Expanding the Design Space for AI Explanations in Human–AI Collaboration

Katelyn Christina Morrison

CMU-HCII-26-101

April 2026

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Adam Perer, Chair, CMU HCII
Motahhare Eslami, CMU HCII
John Zimmerman, CMU HCII
Xiang 'Anthony' Chen, UCLA HCI

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research has been partially or fully funded by the Software Engineering Institute, the Center for Machine Learning and Health, the Office of Graduate & Postdoctoral Affairs, and the NIH.

Copyright © 2026 **Katelyn Christina Morrison**

Keywords: Human–Computer Interaction, Human–AI Collaboration, Explainable AI, Decision-Making, Empirical Studies, Generative AI, Vision–Language Models

For Sunshine (my dad),
for his unconditional love, bravery, and support.
Rest in peace, Marine. February 23rd, 2026.

For my mom,
for teaching me how to navigate life
with grace and strength.

For Jenn, Scott, and Brett,
for always being there for me.

Abstract

Explainable AI (XAI) methods have traditionally aimed to provide people with an understanding of why and how AI systems behave. Recognizing the potential benefits this can have for AI-assisted decision-making, researchers, media, and funding agencies have turned to XAI as a solution to help people validate, interpret, or justify AI outputs in high-stakes decision-making contexts. We refer to this paradigm as output-oriented XAI. While some controlled studies have shown that advancements in XAI techniques can lead to improved human–AI decision-making, other studies have demonstrated that these efforts have resulted in negative outcomes. Human-centered AI researchers argue that these inconsistent findings stem from a "one-size-fits-all" approach that ignores decision-makers' diverse expertise, roles, and information needs. Building on the one-size-fits-all critique, this thesis argues that these mixed results stem from two additional limitations: the assumption that XAI methods lack imperfections and the narrowly scoped design space centered on output-oriented XAI. The existing focus on designing output-oriented XAI risks overlooking the emerging uses of AI explanations that can support complex, human–AI decision-making processes beyond validating, interpreting, or justifying AI outputs.

To address the ongoing challenges that XAI faces in human–AI collaboration, Part I of this thesis presents two empirical studies with crowdworkers and domain stakeholders that expose the impacts of existing output-oriented XAI designs. These studies show how such explanations can be misaligned, misleading, or imperfect, shaping decision-makers' reliance on AI. Although these findings may suggest XAI is not well-suited for human–AI collaboration, Part II presents two exploratory studies that investigate the emerging uses of AI explanations in two different human–AI decision-making contexts. Across two different AI systems, these studies highlight three emerging uses of AI explanations: supporting sensemaking, augmenting communication, and facilitating knowledge transfer. These emerging uses reveal needs that AI explanations could support, but that are not well supported by existing output-oriented XAI designs. However, recent work suggests that generative AI (GenAI) may be able to support such needs. Therefore, Part III first examines the explanatory potential of multimodal GenAI in medical imaging interpretation workflows. Then, the second half of Part III presents process-oriented XAI, a reframing of the XAI design space that positions XAI techniques as a support tool for users throughout workflows. We build a process-oriented XAI system for medical imaging interpretation and investigate its impact on the clinical reasoning processes of medical students and radiology residents.

This thesis expands the design space for AI explanations by acknowledging the impact of their imperfections, exploring their emerging uses, and proposing a process-oriented XAI paradigm. Across six empirical studies, this work challenges traditional conceptions of XAI and advances our understanding of how to design for, evaluate, and think about XAI in human–AI collaborations. In conclusion, we encourage researchers to design XAI for uses beyond validation, interpretation, or justification of AI outputs.

Acknowledgments

As a child, my dad often said I was always asking questions, seeking explanations for everything, to the point that he had to limit the number of questions I could ask him each day. It is hard to say for sure, but it is possible that my ongoing search for an explanation for everything is part of the reason I gravitated toward researching methods for explaining AI or using AI to help explain abstract concepts. Nonetheless, the completion of this thesis is thanks to an incredible group of mentors, friends, collaborators, and believers.

First, I would like to thank my advisor, Adam Perer, for his unwavering support, guidance, and trust, as well as for giving me the space to learn and grow as a researcher. Coming from a family where no one had pursued graduate education, I feel incredibly fortunate to have had Adam as my advisor over the last five years. He made sure I had the necessary connections and collaborations to be successful in the domains I pursued for my projects throughout the PhD. He has helped me clarify what rigorous research in human–computer interaction looks like, and he has been instrumental in shaping me into a confident, independent researcher. I am thankful for our numerous conversations and brainstorming sessions questioning the purpose of explainable AI (XAI), which ultimately set the direction for my thesis work.

I am also deeply grateful to my committee for their thoughtful feedback, guidance, and support throughout the process of bringing this thesis together. John Zimmerman's mini course on Design Perspectives in HCI and his thought-provoking discussions, as I was putting my proposal together, gave me confidence to challenge the status-quo in my domain. Our discussions significantly contributed to the reframing of XAI that I argue for in this thesis. I also have to thank Motahhare Eslami for bringing the radiology project (Chapters 7–8) to Adam and me, which was initially intended as a small model-evaluation study with radiologists. Motahhare's enthusiasm and brainstorming during every meeting not only energized me but also helped me push through with my ideas, even when I sometimes doubted them myself. Lastly, Xiang 'Anthony' Chen's feedback and guidance during the development of my proposed work gave me the confidence to embrace my unique perspectives on XAI. Thank you to the entire committee for your support and guidance.

Throughout my Ph.D., I was fortunate to complete several industry internships at Microsoft Research and IBM Research. Although the work from these internships does not appear directly in this thesis, these experiences had a lasting impact on my technical and research skills, which undoubtedly contributed to the success of my thesis. I would therefore like to thank Eric Horvitz, Shamsi Iqbal, Justin Weisz, Zahra Ashktorab, Djallel Bouneffouf, Maria Chang, Andrew Anderson, Qiaosi (Chelsea) Wang, Michael Muller, Matt Reimer, and Gabriel Gonzalez.

Along with my academic and industry collaborators, my family has played a significant role in my determination to finish my Ph.D. First, I would like to give a special acknowledgment to my dad, whom those closest to me know I lovingly

called “Sunshine”. Above all the lessons he shared with me throughout the 27 years I was fortunate enough to have with him, he taught me what strength, bravery, and perseverance truly look like. Given that a Ph.D. is a long and difficult journey often filled with rejection, perseverance became especially important. He proudly shared my thesis work, particularly Chapters 6–8, with his physicians during his many doctor visits. Chapters 7 and 8 of this thesis are especially dedicated to him, in the hope that one day no child will have to lose a parent to the devastating disease of lung cancer.

Second, I would like to acknowledge the importance of my mom’s unwavering support and inspiration. Her “daily inspiration” messages have been deeply meaningful (and remarkably timely), helping me remain persistent, grounded, and determined throughout my Ph.D. Beyond those daily messages, she has shown me unconditional love and support in countless ways, from proofreading my first paper to consoling me through moments of worry and rejection. She taught me the importance of allowing myself a set amount of time to worry or feel sad before moving forward with the day, and she has shown me how to navigate life’s tragedies with grace.

I would also like to acknowledge my Aunt Kate for the career coaching she provided throughout my Ph.D. She helped me prepare for countless interviews, from internships to full-time roles, and guided me through many difficult conversations along the way (often by prescribing *Difficult Conversations* by Douglas Stone, Bruce Patton, and Sheila Heen). Among the many pieces of advice she has given me, one that has stayed with me is: “*You don’t know what you don’t know.*” Research often begins in that very space of uncertainty, and she has encouraged me to embrace the unknown while continuing to move forward.

Lastly, I cannot go without thanking my siblings, my niece and nephews, my in-laws, and my soon-to-be husband (Álvaro Fernández García) for surrounding me with love, patience, laughter, and strength.

Beyond the unwavering support of my family, many once-strangers-now-colleagues, friends, and mentors helped make this winding path far easier to navigate. While some of these individuals were present for only a portion of my Ph.D., others remain close, and I hope they continue to be part of my life beyond my time at Carnegie Mellon. As I reflect on everyone I have interacted with over the past five years, I am undoubtedly influenced by recency bias as I write this, so I apologize if I fail to acknowledge your impact on my life or on this thesis.

Early in my Ph.D., I attended my first human–computer interaction conference as a student volunteer (CHI 2023, Germany), where I met Philipp Spitzer. We quickly discovered shared research interests (at the time: “birding” in XAI) and have since collaborated on several successful papers and workshops. Also, a shoutout to Niklas Kühl, whom Philipp introduced me to and became a close collaborator and mentor. Additionally, much of my thesis work is thanks to my numerous undergraduate and Master’s research assistants (including, but not limited to, Zexuan Li, Isabelle Rhee, Michelle Feng, Ray Xia, Aidan Bradshaw, Maggie Chen, and Tom Wartmann) as well as my clinical collaborators who have played a significant role in helping me recruit

participants for my studies. I would also like to thank all of the collaborators with whom I have worked over the last five years.

I would like to thank Rushil Khurana for first introducing me to human–computer interaction research and encouraging me to apply to the HCII Ph.D. program; all of the DIG Lab Ph.D. students (Will Epperson, Alex Cabrera, Venkat Sivaraman, Arpit Mathur, Eric Mason, Frank Elavsky, and Hwei-Shin Harriman) for the company down in the basement offices of Newell–Simon Hall; my travel companions and life support group (Lisa Egede, Sanika Moharana, Adinawa Adjagbodjou, and Anastasia Sosnovski) for traveling the world with me during and after conferences and giving me a space to gossip about reality TV; my Duolingo family for keeping me accountable with my Spanish learning streak; my close friends from the Ph.D. program (Anna Kawakami, Kimi Wenzel, Wesley Deng, Luke Guerdan, Yi Fei Cheng, Ezra Awumey, Alicia DeVrio, Hugo Simão, and Jordan Taylor) for making every house party more enjoyable; and my birthday twins (Emanuel Tewolde and Juan Rodolfo Alvarez-Padilla) for sharing your birthdays with me and throwing two unforgettable Halloween birthday bashes.

Use of Generative AI Disclosure. *AI tools (such as Grammarly, Microsoft Copilot, and OpenAI GPT) were used to check for grammar and spelling mistakes; improve the clarity or flow of sentences or ideas; and to remove duplicate references from the bibliography throughout the chapters in this thesis. Any text that was generated by an AI tool was reviewed and modified to represent my own voice and interpretations.*

Acronyms

ADS Algorithmic Decision Support
AI Artificial Intelligence
CDSS Clinical Decision-Support System
CSCW Computer-Supported Cooperative Work
GenAI Generative AI
GenXAI Generative eXplainable AI
HAIC Human–AI Collaboration
HCI Human–Computer Interaction
HCXAI Human-centered Explainable AI
T2I Text-to-Image
VLM Vision–Language Model
XAI eXplainable AI

Table of Contents

Keywords ii

Abstract v

Acknowledgments vii

Acronyms xi

List of Figures xviii

List of Tables xxiv

1 Introduction 1
- 1.1 Thesis Motivation . . . 3
- 1.2 Process-Oriented XAI Paradigm . . . 4
- 1.3 Thesis Statement . . . 6
- 1.4 Thesis Overview . . . 6
- 1.5 Prior Publications and Authorship . . . 8

2 Background & Related Work 10
- 2.1 Explainable AI (XAI) Techniques . . . 11
- 2.2 Output-oriented XAI in Human–AI Collaboration . . . 12
- 2.3 Imperfect AI/XAI . . . 13
- 2.4 Emerging Uses of AI Explanations . . . 14
- 2.5 The Explanatory Potential of Generative AI . . . 14
- 2.6 Process-Oriented (X)AI Support . . . 15

Part I: Imperfect Output-Oriented Explainable AI in Human–AI Collaboration 17

3 Human Explanation Strategies for Output-Oriented XAI 18
- 3.1 Introduction . . . 18
- 3.2 Study Context: Building Damage Assessment . . . 20
  - 3.2.1 Task Expertise . . . 21
  - 3.2.2 Task Generalization . . . 22
  - 3.2.3 Selected Dataset . . . 22
- 3.3 Study 1: Identifying Human Explanation Strategies . . . 22
  - 3.3.1 Study Design . . . 23
  - 3.3.2 Results . . . 26
- 3.4 Study 2: Impact of Human Explanation Strategies . . . 30
  - 3.4.1 Study Design . . . 31
  - 3.4.2 Quantitative Results . . . 35

3.4.3 Qualitative Results . . . 39
3.5 Discussion . . . 41
3.5.1 The Mixed Effect of Explanations . . . 42
3.5.2 Limitations . . . 43
3.6 Future work . . . 44
3.7 Conclusion . . . 45

4 The Role of Human Factors on Decision-Making with Imperfect XAI 46
4.1 Introduction . . . 46
4.2 Theoretical Development . . . 50
4.3 Methodology . . . 53
4.3.1 Task Domain: Bird Species Identification . . . 53
4.3.2 Study Design . . . 54
4.3.3 Data Selection . . . 56
4.3.4 Explanation Modalities . . . 58
4.3.5 Recruitment . . . 60
4.3.6 Quantitative and Qualitative Metrics . . . 60
4.4 Results . . . 62
4.4.1 Participant Statistics . . . 62
4.4.2 Moderating Effects in Imperfect XAI Research Model . . . 63
4.4.3 Human–AI Team Performance . . . 65
4.4.4 Deception caused by Imperfect XAI . . . 67
4.4.5 Designing for Imperfect XAI . . . 69
4.5 Discussion . . . 72
4.6 Limitations & Future Work . . . 75
4.7 Conclusion . . . 77

Part II: Emerging Uses of AI Explanations 78

5 Navigating Appropriations of XAI in a Deployed AI System 79
5.1 Introduction . . . 79
5.2 Case Study Overview . . . 81
5.2.1 Wild Me's AI Decision-Support Tool: Wildbooks . . . 81
5.2.2 Need-Finding Methods . . . 82
5.3 Findings . . . 84
5.3.1 Usage of Existing Transparency Mechanisms . . . 86
5.3.2 Challenges with the Existing Transparency Mechanisms . . . 89
5.3.3 Broader End-User Needs . . . 91
5.4 Reflections & Insights on Navigating Transparency in Deployed AI Systems . . . 93
5.5 Conclusion . . . 96

6 Exploring Emerging Uses of XAI in Clinical Decision-Making 98

6.1 Introduction . . . 98
6.2 Study Context: Pulmonary Arterial Hypertension . . . 100
6.3 Phase 1: Iterative Prototyping of XAI Designs . . . 102
6.3.1 Methods . . . 102
6.3.2 Findings: XAI Design Concepts . . . 104
6.4 Phase 2: Exploring Emerging Uses of XAI . . . 109
6.4.1 High-Fidelity AI-CDSS Design & Development . . . 109
6.4.2 Method . . . 110
6.4.3 Participant Recruitment & Analysis Approach . . . 111
6.5 Findings on Emerging Uses of XAI . . . 111
6.5.1 XAI Use in the Diagnosis Workflow . . . 112
6.5.2 XAI Use in the Treatment Planning Workflow . . . 113
6.5.3 XAI Uses in other parts of the workflow . . . 115
6.5.4 Perspectives on Trusting & Relying on the AI-CDSS . . . 116
6.6 Discussion & Limitations . . . 117
6.6.1 Designing Process-Oriented XAI to Support Emerging Uses of XAI in Clinical Decision-Making . . . 117
6.6.2 XAI for AI-CDSS Ethical Implications & Considerations . . . 120
6.6.3 Limitations & Future Work . . . 121
6.7 Conclusion . . . 122

Part III: Expanding the Design Space with Process-Oriented XAI 123

7 Explanatory Potential of Multimodal GenAI in Radiology Workflows 124
7.1 Introduction . . . 125
7.2 Formative Study . . . 126
7.2.1 Semi-structured Interviews . . . 127
7.2.2 Understanding Perceptions, Needs, & Workflows . . . 128
7.2.3 Formative Discussion Takeaways . . . 130
7.3 Study Design . . . 130
7.3.1 Text-to-CT Scan GenAI Plugin . . . 130
7.3.2 Semi-structured Interviews . . . 132
7.3.3 Participant Recruitment . . . 133
7.3.4 Analysis . . . 133
7.4 Findings . . . 134
7.4.1 Promises of T2I in Medical Education & Training . . . 134
7.4.2 Promises of T2I in Clinical Practice . . . 137
7.4.3 Risks & Challenges with Medical T2I GenAI . . . 138
7.5 Discussion . . . 141
7.5.1 Evaluating Domain-Specific Text-to-Image GenAI . . . 141
7.5.2 Implications of Text-to-CT Scan GenAI in Medical Education, Training, & Practice . . . 142

