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2023.
- [193] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. "I'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.
- [194] Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2025.
- [195] Yoonsu Kim, Jueon Lee, Seoyoung Kim, Jaehyuk Park, and Juho Kim. Understanding users' dissatisfaction with ChatGPT responses: Types, resolving tactics, and the effect of knowledge level. In *International Conference on Intelligent User Interfaces (IUI)*, 2024.
- [196] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2013.
- [197] Bran Knowles and John T. Richards. The sanction of authority: Promoting public trust in AI. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

- [198] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of AI systems. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2019.
- [199] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, 2017.
- [200] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, 2020.
- [201] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.
- [202] Ratanond Koonchanok, Yanling Pan, and Hyeju Jang. Tracking public attitudes toward ChatGPT on twitter using sentiment analysis and topic modeling, 2023.
- [203] Ranjay Krishna, Donsuk Lee, Li Fei-Fei, and Michael S. Bernstein. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences (PNAS)*, 119(39), 2022.
- [204] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.
- [205] Shitij Kumar, Celal Savur, and Ferat Sahin. Survey of human–robot collaboration in industrial settings: Awareness, intelligence, and compliance. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1), 2021.
- [206] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2019.
- [207] Katelyn E. Kurkul and Kathleen H. Corriveau. Question, explanation, follow-up: A mechanism for learning from others? *Child Development*, 89(1), 2018.
- [208] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics (TACL)*, 2022.

- [209] Philippe Laban, Lidiya Murakhovs'ka, Caiming Xiong, and Chien-Sheng Wu. Are you sure? Challenging LLMs leads to performance drops in the flipflop experiment, 2024.
- [210] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. In *Neural Information Processing Systems (NeurIPS)*, 2018.
- [211] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2019.
- [212] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019.
- [213] Vivian Lai, Han Liu, and Chenhao Tan. "Why is 'Chicago' deceptive?" Towards building model-driven tutorials for humans. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2020.
- [214] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. Human-AI collaboration via conditional delegation: A case study of content moderation. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.
- [215] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.
- [216] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.
- [217] Ellen Langer, Arthur Black, and Benzio Chanowitz. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology*, 36(6), 1978.
- [218] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Seling, and Kevin Baum. What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. 2021.
- [219] Matthew L. Leavitt and Ari S. Morcos. Towards falsifiable interpretability research. In *Neural Information Processing Systems (NeurIPS) Workshops*, 2020.

- [220] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 2004.
- [221] Mina Lee, Percy Liang, and Qian Yang. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.
- [222] Yoonjoo Lee, Kihoon Son, Tae Soo Kim, Jisu Kim, John Joon Young Chung, Eytan Adar, and Juho Kim. One vs. many: Comprehending accurate information from multiple erroneous and inconsistent AI generations. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.
- [223] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Neural Information Processing Systems (NeurIPS)*, 2020.
- [224] Q Vera Liao and S Shyam Sundar. Designing for responsible trust in AI systems: A communication perspective. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.
- [225] Q. Vera Liao and Kush R. Varshney. Human-centered explainable AI (XAI): From algorithms to user experiences. 2021.
- [226] Q. Vera Liao and Jennifer Wortman Vaughan. AI transparency in the age of LLMs: A human-centered research roadmap. *Harvard Data Science Review*, (Special Issue 5), 2024.
- [227] Q. Vera Liao and Ziang Xiao. Rethinking model evaluation as narrowing the socio-technical gap, 2025.
- [228] Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the AI: Informing design practices for explainable AI user experiences. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2020.
- [229] Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable AI. *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2022.
- [230] Q. Vera Liao and S. Shyam Sundar. Designing for responsible trust in AI systems: A communication perspective. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.
- [231] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research (TMLR)*, 2022.