7.5.3 Safety Implications of Medical T2I GenAI 143
7.6 Limitations & Future Work 144
7.7 Conclusion 144

8 Augmenting Reasoning: Designing Process-Oriented Explainable AI 146
8.1 Introduction 146
8.2 Process-Oriented XAI System 149
8.2.1 Design Goals 149
8.2.2 System Overview 150
8.2.3 Translating Reasoning Processes to GenXAI 151
8.3 User Study 154
8.3.1 Protocol 154
8.3.2 Patient Vignettes 155
8.3.3 Study Interface 156
8.3.4 Recruitment & Data Analysis 157
8.4 Findings 157
8.4.1 RQ1: Interaction Patterns & Reasoning 158
8.4.2 RQ2: Benefits & Challenges of GenXAI 163
8.4.3 Other Challenges with GenXAI 165
8.5 Discussion 165
8.6 Limitations & Future Work 166
8.7 Conclusion 167

Part IV: Conclusions 168

9 Discussion & Conclusion 169
9.1 Summary of Research Contributions 169
9.1.1 Part I: Imperfect Output-Oriented XAI in Human–AI Collaboration 169
9.1.2 Part II: Emerging Uses of AI Explanations 169
9.1.3 Part III: Expanding the Design Space w/ Process-Oriented XAI 170
9.2 Discussion & Future Research Directions 170
9.2.1 Navigating the Benefits and Harms of Imperfect AI Explanations. 170
9.2.2 Designing AI Explanations to Align with Stakeholders' Workflow Processes. 171
9.2.3 Expanding the Design Space Responsibly. 172
9.3 Conclusion 172

References 173

List of Figures

1.1 A conceptual overview of output-oriented XAI in human–AI decision-making workflows. CT scans generated using MedSyn [380]. 2

1.2 A conceptual overview of process-oriented XAI in human–AI decision-making workflows. CT scans generated using MedSyn [380]. 5

1.3 This thesis consists of three parts: capturing the impact of imperfect, output-oriented XAI (Part I), discovering emerging uses of AI explanations in human–AI collaboration (Part II), and expanding the design space for AI explanations through process-oriented XAI with explainable generative AI (Part III). Figma Make was used to help design this figure. 7

3.1 Interface of the training session used to familiarize users with the task of building damage assessment. (a) Two satellite images are presented side-by-side, and (c) the user assesses damage based on four options: no damage, minor damage, major damage, and destroyed. For the first and second sub-sessions, (b) users are shown an overlay in the image highlighting where the damage is when hovering over the image to help them locate the building 24

3.2 Overview of the interface for the main task. Participants are presented with several different annotation tools to highlight evidence that helps explain their damage assessment. Participants provide explanations for each annotation they add and a global explanation for the overall pair of images. 25

3.3 The number of participants who used each core explanation strategy for each image set was reported by both coders. Some participants used multiple explanation strategies in their response to a given image set. 29

3.4 Examples of sub-strategies B-1, B-2, B-3, B-U, and C 30

3.5 High-level overview of the study design and data analysis done for the two studies. Participants of both studies went through the same training session. Study 1 participants assessed building damage and provided explanations for their assessment. We did a thematic analysis of the explanations provided by participants and identified four core strategies. Representative explanations of each core strategy from Study 1 participants were used in Stages 3 through 5 in the second study, while a separate set of participants assessed building damage. We analyzed how the participants assessed building damage throughout the five stages to understand how the core strategies impacted accuracy and reliance. 31

3.6 The second study consisted of five stages. The five stages are shown four times, one time for each scenario. Stage 1 shows the pre- and post-disaster image, with no information from the AI; Stage 2 adds information by showing how the AI assessed the damage; Stage 3 adds visual annotations, highlighting points or regions of the image where the AI thinks it is important to look; Stage 4 adds local text-based annotations to elaborate on each visual annotation, and Stage 5 adds a global explanation detailing why the AI made the damage assessment that it did. Each participant went through these five stages for each of the four scenarios. 34

3.7 Assessment accuracy, localization accuracy, and the participants' confidence. The facet on the y-axis shows the four different scenarios that the participants were presented with. The percent on the y-axis represents the ratio of participants (a) that got the damage assessment correct, (b) that placed the pinpoint on the damaged structure, and (c) the confidence level for correct assessments. The x-axis shows the five stages described in Figure 3.6. 36

3.8 Measurements to evaluate reliance on AI. The facet on the y-axis shows the four different scenarios that the participants were presented with. The percent on the y-axis represents the ratio of participants (a) who agreed with the AI's assessment, (b) who changed their assessment to agree with the AI's assessment, (c) who agreed with the AI's localization of the damage, and (d) who changed their pinpoint to agree with the AI's localization of the damage. The x-axis shows the five stages described in Figure 3.6. 37

4.1 Different paths that human decision-makers could follow based on receiving AI and XAI advice. This figure expands the model presented by Schemmer et al. [301] by adding an XAI advice dimension. Here, XAI advice is simplified into correct and incorrect explanations. Green check marks represent correct advice or decisions, while red x marks represent incorrect advice or decisions. 51

4.2 Research model for collaborating with imperfect XAI systems. We analyze the moderation of the level of expertise and assertiveness on the effect of the correctness of explanation on RAIR and RSR. 52

4.3 We conduct a rigorous mixed-methods study leveraging a mixed design. Before participants start the task, they are shown a screening test (A). For the human–AI bird identification task, participants are assigned an explanation modality (B). During the task, participants are shown explanations with different levels of assertiveness and different scenarios of correctness (C). Lastly, participants complete a post-survey (D). . . 54

4.4 Example of the two phases for a single bird image that a participant is shown in the study. This specifically shows a Magnolia Warbler (correct prediction, correct explanation), and this participant is assigned to the example-based explanations. For this bird, the participant is shown an assertive explanation. 56

4.5 Representative examples of the example-based and natural language explanations for each scenario: CC, CI, IC, and II. The class of the example-based images in the explanation is not shown to participants during the study. The red and green coloring on the natural language explanations was not shown during the study. This is only provided in the figure to guide the reader. The natural language explanation for the Cerulean Warbler is incorrect because this bird species does not have a grey head. The natural language explanation for the Yellow-billed Cuckoo is incorrect because this bird species is brown with a white belly, has a gold and black beak, and does not have black wings. 58

4.6 The three different language tones that an explanation could have in terms of assertiveness for the natural language and example-based explanation modality. Non-assertive explanations included the words "might be" and "appears to be". Assertive explanations included the words "definitely" and "clearly". Neutral explanations did not include any additional adjectives. 59

4.7 The average overall performance of the human, AI, and human–AI teams for identifying 12 birds. The bar chart on the left shows the performance of the non-experts, while the bar chart on the right shows the performance of the experts. 66

4.8 Average observed RAIR (correct AI advice) and RSR (incorrect AI advice) for example-based explanations (on the left side) and natural language explanations (on the right side). We show the average RAIR and RSR for both levels of expertise, as well as correct and incorrect explanations. 68

4.9 Findings from the qualitative analysis of the survey question: "Under what circumstances would you prefer assertive (*e.g.*, "definitely", "clearly") versus non-assertive (*e.g.*, "might be", "appears to be") versus neutral explanations and why?". Factors that should be taken into consideration when designing explanations for imperfect XAI systems. 70

5.1 Workflow and features represented from the Internet of Turtles Wildbook. For every encounter ("*a sighting of a single animal at a specific location and time*"), users are shown the image/s and can add metadata, such as the sex. Users can attempt to match the unidentified individual (A1) to another already in the database. Matched results appear in a new tab (B) and will show separate results for each unique image in the encounter and/or each unique model available for that species. (B1) Users can inspect a match result, which reveals a saliency map to rationalize the model's result. (B2) Users can also rely on the similarity "score" to determine the best match result. 82

5.2 Our need-finding methods consisted of meeting the ML and product teams and experimenting with the MiewID model. These three sources provided information related to model behaviors, technical limitations, and rationales for technical and design choices. We interviewed end-users and analyzed their training protocols to understand their mental models, use cases, and expectations for Wildbooks. Lastly, we analyzed the community forum posts as a window into interactions across and within stakeholder groups. 84

6.1 Single PAH patient management journey. Journey validated by PAH physicians. 101

6.2 Examples of existing PAH risk calculators. (Left) Interface for the ESC/ERS Guidelines risk assessment tool. (Middle) Non-Invasive French Criteria risk assessment tool. (Right) Reveal Lite 2 risk assessment tool. PAH physicians have referred us to these interfaces, explaining how they aid them in PAH diagnosis and treatment planning. These interfaces are offered by the PAH Initiative [74]. 101

6.3 Inherited AI-CDSS design. (Left) Visualizations designed by a subset of the clinical team to visualize how models would predict risk differently when values of the features were modified. (Right) A navigation tab for different risk prediction models and the features for each model. Each feature's value can be overwritten to create a new scenario. 103

6.4 Evolution of the Subgroup Insights Design. (Left) Inherited AI-CDSS design with donut chart visualizations showing the 'survival patient frequency index'. (Right) An iteration of Phase 1 designs showing text callouts for the instance frequency is integrated into the patient's historical risk trend line chart. 105

6.5 Evolution of Designs for Understanding Driving Factors of Risk Prediction. (Left) One iteration of the Phase 1 low-fidelity prototype design used a three-dot system to represent the influence of a feature on the prediction. (Right) Phase 2 high-fidelity prototype uses partial dependence bars to visualize the impact of features on each other and the prediction simultaneously. 106
6.6 Evolution of the Hypothesis Testing Capabilities Design. (Left) The inherited AI-CDSS design shows bar charts to compare against what-if explanations. (Middle) The Phase 1 low-fidelity prototype design adapted line charts as a historical trend to compare against what-if explanations. (Right) The Phase 2 high-fidelity prototype separates the what-if explanations from the patient's historical trend visualization, enabling emerging uses by offering a comparison of scenarios and a summary of variables modified. 108
6.7 Summary of Phase 1 and Phase 2 methods. Phase 1 consisted of iterative prototyping with feedback from key stakeholders. Informed by developed insights, we develop a high-fidelity AI-CDSS and use it as a design probe for Phase 2. Phase 2 consisted of semi-structured interviews that included a pre-probe, a probe, and a post-probe component. . . . 109
6.8 Final Phase 2 high-fidelity AI-CDSS prototype. (1) List of clinical variables used in the risk model. (1a) Historical trends graph for each quantitative clinical variable. (2) Risk model output and (2a) historical trend of predicted risk. (3) Main XAI component informed by Phase 1 findings. 110

7.1 Input/Output of the Text-to-CT Scan GenAI Model [380]. The text-to-CT Scan GenAI model that we use, called MedSyn [380], takes in a text prompt from the user to generate a 3D CT Scan. *Brain icon by Icons8.* 131
7.2 Open-sourced medical imaging interface [6] with our GenAI plugin powered by MedSyn [380]. (2) Patient Case: The first column shows a real case consistent with severe right pleural effusion (sampled from CT-RATE dataset [129, 130, 131]). (3) CT Scan Viewing Window: The second and third columns show a CT scan comparison viewing window, where a human's CT scan (from [129, 130, 131]) is on the left, and a generated CT scan from MedSyn [380] is on the right. (1) Prompt UI: The fourth column is our text-to-CT Scan GenAI plugin, allowing users to generate a CT scan. 131

8.1 Output-oriented XAI is framed as a way to help humans decide whether to accept, modify, or reject an AI's recommendation or generated artifact. Process-oriented XAI, in contrast, considers a workflow in which humans first reason without AI support, then generate AI explanations and use them to refine or shape their reasoning before finalizing their decision. 147

8.2 GenXAI Clinical Reasoning Support Tool. (1) shows the patient vignette. (2) shows the viewing window: (a) shows the vignette CT scan and (b) shows a generated CT scan with important regions. (3) features the GenXAI Tools, shown below. 151
8.3 Interface designs and example AI explanations for Similar Patients (8.3a) and Variations (8.3b). 152
8.4 Interface designs and example AI explanations from the Important Regions (8.4a) and Natural Language Q&A tools (8.4b). 153
8.5 Overview of user study protocol. Participants were assigned to either group one or group two to determine the order in which they would view the two patient vignettes. Phase 1, the Intro Phase, consisted of a 10-question survey and an interactive walkthrough tour of the interface. Phase 2, the Think-Aloud Phase, consisted of reviewing two patient vignettes and creating case presentations for each. Phase 3, the Debrief Phase, gave participants a space to reflect on the tools. 154
8.6 Slices of key findings marked by red arrows in the CT scans for the two patient vignettes. 156
8.7 Main User Interface where participants completed the user study. The navigation bar directs participants between the different phases of the study. 157
8.8 Example flow of using the GenXAI Tools to refine an impression. . . . 160
8.9 Example flow of using the GenXAI Tools to get a second opinion. . . . 161
8.10 Example flow of interactions that represent the process-oriented XAI system being used as a second-opinion tool. 162

List of Tables

3.1 Summary of the strategies the research team identified from doing a thematic analysis on the responses from Study 1. Grey rows indicate major codes, each of which is followed by the sub-codes (except for code C, which contains no sub-codes.) 26
3.2 Four different groups were created to evaluate each explanation code in each scenario. We show which code was assigned to which scenario for each group and which image set was used for that explanation code/scenario. 33
3.3 Post-hoc Tukey HSD tests for determining statistically significant differences between the participant's localalization and assessment accuracy. 35
3.4 Difference between Stage 2 and Stage 5. P-values from post-hoc Tukey HSD test for the *incorrect assessment* scenario measuring assessment agreement from Stage 2 compared to Stage 5. There are statistically significant differences between strategies A to B and A to F for the participant's assessment agreement. 39

4.1 Moderation analyses of the correctness of natural language and example-based explanations on RAIR and RSR with the level of expertise and assertiveness as moderators. The coding of assertiveness used for the moderation analyses is provided. 63

5.1 A summary of the five end-users interviewed, each using the tool for a different animal: sharks (Sharkbook), whales (Flukebook), sea bass (Giant Sea Bass Wildbook), sea turtles (Internet of Turtles), and snow leopards (WhiskerBook). We summarize their main role in their organization and motivation to use the Wildbook. 85
5.2 Wild Me's tool currently offers two transparency mechanisms: saliency maps and match scores. This table provides a link between the need-finding methods and the key uses of and challenges with the transparency mechanisms. Insights are mapped to the community forum analyses [CF], end-users' training documents [TD], end-user interviews [EU], meetings with the product team [PT], meetings with the ML team [MT], and model experiments [ML]. 87

6.1 Participant demographics for Phase 1. 104
6.2 Mapping Phase 1 insights to design decisions in the Phase 2 AI-CDSS. 108
6.3 Participant demographics for Phase 2. 111
6.4 Participants from both phases are mapped to the emerging use of XAI they imagined across the different phases in a PAH physician's journey of managing a single patient. 112

6.5 Four emerging uses of XAI that expand traditional uses, grounded in theory: multi-stakeholder collaboration, hypothesis exploration and reflection, trust and motivation building, and bridging stakeholder knowledge gaps. 118

7.1 Participant Demographics for the formative study. 128
7.2 Participant Demographics for interview sessions. *Fourth-year med student. **Third-year med student. 133
7.3 (*RQ1*) Participants mapped to which applications of text-to-CT Scan GenAI they imagined across the different phases. 134
7.4 *RQ2*: Participants mapped to which technical challenges and domain-specific risks they brought up. 139

8.1 Participant Demographics and scores for the Perception of AI in Healthcare [311] survey. 158
8.2 Categorization of how participants interacted with the GenXAI Tools. 159

Chapter 1

Introduction

Imperfect artificial intelligence (AI) systems are increasingly integrated into our personal and professional lives, from predicting a person's risk of cancer to identifying relevant publications to cite. However, these imperfect systems can be misleading to those on the receiving end. Imagine a high-stakes scenario where a radiologist collaborates with an AI system to detect the presence of lung nodules in a chest X-ray (*e.g.*, [146]). The AI incorrectly predicts that there are no lung nodules, and the rushed radiologist inappropriately relies on the AI, missing the presence of a nodule in the patient's lung. This can lead to delays in necessary follow-up imaging and treatment for the patient. However, if the radiologist were shown why the AI predicted the absence of lung nodules, they might have seen that the model focused on an irrelevant region of the X-ray where nodules would never be present. Scenarios like these illustrate why researchers, media, and funding agencies began investing in the development of techniques aimed at explaining the reasoning behind AI's predictions [34, 124].