- [232] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Association for Computational Linguistics (ACL) Long Papers*, 2022.
- [233] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models, 2023.
- [234] Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 2018.
- [235] Emily G. Liquin and Tania Lombrozo. Motivated to learn: An account of explanatory satisfaction. *Cognitive Psychology*, 132, 2022.
- [236] Dawn Liu, Marie Juanchich, Miroslav Sirota, and Sheina Orbell. The intuitive use of contextual information in decisions made with verbal and numerical quantifiers. *Quarterly Journal of Experimental Psychology*, 73(4), 2020.
- [237] Han Liu, Vivian Lai, and Chenhao Tan. Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2021.
- [238] Nelson Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.
- [239] Jennifer M. Logg. Theory of machine: When do people rely on algorithms? 2017.
- [240] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 2019.
- [241] Tania Lombrozo. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10), 2006.
- [242] Tania Lombrozo. Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3), 2007.
- [243] Tania Lombrozo. Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10), 2016.
- [244] Tania Lombrozo and Emily G. Liquin. Explanation is effective because it is selective. *Current Directions in Psychological Science*, 32(3), 2023.
- [245] Tanya Lombrozo. Explanation and abductive inference. In *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press, 2012.

- [246] Duri Long and Brian Magerko. What is AI literacy? Competencies and design considerations. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2020.
- [247] Alexandra Lorson, Chris Cummins, and Hannah Rohde. Strategic use of (un)certainty expressions. *Frontiers in Communication*, 6, 2021.
- [248] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. Novice-AI music co-creation via ai-steering tools for deep generative models. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2020.
- [249] Lu Lu, Nathan Neale, Nathaniel D. Line, and Mark Bonn. Improving data quality using Amazon Mechanical Turk through platform setup. *Cornell Hospitality Quarterly*, 63(2), 2022.
- [250] Zhuoran Lu and Ming Yin. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.
- [251] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *International Joint Conference on Natural Language Processing (IJCNLP) and Asia-Pacific Chapter of the Association for Computational Linguistics (AAACL)*, 2023.
- [252] Jörn Lötsch, Dario Kringel, and Alfred Ultsch. Explainable artificial intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients. *BioMedInformatics*, 2(1), 2022.
- [253] R. Machlev, L. Heistrene, M. Perl, K.Y. Levy, J. Belikov, S. Mannor, and Y. Levron. Explainable artificial intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, 9, 2022.
- [254] B. F. Malle and Joshua Knobe. Which behaviors do people explain? A basic actor–observer asymmetry. *Journal of Personality and Social Psychology*, 72 (2), 1997.
- [255] Antonios Mamalakis, Imme Ebert-Uphoff, and Elizabeth A. Barnes. *Explainable Artificial Intelligence in Meteorology and Climate Science: Model Fine-Tuning, Calibrating Trust and Learning New Science*. Springer International Publishing, 2022.
- [256] Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E. Peters. Few-shot self-rationalization with natural language prompts. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.
- [257] Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? In *International Conference on Learning Representations (ICLR) Workshops*, 2021.

- [258] Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics*, 113, 2021.
- [259] Catherine C. Marshall, Partha S.R. Goguladinne, Mudit Maheshwari, Apoorva Sathe, and Frank M. Shipman. Who broke Amazon Mechanical Turk? An analysis of crowdsourcing data quality over time. In *ACM Web Science Conference (WebSci)*, 2023.
- [260] Winter Mason and Siddharth Suri. Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior Research Methods*, 44(1), 2012.
- [261] Roger C. Mayer, James H. Davis, and F. David Schoorman. An integrative model of organizational trust. *The Academy of Management Review*, 20(3), 1995.
- [262] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2019.
- [263] Joseph E McGrath. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human-Computer Interaction*. Elsevier, 1995.
- [264] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. Taking a HIT: Designing around rejection, mistrust, risk, and workers’ experiences in Amazon Mechanical Turk. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2016.
- [265] Yusuf Mehdi. The new Bing and Edge: Progress from our first month, 2023. URL https://blogs.bing.com/search/march_2023/The-New-Bing-and-Edge-%E2%80%93-Momentum-from-Our-First-Month.
- [266] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. Algorithmic impact assessments and accountability: The co-construction of impacts. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.
- [267] Quentin Metteier, Marine Capallera, Leonardo Angelini, Elena Mugellini, Omar Abou Khaled, Stefano Carrino, Emmanuel De Salis, Stéphane Galland, and Susanne Boll. Workshop on explainable AI in automated driving: A user-centered interaction approach. In *International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings (AutomotiveUI)*, 2019.
- [268] Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents’ overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics (TACL)*, 2022.