Starting out as explanation facilities and explanation generators – built-in components that could trace and present a system's reasoning process – in expert systems (an earlier term for AI-driven decision-support systems) [333], explanations were conceived as a way to foster trust in, calibrate reliance on, and support users' understanding of expert systems. As expert systems evolved from interpretable rule-based models to opaque black-box AI systems (like those we mainly interact with today), explanation facilities gave way to more advanced mathematical techniques. For example, researchers began visualizing the activated gradients for a given label during image classification [307], as in the X-ray scenario above, to explain model behavior in place of the rule-based reasoning traces provided by explanation facilities. This shift marked the emergence of what is now known as explainable AI (XAI) [9, 91].

The initial hope with XAI was that, if we could explain black-box AI systems, people could use AI more effectively and safely. One prominent definition of XAI builds on this hope by defining it as "*a set of processes and methods that allow human users to comprehend and trust the results and output created by machine learning algorithms*" [4]. Guided by this definition, XAI research efforts have focused on a wide range of applications, such as assisting data scientists with debugging and auditing black-box models as part of their evaluation workflow [80, 141, 165, 285, 370, 381]. More popularly, efforts have focused on developing XAI tools that help non-technical users understand when to consult and rely on predictive systems during high-stakes decision-making [14, 90, 356]. Across both examples, XAI tools have carried forward

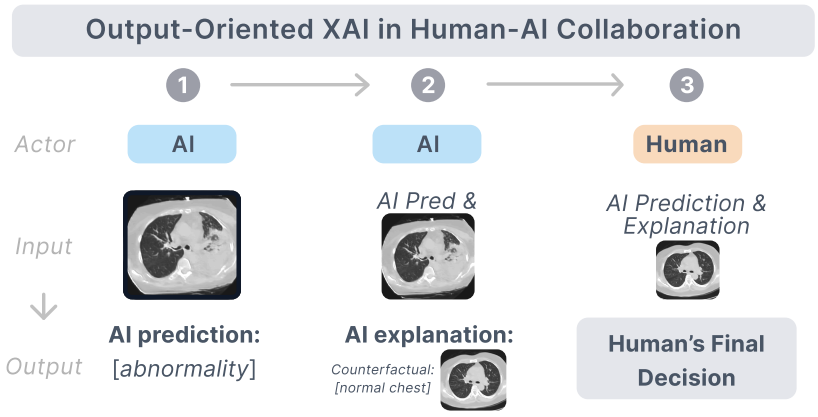


Fig. 1.1: A conceptual overview of output-oriented XAI in human–AI decision-making workflows. CT scans generated using MedSyn [380].

the framing introduced by explanation facilities in expert systems, preserving the traditional uses of explanations: fostering trust in the AI system, calibrating reliance on AI outputs, and supporting users' understanding of model behavior. These traditional uses of AI explanations reflect an output-oriented XAI paradigm, in which explanations are designed to help users interpret, validate, or justify AI outputs, and are commonly evaluated through their effects on trust, reliance, and model understanding (*e.g.*, [49, 150, 200, 364]).

As depicted in Figure 1.1, output-oriented XAI positions decision-makers as validators of AI predictions, focused on interpreting the rationale behind the output for a given input to determine whether they should accept or reject the AI's recommendation. For example, an AI system reviewing a chest image might output the presence or absence of an abnormality, followed by an explanation of that prediction (*i.e.*, a counterfactual explanation showing a chest that it predicted to have no abnormalities). The human, such as a radiologist, must then review both the AI's prediction and its explanation while simultaneously reasoning backward about whether they align with the evidence in the original image. In doing so, this workflow may unintentionally shift the radiologist's attention away from actively reasoning about the underlying data, context, and decision itself. Recent literature has questioned whether these one-shot, interaction paradigms commonly used in empirical studies of human–AI collaboration are appropriate, due to their lack of ecological validity [318] and lack of consideration for decision consequence or multi-stakeholder influence [120]. Furthermore, a recent survey found that explanations have no significant impact on the effectiveness of human–AI collaboration [348]. Motivated by these critiques and limited positive evidence for XAI, this thesis argues for expanding the design space for AI explanations in human–AI collaboration by challenging the output-oriented designs of XAI.

The widespread development of open-source interpretability toolkits (*e.g.*, Srinivasu et al. [328], Wexler et al. [370], and Nori et al. [257]) and XAI methods that are not for

any particular purpose [112] has enabled many HCI researchers, designers, and product developers to easily prototype with, integrate, and deploy XAI techniques into products and workflows across a wide range of industries [152, 159]. Regardless of the domain, explanation methods typically have focused on output-oriented uses with the goal of helping people interrogate AI outputs at the instance and model level (*i.e.*, referred to as local and global explanations [306]). A plethora of empirical work has focused on investigating the impact of levels of interactivity [26, 220], explanatory focus (*i.e.*, local versus global) [278], and explanation techniques (*e.g.*, feature-based, counterfactual, rule-based, example-based, etc. [60, 90, 149]) to determine which configuration best helps different user groups judge a model's recommendation correctness. However, these empirical studies have reinforced the narrow focus on output-oriented XAI designs, aiming to help decision-makers calibrate their reliance on or understand model behavior when using recommendation-centric [389] AI systems.

1.1 Thesis Motivation

Despite years of empirical studies exploring the impact of XAI in human–AI decision-making, literature surveys have shown mixed results and conflicting findings on XAI's value [124, 290, 348, 354]. Some lab studies conclude that explanations are more harmful than helpful when humans use AI to assist with classification tasks, while other studies argue that explanations can improve task performance under certain circumstances. Human-centered AI researchers argue that these inconclusive results stem from taking a one-size-fits-all approach to explanations used in human–AI collaboration, ignoring who would be "opening the black box" and why [96]. Ironically, foundational research on explanations in expert systems echoes the importance of tailoring explanations of rule-based algorithms based on who the explainee is [333]. Ultimately, a culmination of these critiques and stagnant progress laid the foundation for focused efforts on human-centered XAI (HCXAI) [93, 95].

HCXAI efforts have comprised studies that seek to tailor the explanation's presentation style (*i.e.*, graphical vs. textual) [135], the timing of explanation delivery (*i.e.*, proactively vs. on-demand) [45, 59, 86], and the level of interactivity (*i.e.*, static vs. dynamic) [23, 26, 29, 185] to who is requesting the explanation. The rise of multimodal generative AI (GenAI) has supported these ongoing HCXAI efforts by making it easier to generate natural-language explanations tailored to different user groups [55, 249]. This use of GenAI to generate explanations for another AI system's predictions has been described in the literature as Explainable Generative AI (GenXAI) [302]. Ongoing efforts with GenXAI in human–AI collaboration tend to scope the XAI to output-oriented interactions (*e.g.*, [239, 327]).

Recent work has grown critical of the ongoing HCXAI research in human–AI decision-making, noting that tailoring explanations to accurately model infinite user preferences may still fail to address the ongoing challenges with XAI [255]. Drawing

on Redström [282]'s perspective on user-centered design, we argue that efforts to personalize XAI through user-centered design methods may yield overly restrictive explanations. These overly restrictive designs can unintentionally limit users' ability to easily adapt and appropriate AI explanations into their domain-specific reasoning processes and workflows for uses beyond validating, interpreting, or justifying the AI's outputs.

HCXAI studies also tend to assume that AI explanations are "perfect", overlooking the impact that imperfections may have on output-oriented XAI designs. As GenXAI adoption increases, AI explanations are prone to imperfections. Although technical researchers have pointed out that AI can be "right for the wrong reasons" [332] or inconsistent even for the same input and regardless of the AI's prediction accuracy [43, 186, 292], few empirical studies investigate the impact of imperfect AI explanations [136, 267] (*i.e.*, when an AI explanation misaligns with the AI's prediction [239]). Part I of this thesis contributes to the literature by presenting empirical studies on how imperfect output-oriented XAI can lead to inappropriate reliance and distorted reasoning.

Ultimately, the existing XAI literature on human–AI collaboration highlights XAI's ongoing challenge to help decision-makers validate, interpret, or justify AI outputs. One conclusion may be that XAI is just not suitable for human–AI collaboration. However, these limitations are evident only for output-oriented XAI and rarely consider the imperfections of AI explanations in study designs. The design for and evaluation of AI explanations beyond interpreting, validating, and justifying AI outputs is underexplored. Together, these two gaps in the literature motivate the questions this thesis is centered around:

(**RQ1**) How do imperfect, output-oriented AI explanations impact human–AI decision-making?

(**RQ2**) How can AI explanations be used during human–AI collaboration beyond interpreting, validating, and justifying AI outputs?

(**RQ3**) How can we expand the design space to support the emerging uses of AI explanations in human–AI collaboration?

1.2 Process-Oriented XAI Paradigm

This thesis argues that the application of HCXAI has largely focused on output-oriented XAI [354], resulting in an overly restrictive design space in which XAI is framed for interpreting, validating, and justifying AI outputs. Recent literature suggests that the design of AI systems should depart from directly providing users with a recommendation (recommendation-centric) and focus on providing process-oriented support [389]. Designs for such process-oriented AI systems have emerged as a design practice called Tools for Thought [339]. This design practice intentionally builds AI systems that aim to protect and augment human cognition rather than replace or

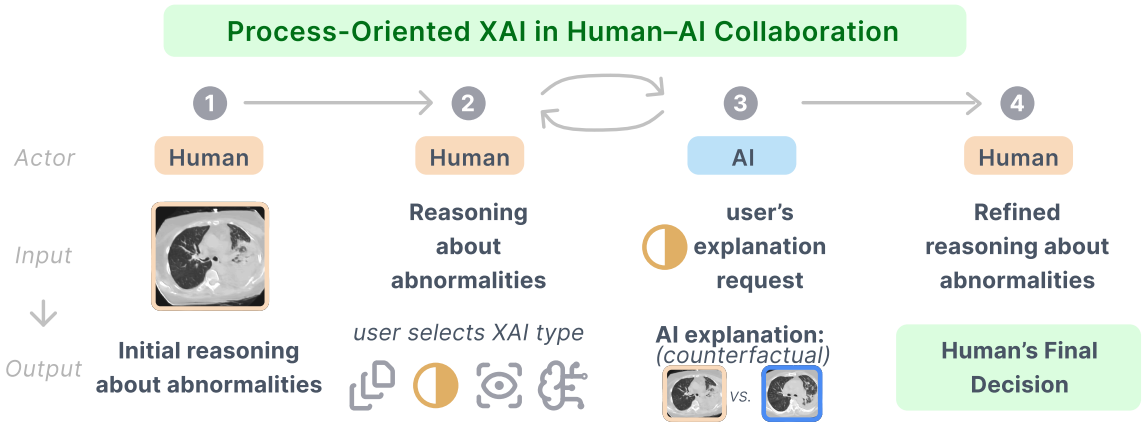


Fig. 1.2: A conceptual overview of process-oriented XAI in human–AI decision-making workflows. CT scans generated using MedSyn [380].

minimize it. However, this reframing leaves the role of XAI innovations unclear, as XAI was initially designed to help decision-makers validate, interpret, and justify outputs from recommendation-centric AI systems.

Much of the XAI literature has treated explanations as static artifacts for explaining AI outputs. However, critiques have argued that explanations are a social [289] and dialogic [234] practice where understanding is co-constructed. Malle's seminal work on explanations [228] (Chapter 3.2) supports these critiques, presenting explanations as a vehicle to facilitate shared understanding among individuals. Complementing this, work in cognitive psychology positions explanations as mechanisms that mediate humans' reasoning processes [222]. Using cognitive psychology of explanations [171, 222], social psychology of explanations [228], and Alan Dix's emphasis on appropriation-aware design [89] as a research lens, **we introduce process-oriented XAI** as a design space that reframes XAI techniques as tools to support human processes throughout workflows, such as reasoning, multi-stakeholder communication, or coordination. Such workflow support can help users refine and shape their reasoning about the underlying data, context, and decision space. This reframing is motivated by the fact that many XAI techniques can emulate aspects of core human reasoning patterns (*i.e.*, counterfactual evidence, example-based reasoning, and localization).

Rather than focusing on tailoring or personalizing explanations to user personas to help people validate, interpret, or justify AI outputs, the latter half of this thesis (Part II and Part III) explores how AI explanations can be designed for and oriented toward different processes that arise throughout workflows. This perspective, depicted in Figure 1.2 as process-oriented XAI, marks a conceptual shift: the role of end-users shifts from validators of AI outputs to active thinkers, and AI explanations shift from validation, interpretation, and justification tools to tools that can support domain-specific sensemaking and reasoning. In contrast to the output-oriented XAI designs, where the AI system predicts abnormalities in a chest image and then provides an

explanation of that prediction for the decision-maker to validate, justify, or interpret the output, a process-oriented XAI design leverages AI's explanatory capabilities to help the decision-maker iteratively shape and refine their reasoning process.

For example, the radiologist may request to see a variation of a chest with a similar-looking abnormality to help them rationalize the abnormality in their patient's chest through visual comparison, refining their reasoning. This design introduces friction by intentionally withholding the AI's prediction or explanation of the data to interpret, and instead facilitates an iterative reasoning process that forces the radiologist to elicit their own reasoning process and use AI explanations to shape it. Within this paradigm, imperfect explanations can be treated as a design goal, as they can introduce friction into the reasoning process, encouraging decision-makers to critically think and reflect on their own reasoning.

The key distinction between output-oriented and process-oriented paradigms is the role AI explanations are designed to play in human–AI collaboration. In output-oriented XAI, the role of AI explanations is to verify, interpret, or justify AI outputs in AI-assisted decision-making, positioning the human as a validator of AI outputs. On the other side, we have process-oriented XAI, where XAI is reframed as a tool to support workflow processes, positioning the human as an active thinker throughout decision-making. Ultimately, these two paradigms address different design objectives towards the same goal: improving human–AI collaborations.

1.3 Thesis Statement

The narrowly scoped design space for AI explanations risks overlooking the broader impact that explanations can bring to human–AI collaboration. My work challenges the traditional ways we think about XAI to expand the design space by exposing the impact of imperfect explanations in output-oriented XAI designs, exploring the emerging uses of AI explanations, and proposing process-oriented XAI as one approach to support those emerging uses.

1.4 Thesis Overview

To address these questions, this dissertation is separated into three parts, depicted in Figure 1.3: Part I evaluates the impact of imperfect, output-oriented XAI designs in human–AI collaboration; Part II explores the emerging uses of AI explanations in human–AI collaboration; and Part III explores process-oriented XAI as one approach to expanding the design space for AI explanations in human–AI collaborations. This thesis details studies across five different contexts: building damage assessment, bird species identification, individual species population counts, chronic disease management, and medical image interpretation.

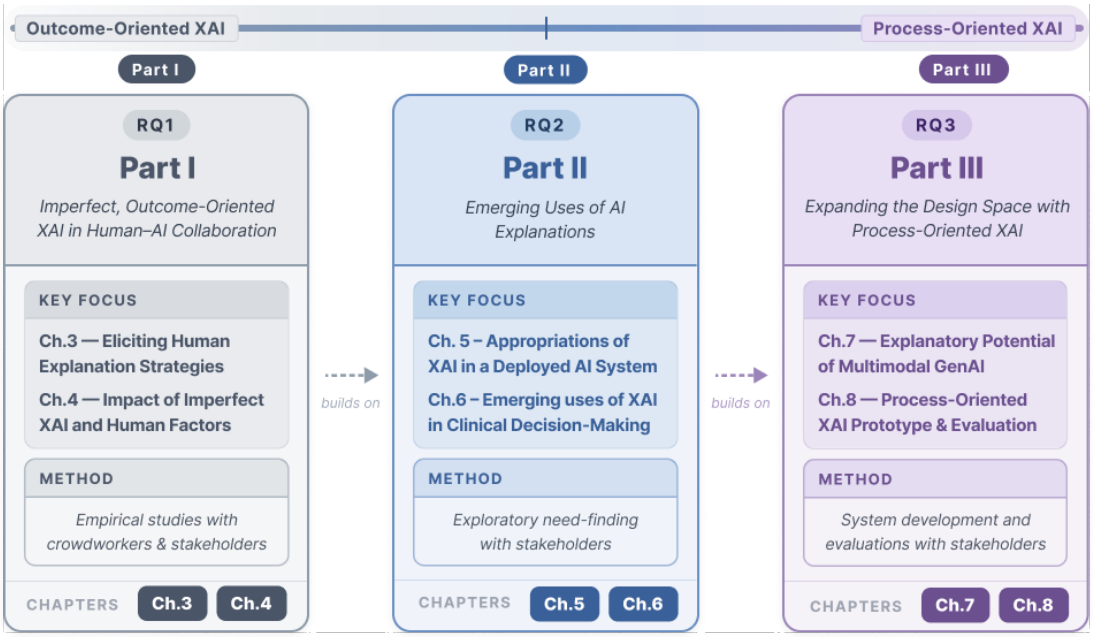


Fig. 1.3: This thesis consists of three parts: capturing the impact of imperfect, output-oriented XAI (Part I), discovering emerging uses of AI explanations in human–AI collaboration (Part II), and expanding the design space for AI explanations through process-oriented XAI with explainable generative AI (Part III). Figma Make was used to help design this figure.

Part I of this dissertation addresses RQ1 by presenting the impact of imperfect, output-oriented explanations on human–AI decision-making when used to assist decision-makers in AI-assisted building damage assessment and bird species classification. Chapter 3 uses a wizard-of-oz AI setup to elicit human explanation strategies and validate how this impacts decision-makers' reliance on AI. Chapter 4 builds on this study by exploring the impact of different types of imperfect XAI techniques (*i.e.*, example-based and natural language) on AI reliance in an AI-assisted bird species classification task for a diverse group of end-users, ranging from crowdworkers to birding hobbyists and ornithologists. Across both studies, we show how imperfect, output-oriented XAI can lead to overreliance on AI, even when the AI's explanations do not reflect the AI's prediction.

To better understand how else AI explanations can be used, Part II addresses RQ2 by conducting two exploratory studies with domain stakeholders who use AI decision-support tools in their workflows. Chapter 5 presents a case study with a

non-profit wildlife conservation organization in which we conducted semi-structured interviews to observe how domain stakeholders currently operationalize and adapt XAI in their decision-support tool for individual species population counting. We continue this line of investigation in Chapter 6 with clinicians who care for patients with chronic diseases. Using a design probe of an interactive XAI system, we probe clinicians on how they might use it throughout their workflow. The findings of these two studies mark a turning point in how we design for and conceptualize the use of AI explanations: XAI can be used as a tool to augment processes throughout complex human–human collaboration by supporting communication, enabling sensemaking or reflection, and facilitating knowledge transfer.

Rooted in the emerging uses highlighted in Part II, we address RQ3 in Part III by expanding the design space for AI explanations by shifting from output-oriented designs to process-oriented interactions using multimodal explainable generative AI as a design medium. Chapter 7 captures the explanatory potential of multimodal generative AI in the context of medical image interpretation. Chapter 8 then presents a process-oriented XAI system for leveraging explainable generative AI techniques for medical image interpretation. We capture the impact of process-oriented XAI on reasoning processes through a think-aloud and semi-structured interview study with clinicians-in-training.

Chapter 9 presents a discussion on the implications for output-oriented and process-oriented XAI designs in human–AI collaboration contexts. This chapter also highlights future directions for further evaluating and refining process-oriented XAI designs in human–AI collaborations.

1.5 Prior Publications and Authorship

This thesis includes adapted versions of papers that I have already published:

Morrison, K., Shin, D., Holstein, K., & Perer, A. (2023). Evaluating the impact of human explanation strategies on human–AI visual decision-making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1-37.

Morrison, K., Spitzer, P. (*co-first authors*), Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The impact of imperfect XAI on human–AI decision-making. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), 1-39.

Turri, V., **Morrison, K.** (*co-first authors*), Robinson, K. M., Abidi, C., Perer, A., Forlizzi, J., & Dzombak, R. (2024, June). Transparency in the Wild: Navigating Transparency in a Deployed AI System to Broaden Need-Finding Approaches. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1494-1514).

Morrison, K., Mathur, A., Bradshaw, A., Wartmann, T., Lundi, S., Zandifar, A., Dai, W., Batmanghelich, K., Eslami, M., & Perer, A. (2025, July). Opportunities & Challenges

in Designing Text-to-Image Generative AI for Medical Education, Training, & Practice. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. Vol. 8. No. 2. 2025.

Morrison, K., Li, Z., Kim, M., Hong, C., Kraisangka, J., Correa-Jaque, P., Fauvel, C., Sahay, S., Vanderpool, R., Everett, A., Lin, S., Kanwar, M., Benza, R., & Perer, A. (2026, July). It's Not Just for Trust: Designing for Emerging Uses of Explainable AI in Clinical Decision-Making. *Under Major Revisions* at ACM HEALTH 2026.

Morrison, K., Mason, E., Chen, M., Lundi, S., Zandifar, A., Kim, D., Dai, W., Eslami, M., & Perer, A. (2026, April). Reframing Explainable Generative AI as a Tool for Thought. In ACM Conference on Human Factors in Computing Systems' Workshop on Tools for Thought (2026).

Morrison, K., Mason, E., Chen, M., Lundi, S., Zandifar, A., Kim, D., Dai, W., Eslami, M., & Perer, A. (2026, April). Augmenting Reasoning, Not Validating Outputs: Designing a Process-Oriented XAI System with Multimodal Generative AI. Manuscript in submission (2026).

Chapter 2

Background & Related Work

This chapter is adapted from the related works of the following papers:

Morrison, K., Shin, D., Holstein, K., & Perer, A. (2023). Evaluating the impact of human explanation strategies on human–AI visual decision-making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1-37.

Morrison, K., Spitzer, P. (*co-first authors*), Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The impact of imperfect XAI on human–AI decision-making. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), 1-39.

Turri, V., **Morrison, K.** (*co-first authors*), Robinson, K. M., Abidi, C., Perer, A., Forlizzi, J., & Dzombak, R. (2024, June). Transparency in the Wild: Navigating Transparency in a Deployed AI System to Broaden Need-Finding Approaches. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1494-1514).

Morrison, K., Mathur, A., Bradshaw, A., Wartmann, T., Lundi, S., Zandifar, A., Dai, W., Batmanghelich, K., Eslami, M., & Perer, A. (2025, July). Opportunities & Challenges in Designing Text-to-Image Generative AI for Medical Education, Training, & Practice. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. Vol. 8. No. 2. 2025.

Morrison, K., Li, Z., Kim, M., Hong, C., Kraisangka, J., Correa-Jaque, P., Fauvel, C., Sahay, S., Vanderpool, R., Everett, A., Lin, S., Kanwar, M., Benza, R., & Perer, A. (2026, July). It's Not Just for Trust: Designing for Emerging Uses of Explainable AI in Clinical Decision-Making. *Under Major Revisions* at ACM HEALTH 2026.

Morrison, K., Mason, E., Chen, M., Lundi, S., Zandifar, A., Kim, D., Dai, W., Eslami, M., & Perer, A. (2026, April). Reframing Explainable Generative AI as a Tool for Thought. In ACM Conference on Human Factors in Computing Systems' Workshop on Tools for Thought (2026).

Morrison, K., Mason, E., Chen, M., Lundi, S., Zandifar, A., Kim, D., Dai, W., Eslami, M., & Perer, A. (2026, April). Augmenting Reasoning, Not Validating Outputs: Designing a Process-Oriented XAI System with Multimodal Generative AI. Manuscript in submission (2026).

This thesis is built on the substantial body of prior work in explainable AI (XAI). Section 2.1 provides an overview of the different types of XAI techniques that have been developed, and Section 2.2 synthesizes the body of work on output-oriented XAI, from efforts on evaluating XAI to designing human-centered XAI. Section 2.3 captures the extent to which literature has acknowledged imperfect AI explanations in empirical studies. Motivating Part II, Section 2.4 highlights the existing gap in literature on studies that design for or understand the emerging uses of AI explanations in human–AI collaboration. Section 2.5 provides background for Part III and motivates the design choice to use generative AI as an explanatory tool, and Section 2.6 reviews the extent of process-oriented designs in (X)AI systems.

2.1 Explainable AI (XAI) Techniques

Explainable AI (XAI) techniques have long been framed as a way to explain AI outputs. This output-oriented framing is captured by numerous literature surveys on the advancements of XAI techniques (*e.g.*, [9, 236, 275, 329, 331]). Grounded in the terminology used by human–computer interaction (HCI) researchers [211], we organize these techniques to be either solutions that intend to "explain why" or solutions that intend to "explain why not/what if". The XAI field has addressed these questions through a range of techniques, including feature attribution, counterfactual, and example-based explanations [290].

Explaining "Why". Explaining why an AI made a certain prediction has been addressed using several different explanatory techniques, including feature attribution, counterfactual, and example-based techniques [211]. Among these, feature attribution methods are a widely adopted approach for explaining why a specific output was produced [290]. This class of methods is the most direct, as they focus on revealing which features of the given input most contributed to the AI's prediction [211]. An example of a feature attribution technique is SHAP [224], which is a model-agnostic, local explanation method that attributes model predictions to individual input features and produces a one-time, static output. Another approach, LIME [285], fits an interpretable model to the black-box model to explain a given prediction. While SHAP and LIME are two popular model-agnostic, feature attribution methods, several model-specific methods have also been designed for specific architectures or data types [250].

Explaining "Why Not/What if". Although explaining "why" is a very popular question to address, several techniques have also been motivated by trying to explain "why not" or "what if". Techniques for these questions, such as counterfactual explanations and example-based explanations, are designed to explain AI outputs by providing contrastive information that shows why an instance was not predicted to be a certain

class or how changes to the input would result in a different output [357]. These techniques have ranged from automated approaches [344] (*e.g.*, DICE [241]) to interactive, user-driven solutions [23]. However, the automated approaches may unintentionally leverage irrelevant or non-actionable variables [29], making the explanations irrelevant or difficult to interpret. There are also normative and comparative example-based explanations [51] that can show examples from training data that closely match the target class or closely match the predicted class. Beyond providing local explanations, there are some global explanation techniques that aim to explain "why not", such as partial dependence plots (PDPs) [275]. PDPs capture the relationship between an input feature and the prediction [275].

Explainable AI techniques have focused on explaining AI outputs from the start, positioning end-users as validators and inspectors of AI systems. While those roles are appropriate for data scientists or model builders, they may not align with domain-specific decision-makers' workflows. Few works have yet to explore the design space for AI explanations with domain stakeholders beyond this output-oriented XAI paradigm. In the next section, we synthesize the impact that these XAI techniques have on human–AI decision-making.

2.2 Output-oriented XAI in Human–AI Collaboration

The predominant interaction design with XAI in human–AI collaboration systems has been output-oriented, providing tools to help users verify or interpret AI outputs. Human–AI collaboration researchers have spent years manipulating the configurations of XAI designs and interactions in hopes of finding the "best" configuration to help humans verify or interpret AI recommendations more effectively (*e.g.*, [59, 90, 135, 149, 266]). We provide a review of how XAI has been evaluated in human–AI collaboration and highlight efforts towards making XAI more human-centered.

Several surveys in the literature capture the substantial amount of studies that measure the success of XAI techniques by evaluating users' trust, reliance, usability, and model understanding during human–AI decision-making tasks [9, 109, 164, 236, 290, 331, 354]. Seminal works and early evaluations of XAI techniques helped establish these evaluation practices. For example, in 2018, SHAP was presented as a way to explain to clinicians why a risk prediction model classified a given patient as it did [225]. The paper that introduced LIME in 2016 similarly situates its contribution by evaluating its ability to foster trust and support model understanding [285]. Several papers have followed by similarly evaluating how XAI techniques impact users' trust, reliance, model understanding, or task performance (*e.g.*, [13, 117, 313, 349, 364, 365]).

Despite initial hopes that XAI would be the solution to challenges in human–AI collaboration, several studies have shown inconsistent benefits [348, 354] and documented XAI's limitations [69, 70, 165, 191]. For example, XAI has been found to be misinterpreted by data scientists [165], to lead people to misunderstand the AI's

behaviors [70], fail to highlight meaningful information [151], and fail to support humans' explanation needs of complex interdependent data points [191]. Papenkordt [266] and Cau and Spano [59] add to this body of literature by finding that the timing of when an AI explanation is shown (*i.e.*, on-demand vs. immediate) had insignificant effects on AI reliance.

Researchers have argued that these challenges stem from the lack of human-centered XAI (HCXAI) techniques that consider end-users' domain expertise, decision context, and explanation needs [93, 98, 209]. Human-centered XAI (HCXAI) literature acknowledges that explanations can be requested for a variety of reasons from a variety of different end-users; incorporating *who* the explanation is being designed for, and *why* is a core part of human-centered XAI [93, 97, 98, 209]. This sparked a substantial number of empirical studies and need-finding work that explore what human-centered XAI looks like. For example, several HCXAI efforts have defined processes for setting XAI requirements and map these to needs and specific XAI techniques (*e.g.*, [47, 66, 71, 100, 101, 128, 158, 212]). Approaches used across prior work include conducting interviews with end-users (*e.g.*, [71, 212, 355]), prototyping potential solutions with end-users (*e.g.*, [71, 158, 177, 355]), or presenting scenarios to stakeholders (*e.g.*, [71, 95]).

Numerous empirical studies have been published on what information decision-makers need from an AI explanation (*e.g.*, [46, 75, 87, 201, 305, 379]). Other empirical studies look at how decision-makers' domain expertise impacts how they interpret explanations, and in turn impacts their task performance, during human–AI collaborations (*e.g.*, [20, 54, 88, 111, 260, 261, 335, 343, 390]). And, many studies compare and contrast different explanation techniques or modalities to see which results in better human–AI team performance (*e.g.*, [67, 90, 135, 149, 382]). Studies have even explored the interactivity of AI explanations, finding that interactive counterfactual explanations align with human-centered principles because they give users control over changing the input features themselves (*e.g.*, [29, 125, 366, 387]). However, many of these studies similarly face negative impacts or insignificant findings.

Although HCXAI has contributed to expanding the design space for AI explanations, there remains a narrow focus on designing XAI to explain AI outputs and evaluate based on trust, reliance, or model understanding. Furthermore, HCXAI fails to capture the impact that imperfect AI explanations have on human–AI collaboration.

2.3 Imperfect AI/XAI

Numerous studies in CSCW and HCI have investigated the impact that imperfect AI has on human–AI collaboration (*e.g.*, [16, 181, 351]). Kocielnik et al. [181] offer three techniques for setting user expectations about the performance of an imperfect AI system, including an accuracy indicator, example-based explanations, and performance control. Several recent studies investigate how programmers collaborate with Copilot,

an imperfect AI programming assistant (*e.g.*, [19, 79, 351]). One of those studies specifically examines how to convey the uncertainty of Copilot outputs [351]. Recent technical work has proposed approaches to mitigate "noisy" or incorrect natural language explanations [154] and Kroeger et al. [187] proposed metrics to algorithmically evaluate the effectiveness of the generated post-hoc natural language explanations. Roy et al. [292] propose metrics to capture the disagreements across explanation techniques. However, fewer studies investigate the impact that imperfect XAI has on human–AI collaboration [136, 267]. For example, Papenmeier et al. [267] investigate the impact that explanation fidelity has on user trust. Nourani et al. [259] look at how meaningful versus meaningless explanations (similar to, but different from imperfect XAI) can impact users' understanding of the AI's accuracy.