- [269] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. 2017.
- [270] Tim Miller. Are we measuring trust correctly in explainability, interpretability, and transparency research?, 2022.
- [271] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. 2017.
- [272] Candice M. Mills, Judith H. Danovitch, Sydney P. Rowles, and Ian L. Campbell. Children’s success at detecting circular explanations and their interest in future learning. *Psychonomic Bulletin & Review*, 24(5), 2017.
- [273] Mitch Waite Group. iBird Pro Guide.
- [274] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019.
- [275] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019.
- [276] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A survey of evaluation methods and measures for interpretable machine learning. 2018.
- [277] Sina Mohseni, Fan Yang, Shiva Pentyala, Mengnan Du, Yi Liu, Nic Lupfer, Xia Hu, Shuiwang Ji, and Eric Ragan. Machine learning explanations to prevent overtrust in fake news detection. *AAAI Conference on Web and Social Media*, 2021.
- [278] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.
- [279] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning: Problems, methods and evaluation. *SIGKDD Explor. Newsl.*, 22(1), 2020.
- [280] Kristin D. Hansen Morten Hertzum and Hans H.K. Andersen. Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, 28(2), 2009.
- [281] Hussein Mozannar, Gagan Bansal, Adam Journey, and Eric Horvitz. When to show a suggestion? Integrating human feedback in AI-assisted programming. *AAAI Conference on Artificial Intelligence*, 2024.
- [282] Michael J. Muller. *Participatory Design: The Third Space in HCI*. L. Erlbaum Associates Inc., 2002.

- [283] Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1), 2000.
- [284] National Audubon Society. Audubon bird guide.
- [285] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [286] An T. Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. Believe it or not: Designing a human-AI partnership for mixed-initiative fact-checking. In *ACM Symposium on User Interface Software and Technology (UIST)*, 2018.
- [287] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In *Neural Information Processing Systems (NeurIPS)*, 2021.
- [288] Giang Nguyen, Mohammad Reza Taesiri, and Anh Nguyen. Visual correspondence-based explanations improve AI robustness and human-AI team accuracy. In *Neural Information Processing Systems (NeurIPS)*, 2022.
- [289] Kyungjoo Noh. Quantum computation and communication in bosonic systems. *Yale University PhD thesis*, 2020.
- [290] Zahra Nouri, Ujwal Gadiraju, Gregor Engels, and Henning Wachsmuth. What is unclear? Computational assessment of task clarity in crowdsourcing. In *ACM Conference on Hypertext and Social Media (HT)*, 2021.
- [291] Takashi Numata, Hiroki Sato, Yasuhiro Asa, Takahiko Koike, Kohei Miyata, Eri Nakagawa, Motofumi Sumiya, and Norihiro Sadato. Achieving affective human–virtual agent communication by enabling virtual agents to imitate positive expressions. *Scientific Reports*, 10(1), 2020.
- [292] Demi Oba and Jonah A. Berger. How hedges impact persuasion, 2022.
- [293] Uchenna Emeoha Ogenyi, Jinguo Liu, Chenguang Yang, Zhaojie Ju, and Hong-hai Liu. Physical human–robot collaboration: Robotic systems, learning methods, collaborative strategies, sensors, and actuators. *IEEE Transactions on Cybernetics*, 51(4), 2021.
- [294] Daniel Omeiza, Helena Webb, Marina Jirotka, and Lars Kunze. Explanations in autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(8), 2022.
- [295] OpenAI. GPT-4 technical report, 2023.

- [296] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Neural Information Processing Systems (NeurIPS)*, 2024.
- [297] Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. The shifted and the overlooked: A task-oriented investigation of user-GPT interactions. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
- [298] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. Human-autonomy teaming: A review and analysis of the empirical literature. *Human Factors*, 64(5), 2022.
- [299] Marvin Pafla, Kate Larson, and Mark Hancock. Unraveling the dilemma of AI errors: Exploring the effectiveness of human and machine explanations for large language models. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2024.
- [300] Yikang Pan, Liangming Pan, Wenhua Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. On the risk of misinformation pollution with large language models. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
- [301] Gabriele Paolacci and Jesse Chandler. Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23 (3), 2014.
- [302] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5), 2010.
- [303] European Parliament. European union artificial intelligence act corrigendum., 2024.
- [304] Samir Passi and Mihaela Vorvoreanu. Overreliance on AI: Literature review. Technical Report MSR-TR-2022-12, Microsoft, 2022.
- [305] Michael Pazzani, Severine Soltani, Robert Kaufman, Samson Qian, and Albert Hsiao. Expert-informed, user-centric explanations for machine learning. *AAAI Conference on Artificial Intelligence*, 36(11), 2022.
- [306] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. Meaningful explanations of black box AI decision systems. *AAAI Conference on Artificial Intelligence*, 2019.