2.4 Emerging Uses of AI Explanations

Relevant literature has identified emerging uses of XAI, such as Lundberg et al. [225], who note that SHAP could improve clinical understanding of hypoxemia during surgery, and Woensel et al. [375], who argue that AI explanations can help patients with a chronic disease better understand their condition and motivate long-term behavior change. However, neither of these works conducts an empirical study to ground these emerging uses in stakeholders' perspectives, workflows, or behaviors. Other work has commented on or alluded to the potential of using XAI to mediate multi-stakeholder communication [75, 158, 176]. Jin et al. [158]'s framework for designing user-centered XAI exposed how patients might use XAI to communicate with their caregivers and family members about their condition. In contrast, Corti et al. [75] observed clinicians imagining how they could use XAI to communicate with other stakeholders and justify their decisions. Literature surveys have also commented on AI explanations' potential role in being used for knowledge discovery and learning [9, 232]. However, the intentional design for and understanding of the current appropriation and emerging uses of XAI in human–AI collaboration is underexplored.

2.5 The Explanatory Potential of Generative AI

Generative AI (GenAI) is fundamentally reshaping how we conceive and design for AI explanations in human–AI collaboration. As GenAI becomes increasingly competent, researchers are now exploring (1) how to explain GenAI itself and (2) how to leverage these models in advancing XAI techniques [31, 302, 350]. Together, these two goals make up the subfield of Explainable Generative AI (GenXAI) [302]. This generative aspect opens the door to exploring new interaction paradigms with XAI in human–AI collaboration.

A growing body of literature has shown interest in using GenAI to augment existing XAI approaches [31]. This transition towards GenXAI is rooted in empirical

findings that non-experts struggle to comprehend AI explanations and in mixed results regarding the impacts on human–AI collaboration. For example, a series of recent technical papers have introduced explanation methods that leverage GenAI to explain another AI model by generating natural language explanations (*e.g.*, [55, 137, 249]) or generating images as counterfactual explanations (*e.g.*, [157, 169, 197]). Recent work has also leveraged GenAI to augment the presentation of outputs from traditional XAI techniques [254]. For example, Nguyen et al. [254] use GenAI to generate a natural language explanation to further augment the interpretation of another explanation, such as feature attributions.

Despite the potential of GenXAI, research on it has primarily remained focused on output-oriented interactions, framing GenXAI as another tool to help users verify, interpret, or justify AI outputs. In fact, empirical studies exploring the impacts of output-oriented GenXAI in human–AI collaboration face many of the same limitations as traditional output-oriented XAI techniques, as they are prone to generating imperfect explanations [239, 327].

In a separate thread of work, researchers have framed GenAI as a tool to support workflow processes, such as domain-specific sensemaking and learning [295]. Numerous prior works in the medical literature have explored foundation models' ability to generate images suitable for medical education [182, 190, 256, 359, 391]. For example, Zhu et al. [391] probe the potential of OpenAI's DALL-E [280] to generate ECGs, including ECGS that are abnormal. Although they do not present a study on stakeholder probing or interviewing, the authors draw on their professional experience to offer insight into DALL-E's potential as a teaching tool. Ultimately, they deemed DALL-E incompetent at generating medically relevant ECG illustrations. Noel [256] and Kumar et al. [190] similarly probe different text-to-image GenAI models to evaluate their capabilities for generating various medical images. Waikel et al. [359] ran an educational intervention study with students to quantitatively evaluate the impact of generative AI images in education about pediatric genetics. Although the participants did not engage in the generation process, their main conclusion was that real images were more helpful, while generated images played a more supporting role in learning. Overall, the takeaway from this body of work is that few works explore GenXAI techniques in these process-oriented contexts.

2.6 Process-Oriented (X)AI Support

A movement to protect human cognition in the design of human–AI collaboration systems has recently emerged [338, 339]. Zhang and Reicherts [389] refer to this as process-oriented support in a system they designed called *ExtendAI*, where users were prompted to provide their own line of reasoning for a decision before the AI gives feedback on their plan [283]. Although not explicitly tailored to a specific workflow, Buçinca et al. [45] experiment with different cognitive forcing functions,

such as showing the AI recommendation only after the user provides an initial decision or offering AI support as an on-demand feature. Kang et al. [160] foster a similar interaction paradigm in their system, Synergi, where AI is used to expand upon users' initial reasoning when searching for relevant research publications. In a similar vein, Sivaraman et al. [319] reframe decision support systems as a collection of Intelligent Reasoning Cues and finds that different cues can play different roles in decision-making. These studies align with the process-oriented XAI interaction paradigm, which removes the AI's recommendation and gives users agency over which XAI techniques to use.

In the context of XAI systems, research has begun exploring process-oriented interaction designs by introducing new frameworks, such as Evaluative AI [233] and Hypothesis-Driven AI [21], or proposing new explanation designs, such as contrastive explanations that aim to predict gaps in human reasoning [44]. Buçinca et al. [44] even observed improvements in human–AI decision-making from their new explanation design, hinting at the value in aligning explanations with human reasoning processes. However, the development and exploration of a process-oriented XAI system using different XAI techniques remains underexplored.

Part I

Imperfect Output-Oriented Explainable AI in Human–AI Collaboration

Chapter 3

The Impact of Human Explanation Strategies for Output-Oriented XAI

This chapter is adapted from the following paper:

Morrison, K., Shin, D., Holstein, K., & Perer, A. (2023). Evaluating the Impact of Human Explanation Strategies on Human–AI Visual Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 7 no. (CSCW1), 1-37.

3.1 Introduction

Artificial intelligence (AI) systems are increasingly being deployed with the aim of helping practitioners make high-stakes decisions more quickly and accurately. For example, computer vision enables the extraction of meaningful information from images, such as classifying objects within an image or segmenting an image to locate objects [334]. With computer vision models rapidly increasing in accuracy, several tasks requiring image classification, segmentation, or object detection have been automated to aid practitioners. For example, radiologists may work with AI-based systems to evaluate medical images [52, 268], while humanitarian aid and disaster relief workers may work with AI to assess building damage from satellite imagery after natural disasters or civil unrest [126]. These are examples of visual decision-making tasks that can be supported by advanced AI systems and are becoming increasingly prevalent. However, human–AI collaboration is not guaranteed to result in better decision-making in practice.

To support more effective human–AI collaboration, several eXplainable AI (XAI) techniques have been developed with the goal of providing insight into how an AI makes its decisions, helping decision-makers calibrate their trust and reliance upon AI[1]. For instance, some techniques yield saliency maps (heatmaps) indicating the most important regions of the image that contributed to its classification [314]. However,

[1]Throughout this chapter, we distinguish between the terms *reliance* and *trust*: reliance is an observable human behavior that can be measured, whereas human trust is a latent factor which cannot directly be measured from observable human behaviors [81, 153, 199, 352]

many of these methods have been designed for data scientists or machine learning engineers instead of subject matter experts [211]. Furthermore, Brennen [41] and Liao et al. [211] emphasize that several XAI techniques do not present the information that decision-makers actually need to see to inform their decision-making. For instance, in the context of medical imaging, Saporta et al. [298] report that existing XAI techniques rarely highlight clinically meaningful image regions. These empirical findings highlight the need for a more human-centered and empirically informed approach to designing and evaluating AI explanation techniques to understand which explanation techniques, or combination of techniques, can truly support effective human–AI collaborative decision-making.

To tackle this need, our research takes a broader view, shifting focus from current XAI techniques to ask the question: *What types of explanations do humans generate and benefit from in the context of visual decision-making tasks*? We argue there is much to learn from *human-generated* explanations in the context of human–human collaboration [362]. In many real-world decision-making settings, practitioners generate explanations intended for other practitioners to consume to support their decision-making. For example, when radiologists report findings in medical images, they often characterize their observations and inferences for other radiologists or practitioners in other care units [63]. By uncovering strategies humans use to explain decisions effectively to others, we hope to inform the design of novel human-centered XAI approaches that could emulate these strategies. Inspired by these observations, in this study, we focus on three primary questions:

- **RQ1.** What strategies do humans use to explain, or rationalize, their reasoning for a high-stakes image classification task?
- **RQ2.** How do different human explanation strategies improve humans' accuracy in a high-stakes image classification task?
- **RQ3.** How do different human explanation strategies impact decision-makers' reliance on AI?

To ground our experiments in a real task requiring high-stakes decisions, we focus on AI-assisted decision-making for assessing building damage after natural disasters, utilizing satellite imagery. Assessing the damage is typically a bottleneck before responders can help, and AI-based solutions have been proposed to solve this bottleneck [126].

We address our research questions through a series of two studies. In the first study, we address **RQ1** by designing a system that shows participants several pairs of images for a specific geographic location before and after a natural disaster. For each image pair, participants assessed the level of damage and were asked to provide convincing explanations using a wide range of tools that justified their assessment. These results allowed us to uncover explanation strategies that humans employ when explaining their decision to others, which can help inform the style and design of novel human-centered XAI techniques.

In the second study, we address **RQ2** and **RQ3** by conducting a follow-up study in which the explanations elicited from participants in the first study are presented under the guise of AI. This approach allowed us to evaluate how different explanation strategies identified from the first study may impact decision-makers' accuracy and reliance on AI, as if these strategies were implemented and deployed as human-centered XAI techniques.

Across our two studies, we contribute the following:

- We provide insight into the **types of explanation strategies that humans employ** when providing rationales for their decisions, in the context of a visual decision-making task **to inform the design of human-centered XAI techniques**.
- We present a **new approach** to explore the impacts of prospective explainable AI techniques on human–AI decision-making by presenting participants with different types of human-generated explanation strategies, where the human explanation strategies are framed as AI-generated explanations.
- Using this approach, we conduct the **first empirical investigation** in the literature into the **impacts of different human explanation strategies** on task accuracy and reliance upon AI-based assessments.

3.2 Study Context: Building Damage Assessment

To ground our experiments in a real task requiring high-stakes visual-based decisions, we focus on AI-assisted decision-making for assessing building damage after natural disasters, utilizing satellite imagery. Assessing the damage is typically a bottleneck before first responders and government funding agencies can help. As such, AI-based solutions have been proposed to solve this bottleneck [126].

Satellite imagery is abundant, which has resulted in numerous computer vision applications. For example, satellite imagery and computer vision techniques have been used in various high-stakes scenarios, such as identifying economic growth and stability [156, 263, 384], detecting poachers and illegal fishing vessels [353], and identifying damaged buildings after natural disasters, armed conflicts, or other catastrophic events [126].

Natural disasters like hurricanes, floods, wildfires, and earthquakes often devastate humans and their properties. Using satellite imagery to assess building damage after a natural disaster is a critical task. The damaged structures put a huge strain on the local and regional economies, forcing officials to rely on the government for funds to support recovery efforts. Our interviews with HADR experts suggest that in certain situations, rapid and accurate identification of the number of damaged structures is crucial to receive funding in a timely manner for recovery efforts.

3.2.1 Task Expertise

We chose building damage assessment from satellite imagery as our task to address these research questions because prior work indicates that it is possible to train users to perform this task with reasonable accuracy quickly, and platforms already exist to crowdsource building damage assessment from laypeople [8, 172], as discussed above. Unlike other visual decision-making domains, such as radiology, where professionals are often required to have specialized expertise and crowdsourcing is not an option, prior work for building damage assessment has successfully used crowdsourcing approaches to help gather information more quickly after natural disasters [172, 312]. Furthermore, many people have experience viewing and interpreting satellite images in their day-to-day life unlike medical images, such as when using navigation systems or online interfaces like Google Earth.

We had informal discussions with HADR experts to understand the domain better while designing our studies. These discussions helped us understand what HADR experts would look for in satellite imagery to determine if a building is damaged and its level of damage. For example, analysts look for buildings where the roof had completely collapsed or buildings that are partially or completely surrounded by water. Analysts also look for displaced portions of buildings in the immediate surrounding areas. These examples helped us create a change detection [341] training module to quickly train our participants to assess building damage from pre- and post-disaster satellite imagery. These discussions also provided our team with an understanding of the skill level that people had to create the ground truth labels for the building damage assessment dataset (xView2) [126]. The annotations were labeled by the authors of the dataset, HADR experts, and imagery analysts from CrowdAI [126]. Lastly, we learned about the types of tools domain experts use in practice to identify damaged buildings. Many analysts rely on ArcGIS, a general-purpose mapping tool, to view and edit maps, including satellite imagery.

Various crowdsourcing platforms have been developed to engage the public in assessing building damage through satellite imagery. Features from tools that experts use, such as ArcGIS, are included in many of these crowdsourcing platforms. These include segmenting buildings and panning or zooming in on satellite images. For example, MapSwipe provides a platform with such features for volunteers to contribute to several tasks, including damage assessment and change detection in satellite imagery [8]. CrowdAI also partners with governmental institutions, such as the California Air National Guard, to leverage crowdworkers to assess building damage in real-time during wildfires [7]. Given that platforms exist to crowdsource initial building damage assessments, we determined that crowdworkers would be a reasonable cohort for our study.

3.2.2 Task Generalization

Localizing and grading damaged structures from satellite imagery is becoming a popular classification and object detection task in deep learning and computer vision. Building damage assessment in satellite imagery requires the analysts to pan and zoom in on a pair of before and after images to identify any abnormalities. The task of comparing two images to identify and annotate differences between images generalizes beyond damage assessment. For example, radiologists compare multiple views to make diagnostic interpretations in breast imaging [116]. Thus, we believe that the insights derived from satellite imagery assessment tasks could be extensible to other domains that involve computer vision applications.

3.2.3 Selected Dataset

Designed for supporting humanitarian assistance and disaster relief (HADR) research, the xBD dataset covers a wide range of natural disasters that frequently occur, such as floods, earthquakes, wildfires, and hurricanes [126]. This dataset offers before and after images, ground truth assessments, and annotations for the damage. With the ground truth labels available, we were able to train participants to get familiar with the task at hand. This dataset is the official dataset used to develop state-of-the-art object detection and classification models that are currently deployed in a limited capacity in the real world to aid in initial building damage assessments.

Since most participants are likely to be new to assessing building damage from satellite imagery, we deemed it not beneficial to present participants with images that had more than one damaged structure with different levels of damage. This choice was also made to reduce potential confounders. We filtered out 96.53% of images from the xBD dataset that had more than one damaged structure with different levels of damage or no damaged structures.

3.3 Study 1: Identifying Human Explanation Strategies

In this study, we address **RQ1** by eliciting explanations from humans during a visual decision-making task: building damage assessment from satellite imagery. To inform the design of human-centered explainable AI techniques for visual decision-making tasks, we address the following two questions:

- What kinds of strategies do humans employ to explain their rationale for their damage assessment?
- What are the most prevalent strategies among those identified?

The main goal of this study is to uncover strategies that humans employ when explaining their decision to understand human-preferred explanation strategies better.

To uncover such strategies, we created an interface to elicit explanations from crowdworkers while making decisions. Depending on the strategy used, one explanation may reveal different information or features about the decision than another explanation that uses a different strategy. Understanding these strategies and how contextual information is incorporated into such strategies can inform new strategies for designing novel human-centered explainable AI techniques.

3.3.1 Study Design

3.3.1.1 Recruitment & Participants. To collect explanations, we performed our study on an online participant recruiting platform, Prolific. For a consistent participant base, we recruited workers who use English as their first language, currently reside in the United States, have an approval rate of 95%, and have a record of at least 50 task submissions. After screening out incomplete responses, responses from a total of 60 workers ($M_{age} = 34.07$, $SD_{age} = 8.28$, 26 female) were collected. Once the task was completed, each worker was assigned an anonymous ID and paid 6.50 USD for their participation. To motivate the participants, we offered a bonus payment of 1 USD to those who received the highest scores on their assessments (25% of the participants were awarded bonuses).

3.3.1.2 Data Selection Method. About half of the image sets in the xBD dataset contain no building damage (46.37%). Furthermore, of the images with damaged buildings, 89.47% contained multiple buildings. Our experimental design focuses only on images with one damaged building (5.64% of images) or multiple damaged buildings with the same damage level (18.26%), so we could isolate the assessment and localization for one individual target. Furthermore, this allowed us to focus our training on a single target. For the training session, we curated a set of 27 image sets for various natural disasters (seven floods, 13 hurricanes, five wildfires, and two tsunamis). For the main session, we curated a set of 10 cases for a variety of natural disasters (four hurricanes, three floods, two wildfires, and one earthquake), each containing pre- and post-disaster images of a damaged building. The chosen images we used are available on our GitHub site: training set, main set.

3.3.1.3 Training Session. Before starting the main session, participants are required to complete a training session. The objective of the training session is to ease participants into the annotation interface and get familiar with the building damage assessment tasks. Participants are provided with detailed criteria on how to assess building damage to help them learn the differences between the levels of damage (*i.e.*, no damage, minor damage, major damage, destroyed) [126]. With these criteria in mind, participants are asked to complete the training session by assessing damage in several sets of satellite images. The training interface seen in Figure 3.1 shows the pre- and post-disaster satellite images with four different damage assessment options.

The training phase is composed of three sub-sessions, with slightly different func-

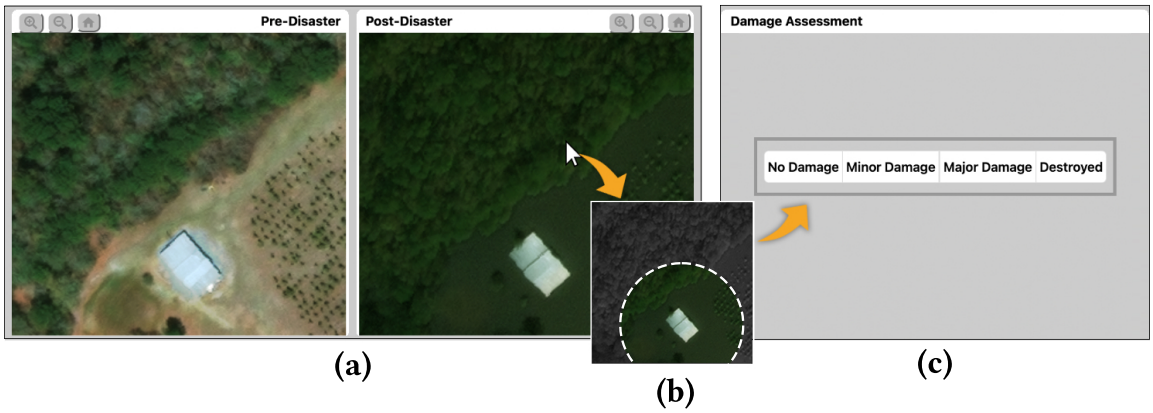


Fig. 3.1: Interface of the training session used to familiarize users with the task of building damage assessment. (a) Two satellite images are presented side-by-side, and (c) the user assesses damage based on four options: no damage, minor damage, major damage, and destroyed. For the first and second sub-sessions, (b) users are shown an overlay in the image highlighting where the damage is when hovering over the image to help them locate the building

tionalities in each to help familiarize participants with the task and navigate the interface. Each sub-session shows nine pairs of satellite images in groups of three. The participant is alerted of which damage assessments they got incorrect and correct. The participant must select the correct damage assessment for each pair of images in a group before seeing the next group. The first sub-session shows a clue (seen in Figure 3.1-(b)) of where they should look to assess the level of damage when they hover over the imagery. However, they do not have zoom or pan functionalities. The second sub-session adds zoom/pan functionality with the clue. Finally, in the third sub-session, the clue is taken away, but the zoom/pan functionality remains. The sub-sessions were designed to help the participant understand what features to look for in satellite images when assessing the damage. The average time each participant spent completing training tasks was 13.26 minutes ($SD = 5.52$).

Even if the participants may have become sufficiently accustomed to the interface and assessing damages in the training session, we presumed that they still may not be familiar with generating annotations on top of the images. Thus, we let participants go through a tour of the annotation interface designed for the main session. This tour introduces participants to the different annotation tools offered and how to add text to rationalize their annotations and assessments.

3.3.1.4 Main session. After successfully completing the training session, each participant was presented with a pair of satellite images of a specific geographical region before and after a natural disaster. We presented 10 sets of satellite images that we curated from the xBD data set [126]. Participants were asked to identify the damaged structure in the post-disaster image, indicate the level of damage (*i.e.*, no damage, minor damage, major damage, destroyed), and explain why they think it has that level of damage. Participants were told that their explanations would be shown to other

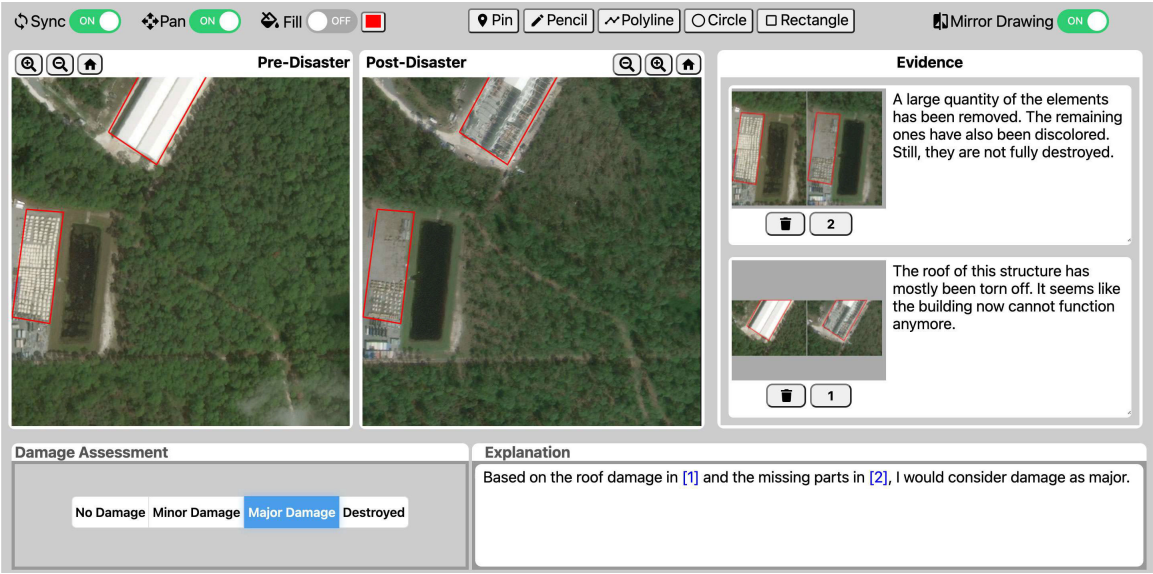


Fig. 3.2: Overview of the interface for the main task. Participants are presented with several different annotation tools to highlight evidence that helps explain their damage assessment. Participants provide explanations for each annotation they add and a global explanation for the overall pair of images.

participants in the study and that their goal was to, "convince the other participants that your judgment of damage is correct". To mitigate possible biases that stem from the sequence, we assigned each participant to one of four different sequences to present pairs of images.

While assessing image sets, participants were offered a wide range of drawing tools (rectangle, polyline, pin, circle, pen) and viewing functionalities (pan, zoom, mirror-drawing) to help explain their assessment choice, as shown in Figure 3.2. Mirror-drawing allows participants to synchronize their annotations for the pre- and post-disaster images. When mirror-drawing is selected, the participant only has to draw their annotation once, which will be reflected on the other image at the same location. The annotation tools allowed participants to have flexibility in locating damaged buildings and choosing how to generate visual components of their explanations (noted as 'Explanation' on the bottom right side of the interface in Figure 3.2).

Once participants draw a particular markup using the provided drawing and viewing tools, our system shows a text field to allow participants to explain the marked-up region via text. We will refer to the marked-up region as the image-based annotation and the accompanying text as the text-based annotation. Participants can reference the image- and text-based annotations by citing them within their global explanation (seen in Figure 3.2).

After completing the task, each participant was asked to assess the clarity of the guidelines we offered, along with their performance based on the 5-point Likert scale. Additionally, to identify possible enhancements for designing the interface, we asked

them to freely note any tools that were difficult to use and challenging tasks, as well as their comments on the study. The average time elapsed for the main phase was 23.71 minutes ($SD = 12.19$). The average number of image- and text-based annotations per participant was 14.73 annotations ($SD = 4.07$) and an average of 1.47 annotations per image ($SD = 0.85$).

3.3.1.5 Analysis. To develop codes that captured the participants' explanation strategies, we adopted a "coding reliability" approach to thematic analysis [40]. Two members of our research team independently analyzed participants' image- and text-based annotations while taking notes on explanation strategies they observed. Our research team discussed these notes and iteratively synthesized them into higher-level codes. Once all team members agreed on a set of codes to characterize high-level qualities of participants' explanations, two team members went back through the full dataset. They applied this set as a code book, labeling each response with one or more of these codes.

3.3.2 Results

Although we provide quantitative results on the frequency of strategies used by participants during decision-making (Figure 3.3), the main findings of this study are the qualitative results summarizing the core strategies employed during decision-making (Table 3.1).

Table 3.1: Summary of the strategies the research team identified from doing a thematic analysis on the responses from Study 1. Grey rows indicate major codes, each of which is followed by the sub-codes (except for code C, which contains no sub-codes.)

Code	Definition	Example
A	Constructing a causal story to explain building damage	
A-1	Use evidence of natural disaster/lower-level cause to argue that there was damage on objects	*"From the evidence of flooding, the structure seems to have been affected"*
A-2	Use evidence of structural/sub-structural damage/surrounding area to argue that there was the effect of natural disaster/lower-level cause	*"Since the structure seems to be underwater, there must have been a flood"*
A-3	Multi-step causal chain argument	*"From the evidence of the flood, I suspect the flood has moved and damaged the structure"*

Code	Definition	Example
B	Contrasting pre- and post-disaster imagery	
B-1	Comparison of the structure	See Figure 3.4-(a)
B-2	Comparison of the surrounding area	See Figure 3.4-(b)
B-3	Comparison of the sub-structures	See Figure 3.4-(c)
B-U	Falls under B, but without using a mirror-drawing or objects unidentifiable	See Figure 3.4-(d)
C	Highlighting affected part of a building	See Figure 3.4-(c)
D	Explanations based on the extent of damage to a specific building	
D-1	Based on the amount or proportion of the structure that appears to be damaged	*"The building is not completely destroyed but it has lost the majority of its roof"*
D-2	Based on their inference about the possibility of being recovered someday	*"Looks like it can be recovered someday"*
E	Explaining reasons for lack of confidence in their own assessment	
E-1	Due to confusing artifacts in images	*"It's hard to tell due to the shadow"*
E-2	Due to the changes irrelevant to disaster between pre- and post-, or the elapsed time between them	*"Seems like new buildings had been built"*
E-U	Other reasons or without any reason	*"This image is a bit mystifying to decode"*
F	Using the number of damaged structures in an image as the measure for severity of the disaster	
F-1	Structures only	*"There are some structures remaining but most have been very damaged"*
F-2	Structure + surrounding area	*"Small area burnt. Home intact"*

Code	Definition	Example
F-3	Surrounding area only	*"Majority of the water has been removed"*
F-U	Non-identifiable	Simply indicating region as 'area'

3.3.2.1 Identified Strategies. We uncovered six explanation strategies and labeled the responses based on these strategies with a high overall inter-rater reliability (Krippendorff's α) of 0.75. The definition and a concrete example for each strategy are listed in Table 3.1. The table consists of three columns: code, definition, and example. The code is a short identifier to represent the strategy. The definitions in the gray rows represent the general strategy observed, and the definitions in the white rows represent more specific subtypes of strategies. The frequency of each strategy identified by the research team is shown in the Appendix of the published manuscript.

3.3.2.2 Representative Strategies. Although we identified six major strategies (A - F), strategies C and E were so sparse (approximately 10% of overall user assessments) that it was impossible to identify the exemplar case for certain pairs of satellite images. Thus, strategies A, B, D, and F were curated as the *core explanation strategies*. In Figure 3.3, we provide a breakdown of the number of participants who used a given explanation strategy for each image. The four strategies for a given image may exceed 60 because some participants used multiple strategies within their response. We observed that code B (comparing pre- and post-disaster images) was the most prevalent among the responses for all images. We acknowledge that this result may be influenced by the interface design showing the two images side by side, making it natural to compare them in their response. The frequency of a strategy observed in Study 1 should not be associated with how that strategy will impact decision-making.

In the rest of this section, we provide detailed descriptions and examples of each major explanation strategy we identified through our analysis.

Strategy A: Constructing a causal argument to explain building damage. Rather than directly referencing the visual characteristics of a building, participants instead pointed to visual evidence of a natural disaster in the surroundings of a building to explain their assessment of building damage (A-1; *e.g.*, *"From the evidence of flooding, I would say the building seems to have been affected"*). In other cases, the participants inferred that a particular type of a natural disaster had occurred based on evidence of damage to a building and then explained their overall assessment of building damage with reference to the type of disaster (A-2; *e.g.*, *"The building has roof damage. Probably a hurricane came and hit it"*). Some participants constructed more complex, multi-step causal arguments (A-3; *e.g.*, *"**(Step 1)** There was a fire and **(Step 2)** it was a wildfire that took everything from the building. **(Step 3)** You can only see the outline of the building"*).

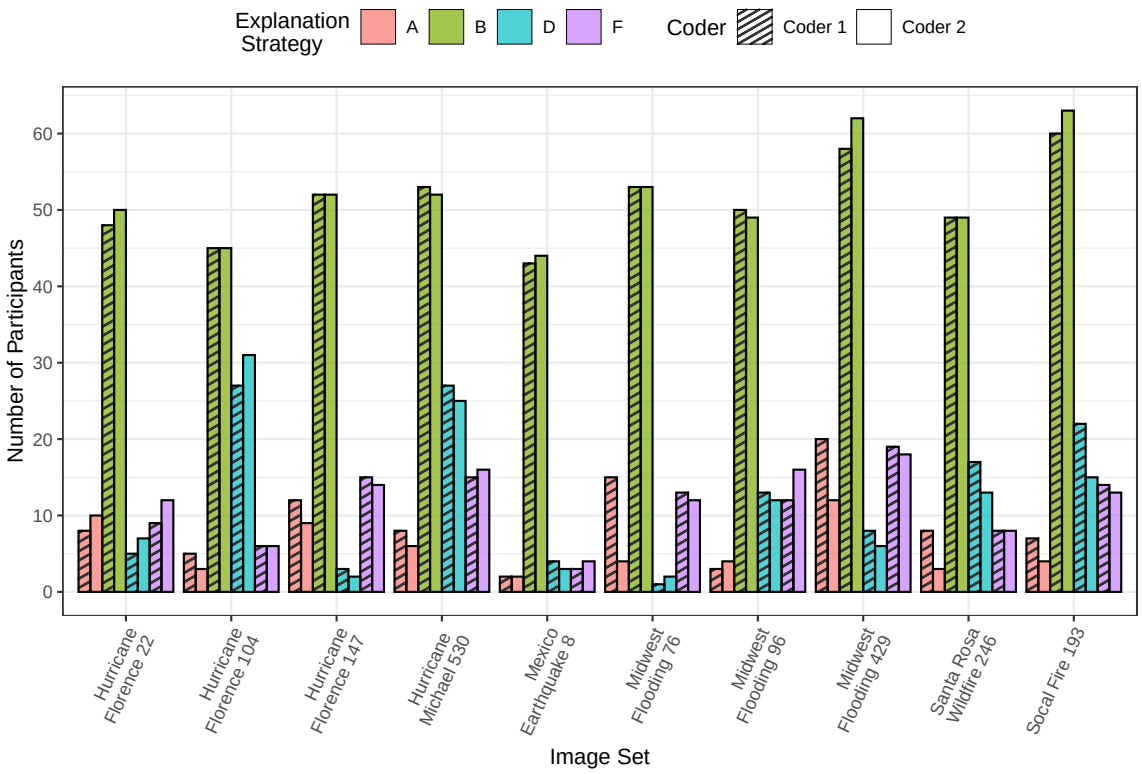
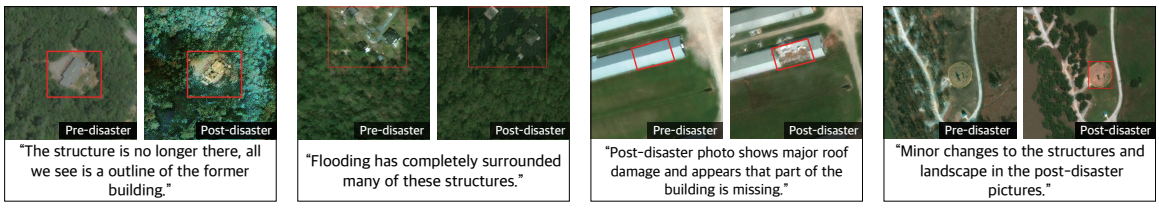


Fig. 3.3: The number of participants who used each core explanation strategy for each image set was reported by both coders. Some participants used multiple explanation strategies in their response to a given image set.