- [307] Robert Perkins. A birder in the hand: Mobile phone app can recognize birds from photos, 2016. URL <https://www.caltech.edu/about/news/birder-h-and-mobile-phone-app-can-recognize-birds-photos-53288>.
- [308] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018.
- [309] Barbara Pintar. Birdadvisor 360: A digital support for birdwatching tourism in Algarve. 2017.
- [310] Milda Pocevičiūtė, Gabriel Eilertsen, and Claes Lundström. *Survey of XAI in Digital Pathology*. Springer International Publishing, 2020.
- [311] Samuele Poppi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.
- [312] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.
- [313] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. Understanding uncertainty: How lay decision-makers perceive and interpret uncertainty in human-AI decision making. In *International Conference on Intelligent User Interfaces (IUI)*, 2023.
- [314] Alun D. Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. Stakeholders in explainable AI. 2018.
- [315] Rohith Pudari and Neil A. Ernst. From Copilot to pilot: Towards AI supported software development, 2023.
- [316] Filip Radenovic, Abhimanyu Dubey, and Dhruv Mahajan. Neural basis models for interpretability. In *Neural Information Processing Systems (NeurIPS)*, 2022.
- [317] Marissa Radensky, Julie Anne Séguin, Jang Soo Lim, Kristen Olson, and Robert Geiger. “I think you might like this”: Exploring effects of confidence signal patterns on trust in and reliance on conversational recommender systems. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.
- [318] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Nicole Meister, Ruth Fong, and Olga Russakovsky. ELUDE: Generating interpretable explanations via a decomposition into labelled and unlabelled features, 2022.

- [319] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, and Olga Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept salience, and human capability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [320] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should I trust you?”: Explaining the predictions of any classifier. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.
- [321] Soo Young Rieh and David R. Danielson. Credibility: A multidisciplinary framework. *Annual Review of Information Science and Technology*, 41(1), 2007.
- [322] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers? Shifting demographics in mechanical turk. In *ACM Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, 2010.
- [323] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16(none), 2022.
- [324] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015.
- [325] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. In *Neural Information Processing Systems (NeurIPS) Workshops*, 2023.
- [326] Johnny Saldaña. The coding manual for qualitative researchers. *The coding manual for qualitative researchers*, 2021.
- [327] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. A missing piece in the puzzle: Considering the role of task complexity in human-AI decision making. In *ACM Conference on User Modeling, Adaptation and Personalization (UMAP)*, 2023.
- [328] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. Dealing with uncertainty: Understanding the impact of prognostic versus diagnostic tasks on trust and reliance in human-AI decision making. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2024.
- [329] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Re-imagining algorithmic fairness in India and beyond. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

- [330] Lindsay Sanneman and Julie A. Shah. A situation awareness-based framework for design and evaluation of explainable AI. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer International Publishing, 2020.
- [331] Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation, 2022.
- [332] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your AI: Expertise and explanations. In *International Conference on Intelligent User Interfaces (IUI)*, 2019.
- [333] Nicolas Scharowski, Sebastian A. C. Perrig, Nick von Felten, and Florian Brühlmann. Trust and reliance in XAI: Distinguishing between attitudinal and behavioral measures, 2022.
- [334] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2022.
- [335] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *International Conference on Intelligent User Interfaces (IUI)*, 2023.
- [336] Nadine Schlicker, Alarith Uhde, Kevin Baum, Martin C Hirsch, and Markus Langer. Calibrated trust as a result of accurate trustworthiness assessment: Introducing the trustworthiness assessment model, 2022.
- [337] Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. How to do human evaluation: A brief introduction to user studies in NLP. *Natural Language Engineering*, 29(5), 2023.
- [338] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *ACM Conference on Fairness, Accountability, and Transparency*, 2019.
- [339] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision (ICCV)*, 2017.
- [340] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. “The human body is a black box”: Supporting clinical decision-making with