Strategy B: Contrasting pre- and post-disaster imagery. As seen in Figure 3.3, most participants explained their damage assessments through direct comparisons across image sets (*e.g.*, by creating an annotation using the mirror drawing tool). In more than 200 cases, participants referenced contrasts in the appearance of a specific building between the pre- and post-disaster images (B-1; see Figure 3.4-(a)) or contrasts in the appearance of specific substructures of a building (B-3; see Figure 3.4-(c)). Additionally, participants often directly compared the appearance before and after the disaster of the area surrounding a building (B-2; see Figure 3.4-(b)). Finally, ambiguous cases in which people generated contrast-based explanations but did not clearly specify which elements of an image they were comparing were marked with the sub-strategy B-U (see Figure 3.4-(d)).

Strategy D: Explanations based on the extent of damage to a specific building. For some cases, participants explained their assessment of the level of damage to a given building based on the *proportion* of the building that appears to be damaged (D-1; *e.g.*, *"Approximately half of the building was collapsed"*). Interestingly, even when significant damage to the building was evident, some participants rationalized lower



(a) Example of B-1 **(b) Example of B-2** **(c) Example of B-3 and C** **(d) Example of B-U**

Fig. 3.4: Examples of sub-strategies B-1, B-2, B-3, B-U, and C

damage assessments by arguing that the damage appeared repairable (D-2; *e.g.*, *"One part (of the building) was hit ... seems like it could be rebuilt"*).

Strategy F: Using the number of damaged structures in an image as a measure of the severity of the disaster. While strategy D captures cases where participants explain their damage assessments with reference to the apparent extent of damage to the building itself, strategy F is when participants explain their damage assessments with reference to the overall extent of damage observed in the image as a whole. For example, some participants explained their building damage assessment with reference to the number of other buildings that appeared to be affected (F-1; *e.g.*, *"It appears that one building has disappeared, leading me to believe it was destroyed. However, the remaining buildings seen are unharmed"*).

Furthermore, other participants explained their damage assessment with reference to the extent of damage visible in the area surrounding a building, including building damage (F-2; *e.g.*, *"None of the large buildings appear to be damaged, but there is evidence of a large mud patch (in the surrounding area), indicating some minor flood damage"*) or excluding building damage (F-3; *e.g.*, *"All trees have been damaged or destroyed"* and *"Flooding has completely surrounded many of these structures"*). Finally, ambiguous cases were marked as F-U (*e.g.*, *"Every area was totally destroyed"*).

3.4 Study 2: Impact of Human Explanation Strategies

Before recommending that researchers start developing techniques to create explanations in a similar manner to humans, as we found in the first study, we need to evaluate whether they might actually be helpful. We uncovered several explanation strategies that people use when explaining their decision for building damage assessment by eliciting explanations from them in Study 1. Based on these core strategies from Study 1, we conducted a follow-up study to explore whether and how each explanation strategy impacts decision-makers' performance if these strategies were to be implemented as actual human-centered XAI (HCXAI) techniques. This study will allow us to further inform the design of HCXAI techniques by quantifying their potential impact on decision-makers during human–AI collaboration. A pictorial overview of

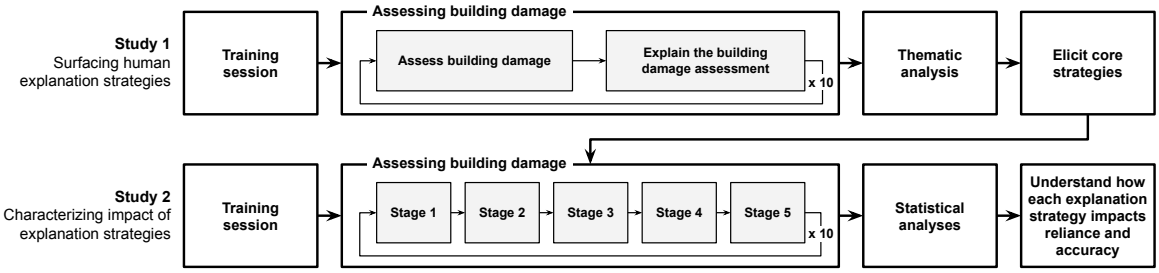


Fig. 3.5: High-level overview of the study design and data analysis done for the two studies. Participants of both studies went through the same training session. Study 1 participants assessed building damage and provided explanations for their assessment. We did a thematic analysis of the explanations provided by participants and identified four core strategies. Representative explanations of each core strategy from Study 1 participants were used in Stages 3 through 5 in the second study, while a separate set of participants assessed building damage. We analyzed how the participants assessed building damage throughout the five stages to understand how the core strategies impacted accuracy and reliance.

how this study uses findings from the first study is shown in Figure 3.5.

3.4.1 Study Design

We explore **RQ2 and RQ3** by framing the explanations from Study 1 as if they were generated by AI and showing them to a new set of participants doing the same task, building damage assessment. The study is designed as if we are using AI to identify how the explanation strategies might impact decision-making if they were developed and implemented as an XAI technique. Specifically, it allowed us to evaluate whether and how particular human explanation strategies impact decision-makers' accuracy and reliance on AI.

3.4.1.1 Recruitment & Participants. We followed the same recruitment procedures as in Study 1: we collected results from 60 participants ($M_{age} = 34.93$, $SD_{age} = 12.51$; 33 female) through Prolific, who have English as their first language, who currently reside in the United States, who have an approval rate of at least 95%, and who have a minimum of 50 submissions. We excluded workers who participated in Study 1 due to the significant overlap in content. We also excluded two participants who failed an attention check. Each worker was compensated 6.50 USD for their participation, with an anonymous ID assigned to analyze their responses. To incentivize active and high-quality participation, we offered a bonus of 1 USD for those who scored within the top 50% of all participants. Of the 60 participants, 30 participants received bonuses.

3.4.1.2 Training Session. The training for this study consisted of two phases: the first phase was the same training phase from the first study (seen in Figure 3.1) where participants were shown nine pairs of satellite images in groups of three to

get familiarized with building damage assessment. In the second training phase, we provide a walk-through of the damage assessment tool and highlight the additional information that is provided in each stage. Each participant is assigned a pair of satellite images based on the satellite images they will not see during the main task. The visual and text-based annotations for the example task were manually created and are available in the Appendix of the published manuscript.

3.4.1.3 Main Task. Participants were asked to assess building damage from several pairs of pre- and post-disaster satellite images. During the task, participants in this study were presented with human-generated outputs (framed as AI outputs) sampled from four different scenarios that emerged in the data from the first study:

- **Correct assessment with correct localization**: The Study 1 participant's assessment and visual annotation matched the ground truth assessment and the ground truth position of the damaged building.
- **Correct assessment with partially correct localization**: The Study 1 participant's assessment matched the ground truth assessment, while their visual annotation included a portion of the ground truth position of the damaged building.
- **Correct assessment with incorrect localization**: The Study 1 participant's assessment matched the ground truth assessment, while their visual annotation was completely wrong.
- **Incorrect assessment**: The Study 1 participant's assessment was incorrect.

We identified a set of representative examples for each explanation strategy and each scenario detailed in Section 3.4.1.4. Participants were randomly assigned to one of four groups (Table 3.2) with a total of 15 participants for each group. Each participant saw a total of four different image sets, each corresponding to one of the four scenarios. Different explanation strategies were paired with different scenarios across the four groups. In addition, within each group, these image sets were presented to the participant in a random order to minimize possible order effects.

The user went through five stages within each scenario, as shown in Figure 3.6. We identified five stages for the participant to go through so that we could observe the impact of each component separately. Stage 1 shows the participants' baseline accuracy on the images presented in a given scenario *before* they are shown any information from the AI. Stage 2 shows participants' accuracy after being presented with the AI's damage assessment (but before they can see the AI's annotations or explanation). Stage 3 adds localization, Stage 4 adds local explanations, and Stage 5 adds global explanations. By isolating these types of explanations, we can quantify the individual impact as if the AI presented them to the participant. These stages are the same across the four scenarios, revealing the same type of information but for different cases. In some scenarios, the ground truth label is incorrect, or the ground truth localization is incorrect, so our study is realistic, like an imperfect AI. We compare the reliance and performance in each stage across these scenarios to identify how the type

Table 3.2: Four different groups were created to evaluate each explanation code in each scenario. We show which code was assigned to which scenario for each group and which image set was used for that explanation code/scenario.

Group #	Scenario	Explanation Strategy	Image Set
1	*Correct Assessment, Correct Localization*	A	*Hurricane Florence 22*
	Correct Assessment, Partially Correct Localization	B	*Hurricane Michael 530*
	Correct Assessment, Incorrect Localization	D	*Midwest Flooding 96*
	Incorrect Assessment	F	*Midwest Flooding 76*
2	*Correct Assessment, Correct Localization*	B	*Socal Fire 193*
	Correct Assessment, Partially Correct Localization	D	*Hurricane Florence 22*
	Correct Assessment, Incorrect Localization	F	*Hurricane Michael 530*
	Incorrect Assessment	A	*Santa Rosa Wildfire 246*
3	*Correct Assessment, Correct Localization*	D	*Hurricane Michael 530*
	Correct Assessment, Partially Correct Localization	F	*Midwest Flooding 96*
	Correct Assessment, Incorrect Localization	A	*Hurricane Florence 22*
	Incorrect Assessment	B	*Hurricane Florence 147*
4	*Correct Assessment, Correct Localization*	F	*Hurricane Florence 104*
	Correct Assessment, Partially Correct Localization	A	*Hurricane Florence 22*
	Correct Assessment, Incorrect Localization	B	*Mexico Earthquake 8*
	Incorrect Assessment	D	*Midwest Flooding 429*

of information provided impacts the decision-making because the only factor changing is the accuracy of the AI; the type of information provided remains unchanged.

At each stage, the participant was asked to provide their damage assessment and use a pinpoint marker to identify exactly where they thought the damage was in the image. For each image, we also asked the user to indicate how confident they felt in their damage assessment (on a 4-point Likert scale from very unsure to very confident) and how helpful they thought the added information was (on a 4-point Likert scale from very unhelpful to very helpful) when making their damage assessment. Finally, we provided an optional text box to allow the user to briefly detail whether and how they believed the AI assessment and explanations may have affected their own damage assessment. Since this was optional, not all participants responded. The task took on average 28.79 minutes (SD_{time} = 13.37 minutes).

3.4.1.4 Data Selection Method. The satellite images and explanations presented to participants in this study were selected from the responses from the first study. The curated set is available on our GitHub site. One coder analyzed every response from the first study to evaluate the quality of the image- and text-based annotations

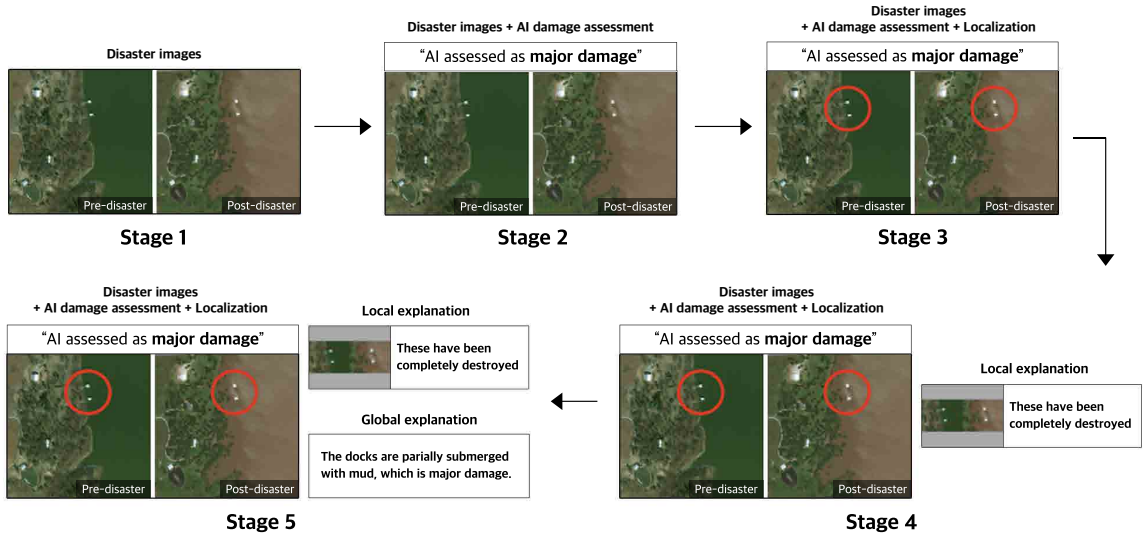


Fig. 3.6: The second study consisted of five stages. The five stages are shown four times, one time for each scenario. Stage 1 shows the pre- and post-disaster image, with no information from the AI; Stage 2 adds information by showing how the AI assessed the damage; Stage 3 adds visual annotations, highlighting points or regions of the image where the AI thinks it is important to look; Stage 4 adds local text-based annotations to elaborate on each visual annotation, and Stage 5 adds a global explanation detailing why the AI made the damage assessment that it did. Each participant went through these five stages for each of the four scenarios.

provided. The coder reviewed observations where the participant's damage assessment was correct, both coders from Study 1 agreed on the explanation strategy, and the explanation strategy was a core explanation strategy from Study 1. From the filtered observations, the coder mapped the image-based annotations into three categories:

- **Correct Localization**: The participant identified the ground truth position of the damaged building with an image-based annotation.
- **Partially Correct Localization**: The participant's image-based annotation included a portion of the ground-truth position of the damaged building.
- **Incorrect Localization**: The participant's image-based annotation does not include any portion of the ground truth position of the damaged building.

After coding the annotations in each observation, the coder evaluated the quality of the text-based annotations for observations where the participant's damage assessment was correct. We specifically selected observations that used only one explanation strategy to allow us to isolate the effectiveness of each strategy by itself. We chose from observations that did not cite the annotations[2] (66% of the observations and 63% of the participants did not cite the annotations[3]) to limit possible confusion in references and separate the impact of text-based annotations from global explanations.

[2]As seen in the explanation in Figure 3.2
[3]See Appendix in the published manuscript for full results.

3.4.2 Quantitative Results

We evaluated the core explanation strategies across different scenarios (correct, partially correct, and incorrect localizations) to identify whether and how these may have impacted participants' assessment accuracy and reliance on AI assessments. To address **RQ2**, we present participants' accuracy across the five stages for each explanation strategy and scenario in Figure 3.7. The accuracy of the assessment is calculated by the percentage of participants who selected the ground truth assessment of the damage. Localization accuracy is calculated by the percentage of participants who placed the pinpoint on the ground truth damaged area.

Throughout the remainder of this section, we evaluate the impacts of a given explanation strategy by examining the change in a given measure between Stage 5 (*i.e.*, all AI outputs shown) and Stage 2 (*i.e.*, AI damage assessment only). Where appropriate, we also examine changes between other stages, for example, to understand the impacts of presenting text-based annotations over and above visual annotations. We provide a detailed interpretation of our results below.

3.4.2.1 Causal explanation strategies mislead humans less. In scenarios where the AI damage assessment was actually incorrect, we found that causal explanations (strategy A) misled humans less than other explanation strategies. An ANOVA for

(a) **Difference between Stage 2 and Stage 5**. P-values from post-hoc Tukey HSD test for the *correct assessment, incorrect localization* scenario measuring localization accuracy from stage 2 compared to Stage 5. There are statistically significant differences between strategies A to B and D for localization accuracy.