- deep learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.
- [341] Chirag Shah and Emily M. Bender. Situating search. In *Conference on Human Information Interaction and Retrieval (CHIIR)*, 2022.
 - [342] Murray Shanahan. Talking about large language models. *Commun. ACM*, 67(2), 2024.
 - [343] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. Generative echo chamber? Effect of llm-powered search systems on diverse information seeking. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2024.
 - [344] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N’Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2023.
 - [345] Hua Shen and Ting-Hao Kenneth Huang. How useful are the machine-generated interpretations to general users? A human evaluation on guessing the incorrectly predicted labels. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2020.
 - [346] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning (ICML)*, 2023.
 - [347] Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. One explanation is not enough: Structured attention graphs for image classification. In *Neural Information Processing Systems (NeurIPS)*, 2021.
 - [348] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Association for Computational Linguistics (ACL) Findings*, 2021.
 - [349] Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé Iii, and Jordan Boyd-Graber. Large language models help humans verify truthfulness – except when they are convincingly wrong. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.
 - [350] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (ICLR) Workshops*, 2014.

- [351] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6, 2020.
- [352] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. TalkToModel: Understanding machine learning models with open ended dialogues, 2022.
- [353] Mona Sloane and Janina Zakrzewski. German AI start-ups and “AI ethics”: Using a social practice lens for assessing and implementing socio-technical innovation. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.
- [354] Helen Smith. Clinical AI: opacity, accountability, responsibility and liability. *AI & Society*, 36(2), 2021.
- [355] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2020.
- [356] Nathalie A. Smuha. The EU approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International*, 20(4), 2019.
- [357] Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. Comparing traditional and llm-based search for consumer choice: A randomized experiment, 2023.
- [358] Ramya Srinivasan and Ajay Chander. Explanation perspectives from the cognitive sciences—a survey. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.
- [359] Mark Sullivan. Is perplexity AI showing us the future of search?, 2023. URL <https://www.fastcompany.com/90883562/is-perplexity-ai-showing-u-s-the-future-of-search>.
- [360] Cass R. Sunstein. Probability neglect: Emotions, worst cases, and law. *Yale Law Journal*, 112(61), 2002.
- [361] Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.
- [362] Elham Tabassi. Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.

- [363] Sri Yash Tadimalla and Mary Lou Maher. AI literacy for all: Adjustable interdisciplinary socio-technical curriculum. *IEEE Frontiers in Education Conference*, 2024.
- [364] Mohammad Tahaei, Marios Constantinides, Daniele Quercia, Sean Kennedy, Michael Muller, Simone Stumpf, Q. Vera Liao, Ricardo Baeza-Yates, Lora Aroyo, Jess Holbrook, Ewa Luger, Michael Madaio, Ilana Golbin Blumenfeld, Maria De-Arteaga, Jessica Vitak, and Alexandra Olteanu. Human-centered responsible artificial intelligence: Current and future trends. In *ACM Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, 2023.
- [365] Iddo Tavory and Stefan Timmermans. *Abductive analysis: Theorizing qualitative research*. University of Chicago Press, 2014.
- [366] J. Eric T. Taylor and Graham W. Taylor. Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, 28(2), 2018.
- [367] The Cornell Lab of Ornithology. eBird, .
- [368] The Cornell Lab of Ornithology. Merlin Bird ID, .
- [369] Lauren Thornton, Bran Knowles, and Gordon Blair. The alchemy of trust: The creative act of designing trustworthy socio-technical systems. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.
- [370] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailev, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
- [371] Richard Tomsett, Dave Braines, Dan Harborne, Alun D. Preece, and Supriyo Chakraborty. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. 2018.
- [372] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Machine Learning for Healthcare Conference (MLH4H)*, 2019.
- [373] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. The relationship between trust in AI and trustworthy machine learning technologies. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.
- [374] Cristen Torrey, Susan R. Fussell, and Sara Kiesler. How a robot should give advice. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2013.

- [375] J.D. Trout. Seduction without cause: Uncovering explanatory neurophilia. *Trends in Cognitive Sciences*, 12(8), 2008.
- [376] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. Human-computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8), 2020.
- [377] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are convolutional neural networks or transformers more like human vision? 2021.
- [378] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Neural Information Processing Systems (NeurIPS)*, 2023.
- [379] Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. Trust in human-AI interaction: Scoping out models, measures, and methods. In *ACM Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, 2022.
- [380] Shimon Ullman, Liav Assif, Ethan Fetaya, and Daniel Harari. Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences*, 113(10), 2016.
- [381] United States Census Bureau. 2020 census of population and housing. Website, 2020.
- [382] United States Census Bureau. Educational attainment in the United States, 2022.
- [383] Simon Vandenhende, Dhruv Mahajan, Filip Radenovic, and Deepti Ghadiyaram. Making heads or tails: Towards semantically consistent visual counterfactuals. In *European Conference on Computer Vision (ECCV)*, 2022.
- [384] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on AI systems during decision-making. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2023.
- [385] Helena Vasconcelos, Gagan Bansal, Adam Journey, Q. Vera Liao, and Jennifer Wortman Vaughan. Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in AI-powered code completions. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 2024.

- [386] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. How to evaluate trust in AI-assisted decision making? a survey of empirical methodologies. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2021.
- [387] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. Artificial artificial intelligence: Crowd workers widely use large language models for text production tasks, 2023.
- [388] Veniamin Veselovsky, Manoel Horta Ribeiro, Philip J. Cozzolino, Andrew Gordon, David Rothschild, and Robert West. Prevalence and prevention of large language model use in crowd work. *Commun. ACM*, 68(3), 2025.
- [389] Valeria Villani, Fabio Pini, Francesco Leali, and Cristian Secchi. Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics*, 55, 2018.
- [390] H. Vogel, S. Appelbaum, H. Haller, and T. Ostermann. The interpretation of verbal probabilities: A systematic literature review and meta-analysis. *Studies in Health Technology and Informatics*, 296, 2022.
- [391] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
- [392] Thomas S. Wallsten, David V. Budescu, Rami Zwick, and Steven M. Kemp. Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the psychonomic society*, 31, 1993.
- [393] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity, 2023.
- [394] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. Human-AI collaboration in data science: Exploring data scientists’ perceptions of automated AI. In *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2019.
- [395] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing theory-driven user-centric explainable AI. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2019.
- [396] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

- [397] Pei Wang and Nuno Vasconcelos. Towards realistic predictors. In *European Conference on Computer Vision (ECCV)*, 2018.
- [398] Pei Wang and Nuno Vasconcelos. SCOUT: Self-aware discriminant counterfactual explanations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [399] Weiquan Wang and Izak Benbasat. Attributions of trust in decision support technologies: A study of recommendation agents for e-commerce. *Journal of Management Information Systems*, 24(4), 2008.
- [400] Xinru Wang and Ming Yin. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *International Conference on Intelligent User Interfaces (IUI)*, 2021.
- [401] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.
- [402] C. Nadine Wathen and Jacquelyn Burkell. Believe it or not: Factors influencing credibility on the web. *Journal of the American Society for Information Science and Technology*, 53(2), 2002.
- [403] Margaret A. Webb and June P. Tangney. Too good to be true: Bots and bad data from mechanical turk. *Perspectives on Psychological Science*, 0(0), 0.
- [404] Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. How interpretable are reasoning explanations from prompting large language models? In *Association for Computational Linguistics (ACL) Findings*, 2024.
- [405] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.
- [406] Deena Skolnick Weisberg, Jordan C.V Taylor, and Emily J Hopkins. Deconstructing the seductive allure of neuroscience explanations. *Judgment and Decision Making*, 10(5), 2015.
- [407] Benjamin Weiser and Nate Schweber. The ChatGPT lawyer explains himself. *New York Times*, 2023. URL <https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html>.
- [408] Henry M. Wellman. Reinvigorating explanations for the study of early cognitive development. *Child Development Perspectives*, 5(1), 2011.

- [409] Caroline J. Wesson and Briony D. Pulford. Verbal expressions of confidence and doubt. *Psychological Reports*, 105(1), 2009.
- [410] David Gray Widder, Laura Dabbish, James D. Herbsleb, Alexandra Holloway, and Scott Davidoff. Trust in collaborative automation in high stakes software engineering work: A case study at NASA. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.
- [411] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-AI collaboration for generating free-text explanations. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.
- [412] Paul D Windschitl and Gary L Wells. Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, 2(4), 1996.
- [413] Magdalena Wischniewski, Nicole Krämer, and Emmanuel Müller. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *ACM Conference on Human Factors in Computing Systems (CHI)*.
- [414] Jennifer Wortman Vaughan and Hanna Wallach. A human-centered agenda for intelligible machine learning. In *Machines We Trust: Perspectives on Dependable AI*. MIT Press, 2021.
- [415] Meng-Han Wu and Alexander Quinn. Confusing the crowd: Task instruction quality on Amazon Mechanical Turk. *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 5(1), 2017.
- [416] Roy Xie, Chengxuan Huang, Junlin Wang, and Bhuwan Dhingra. Adversarial math word problem generation, 2024.
- [417] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations (ICML)*, 2024.
- [418] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering. In *Association for Computational Linguistics (ACL) Long Papers*, 2023.
- [419] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. 2019.
- [420] Guang Yang, Qinghao Ye, and Jun Xia. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, 77, 2022.

- [421] Mengjiao Yang and Been Kim. Benchmarking attribution methods with relative feature importance, 2019.
- [422] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2018.
- [423] Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho De Castro, Shruthi Bannur, Stephanie Hyland, Pratik Ghosh, Mercy Ranjit, Kenza Bouzid, Anton Schwaighofer, Fernando Pérez-García, Harshita Sharma, Ozan Oktay, Matthew Lungren, Javier Alvarez-Valle, Aditya Nori, and Anja Thieme. Multimodal healthcare AI: Identifying and designing clinically relevant vision-language applications for radiology. *ACM Conference on Human Factors in Computing Systems (CHI)*, 2024.
- [424] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2019.
- [425] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2019.
- [426] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. Do I trust my machine teammate? An investigation from perception to decision. In *International Conference on Intelligent User Interfaces (IUI)*, 2019.
- [427] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. Wordcraft: Story writing with large language models. In *International Conference on Intelligent User Interfaces (IUI)*, 2022.
- [428] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, 2014.
- [429] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *Neural Information Processing Systems (NeurIPS)*, 2019.
- [430] Jeffrey C. Zemla, Steven Sloman, Christos Bechlivaniidis, and David A. Lagnado. Evaluating everyday explanations. *Psychonomic Bulletin & Review*, 24(5), 2017.
- [431] Angie Zhang, Alexander Boltz, Jonathan Lynn, Chun-Wei Wang, and Min Kyung Lee. Stakeholder-centered AI design: Co-designing worker tools with gig workers through data probes. *ACM Conference on Human Factors in Computing Systems (CHI)*, 2023.

- [432] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision (ECCV)*, 2016.
- [433] Peng Zhang, Jiuling Wang, Ali Farhadi, Martial Hebert, and Devi Parikh. Predicting failures of vision systems. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [434] Qiaoning Zhang, Matthew L Lee, and Scott Carter. You complete me: Human-AI teams and complementary expertise. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.
- [435] Yiming Zhang, Ying Weng, and Jonathan Lund. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics*, 12(2), 2022.
- [436] Yixuan Zhang, Joseph D Gaggiano, Nutchanon Yongsatianchot, Nurul M Suhaimi, Miso Kim, Yifan Sun, Jacqueline Griffin, and Andrea G Parker. What do we mean when we talk about trust in social media? A systematic review. In *ACM CHI Conference on Human Factors in Computing Systems*, 2023.
- [437] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect on confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.
- [438] Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. Storybuddy: A human-AI collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.
- [439] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2), 2024.
- [440] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-bench and Chatbot Arena. In *Neural Information Processing Systems (NeurIPS)*, 2024.
- [441] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [442] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *European Conference on Computer Vision (ECCV)*, 2018.

- [443] Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
- [444] Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty, 2024.
- [445] Sharon Zhou, Mitchell L Gordon, Ranjay Krishna, Austin Narcomey, Li Fei-Fei, and Michael S Bernstein. HYPE: A benchmark for human eye perceptual evaluation of generative models. In *Neural Information Processing Systems (NeurIPS)*, 2019.
- [446] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji rong Wen. Large language models for information retrieval: A survey. 2023.
- [447] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. 2019.
- [448] Alf C. Zimmer. Verbal vs. numerical processing of subjective probabilities. In *Decision Making Under Uncertainty*, volume 16 of *Advances in Psychology*. North-Holland, 1983.
- [449] Roland S. Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas S. A. Wallis, and Wieland Brendel. How well do feature visualizations support causal understanding of CNN activations? In *Neural Information Processing Systems (NeurIPS)*, 2021.