Correct Assessment, Incorrect Localization Scenario Metric: Localization Accuracy	
Strategy Pairs	**Tukey HSD p-value**
A - B	**0.019**
A - D	**0.019**
A - F	0.275
B - D	0.900
B - F	0.607
D - F	0.607

(b) **Difference between Stage 2 and Stage 5.** P-values from post-hoc Tukey HSD test for the *incorrect assessment* scenario measuring assessment correctness from Stage 2 compared to Stage 5. There are statistically significant differences between strategies A to B, D, and F for the participant's assessment correctness.

Incorrect Assessment Scenario Metric: Assessment Accuracy	
Strategy Pairs	**Tukey HSD p-value**
A - B	**0.008**
A - D	**0.035**
A - F	**0.001**
B - D	0.900
B - F	0.900
D - F	0.662

Table 3.3: Post-hoc Tukey HSD tests for determining statistically significant differences between the participant's localalization and assessment accuracy.

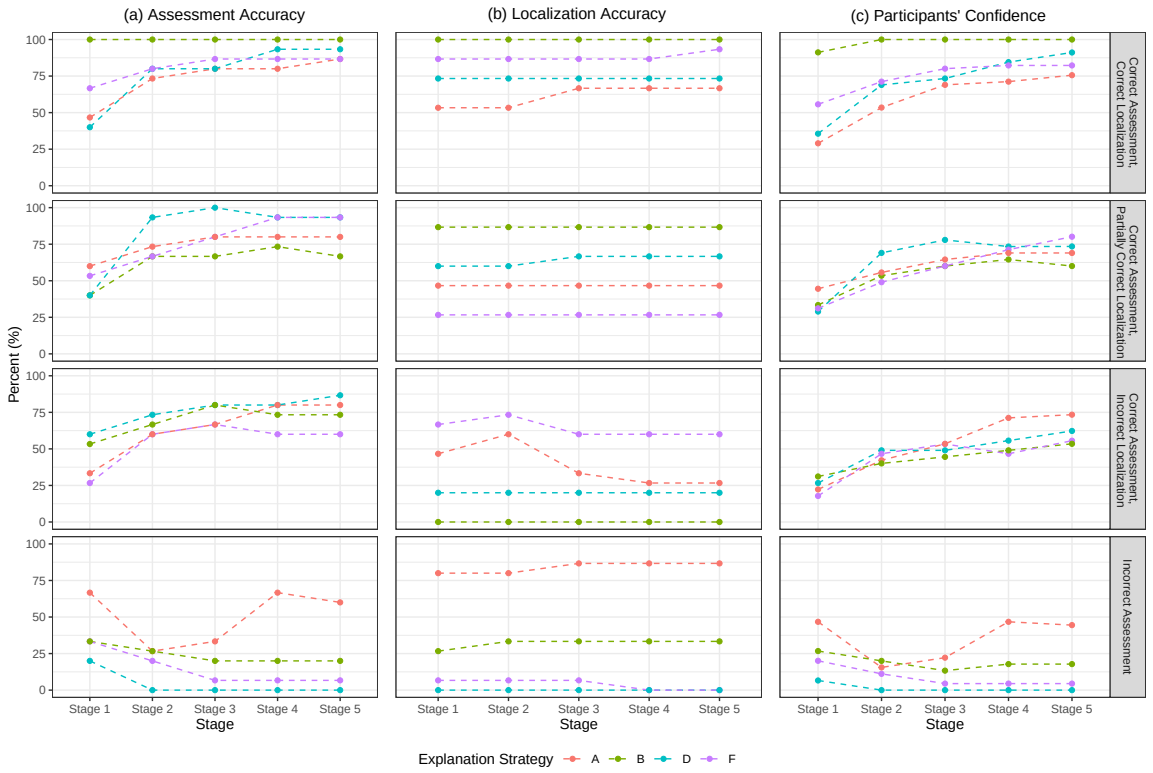


Fig. 3.7: Assessment accuracy, localization accuracy, and the participants' confidence. The facet on the y-axis shows the four different scenarios that the participants were presented with. The percent on the y-axis represents the ratio of participants (a) that got the damage assessment correct, (b) that placed the pinpoint on the damaged structure, and (c) the confidence level for correct assessments. The x-axis shows the five stages described in Figure 3.6.

the *incorrect assessment* scenario ($p < 0.001$) indicated a significant difference in the impacts that different explanation strategies have on assessment accuracy. As shown in Table 3.3b, a post-hoc Tukey HSD test for this scenario revealed a statistically significant difference in the Stage 5 - Stage 2 slope between explanation strategy A (involving a causal argument to explain a building damage assessment) and all other core strategies. This may suggest that humans are better at calibrating their reliance on AI assessments by reasoning about a causal argument versus by assessing other kinds of explanations. In line with this interpretation, one participant who switched back their assessment to their initial belief stated, "*[I] disagree with the reasoning, but it convinced me that my initial assessment was correct*". In this case, the participant inferred that poor explanations might imply poor accuracy for AI assessments.

In contrast to strategy A, the accuracy for all other explanation strategies decreased as participants switched their original correct assessments to agree with the AI's incorrect assessment in the *incorrect assessment* scenario. Strategy A shows a steep increase in accuracy from Stage 3 to Stage 4 (Figure 3.7), suggesting that the local

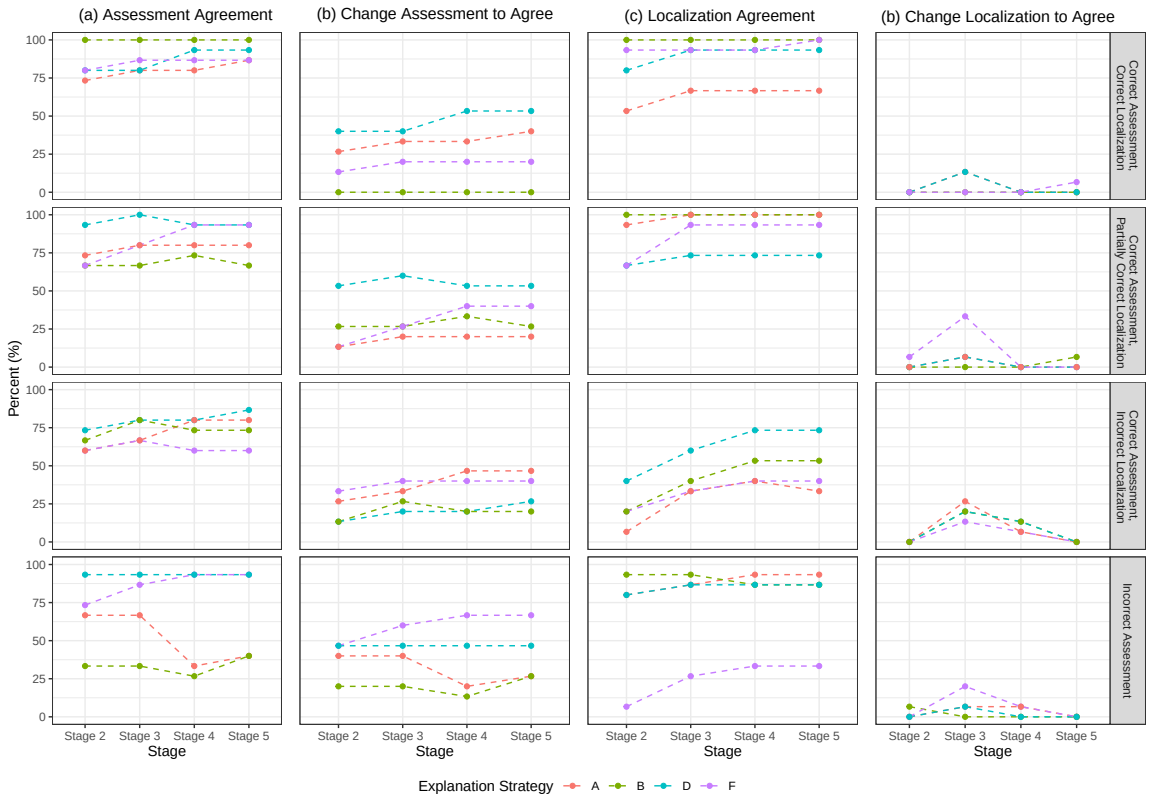


Fig. 3.8: Measurements to evaluate reliance on AI. The facet on the y-axis shows the four different scenarios that the participants were presented with. The percent on the y-axis represents the ratio of participants (a) who agreed with the AI's assessment, (b) who changed their assessment to agree with the AI's assessment, (c) who agreed with the AI's localization of the damage, and (d) who changed their pinpoint to agree with the AI's localization of the damage. The x-axis shows the five stages described in Figure 3.6.

text-based annotations revealed in Stage 4 may have played a large role in helping participants calibrate their reliance on the AI.

3.4.2.2 Incorrect localizations within causal explanations could lead to incorrect rationalizations. In cases where participants were shown a correct assessment, but an incorrect localization from the AI, causal explanations (strategy A) misled human localization more often than other explanation strategies. An ANOVA for the *correct assessment, incorrect localization* scenario indicated a significant difference in the impacts that different explanation strategies have on localization accuracy ($p < 0.01$). A post-hoc Tukey HSD test revealed statistically significant differences in the Stage 5 - Stage 2 slopes between explanation strategies B and D versus A (see Table 3.3a). Strategy A shows a sharp drop in accuracy from Stage 2 to 3 (Figure 3.7), suggesting that the presentation of local visual annotations, as part of a causal explanation,

may have played a large role in encouraging inappropriate reliance upon incorrect AI localizations.

Interestingly, the presentation of the incorrect localization in Stage 3 resulted in several participants changing the location of their pinpoint for strategy A, with a strong increase in overall confidence (Figure 3.7; third row, middle column). In this case, the participants could have associated incorrect features (corresponding to the incorrect localization) with the correct damage assessment, potentially influencing how they made damage assessments later in the study.

3.4.2.3 Causal explanation strategies decrease over-reliance on AI. When shown causal explanations (strategy A), we found that humans relied less on AI when the assessment was incorrect. Drawing on the prior literature, we report multiple related behavioral measures in Figure 3.8 to better understand participants' reliance on AI (**RQ3**) [352, 388]:

- *Assessment Agreement.* How often participants' assessments agree with the AI's assessment.
- *Change Assessment to Agree.* How often participants choose to change their assessment to agree with the AI's assessment.
- *Localization Agreement.* How often participants agree with the AI's localization of damage in the image (based on the location of the participant's pinpoint).
- *Change Localization to Agree.* How often participants choose to change the location of their pinpoint to agree with the AI's localization of damage.

An ANOVA for the *incorrect assessment* scenario showed a significant difference in the impact that different explanation strategies had on assessment agreement (p ¡ 0.05). A post-hoc Tukey HSD test (Table 3.4) for this scenario revealed a statistically significant difference in the Stage 5 - Stage 2 slope between explanation strategy A (causal explanation) and B (comparing before and after features to explain a building damage assessment) as well as between explanation strategy A and F (identifying the number of damaged structures to explain a building damage assessment).

For strategy A in the *incorrect assessment* scenario where the AI localization is correct, we see a huge decline in assessment agreement when the text-based explanations are presented in Stage 4. These trends are also observed in Figure 3.7. The increase in disagreement with the AI at Stage 4 for strategy A could be the result of participants judging that the causal arguments for the incorrect assessment actually do not make much sense. In contrast to strategies A and B, in the *incorrect assessment* scenario, strategy F shows an increase in reliance according to all four reported reliance metrics. For instance, before showing the AI's localization in Stage 3, 6.7% of participants disagreed with the AI's localization. However, after seeing the explanations in Stages 3 and 4, 33% of participants agreed with the AI's localization. The participants were misled by explanation strategy F to adopt the incorrect localization, resulting in a very low assessment accuracy as seen in Figure 3.7.

Table 3.4: Difference between Stage 2 and Stage 5. P-values from post-hoc Tukey HSD test for the *incorrect assessment* scenario measuring assessment agreement from Stage 2 compared to Stage 5. There are statistically significant differences between strategies A to B and A to F for the participant's assessment agreement.

Incorrect Assessment Scenario Metric: Assessment Agreement	
Strategy Pairs	**Tukey HSD p-value**
A - B	**0.041**
A - D	0.140
A - F	**0.002**
B - D	0.900
B - F	0.675
D - F	0.367

3.4.2.4 Strategy type is less important when the AI is accurate and explanations are high quality. We found no significant differences in assessment accuracy or reliance across explanation strategy types in the scenario when the AI assessment was correct. By contrast, as described above in Subsection 3.4.2.1, different explanation strategies had differential impacts on participants' assessment accuracy and reliance in scenarios where the AI was actually incorrect.

3.4.2.5 Pre- and post-disaster comparisons do not have a significant impact. Although the most prevalent type of explanation strategy that the participants used was a comparison between the pre- and post-disaster images, we did not observe that this type of explanation significantly impacted task accuracy or reliance on AI across all scenarios.

3.4.2.6 Participants perceive explanations as most helpful when the AI is accurate. Overall, our results suggest that participants generally found the explanations to be the most helpful in scenarios where the AI assessment and localization were correct, confirming their initial damage assessment. Participants generally found explanations less helpful when the AI's localization was incorrect or when the localization was correct and the assessment was incorrect.

3.4.3 Qualitative Results

In this subsection, we report on participant responses to an optional question that was visible across every stage, asking participants whether and how they believed the AI affected their assessment for that stage. Since this question was optional, we did not have data for all participants. One coder analyzed the responses to

get a general understanding of the themes that emerged from the optional question responses. Overall, participants stated that the AI either confirmed their initial damage assessment or, in the scenario where the localization was incorrect, that the AI did not highlight meaningful regions. Below, we present examples of participant quotes, illustrating trends that emerged for each scenario.

When correct, the "AI" increased participant's confidence. Among participants who selected the correct damage assessment from the very beginning in the *correct assessment, correct localization* scenario, several noted that seeing the AI's assessment and explanations increased their confidence in their original decision. Participants who initially selected an incorrect damage assessment also described why they changed their assessment to match the AI's. For example, one participant who had originally selected the incorrect damage assessment changed to agree with the AI's assessment because "*...[the AI] came to the same conclusion and description of damage as me.*". A couple of participants reported that they chose to agree with the AI's assessment because the AI pointed out specific damage that they previously had not noticed. Interestingly, some participants noted that certain explanations were similar to their own line of reasoning. Four out of 15 participants noted that the AI provides descriptions that match their own thought process for strategy B and two out of 15 participants for strategy D.

When partially correct, participants critiqued the "AI". By contrast, in the *correct assessment, partially correct localization* scenario, very few participants made comments about the AI reinforcing their assessment or providing similar reasoning. Instead, participants focused on pointing out pieces of information that the AI did not mention or realizing a damaged area/structure they previously did not see. For example, one participant who originally had selected the incorrect assessment and later agreed with the AI said, "*It appears that I was looking at the wrong thing, the section that the AI has pointed out does have minor damage*". Another participant pointed out the flaws in the AI's assessment, saying, "*[The AI] was correct about the damages. AI didn't notice the homes destroyed with what looks like only a foundation of a house after the disaster*".

In the *correct assessment, incorrect localization* scenario, participants were split between agreeing with or not agreeing with the AI's localization. Four out of 15 participants for explanation strategies A, B, and D and three out of 15 participants for explanation strategy F noted that the AI did not identify any meaningful regions. One participant said, "*[The] AI annotation is over an area that did not have damaged structure*". However, a few participants were misled by the incorrect annotations, resulting in behaviors such as moving the location of their pinpoint to agree with the AI's localization of damage or changing their damage assessment. One participant stated, "*[The AI] helped to direct the pin location instead of a general area*". Three participants for strategy F, three for strategy D, and two for strategy A made comments

about the AI reinforcing their damage assessment and increasing their confidence, despite the AI being incorrect.

When the "AI" prediction was incorrect, the "AI" incorrectly reassured or influenced participants. For example, when the AI's damage assessment was *incorrect* and the participant's initial damage assessment matched the AI's assessment shown in Stage 2, several participants were incorrectly reassured by this and noted that the AI reinforced their original assessment. One participant said, "*It validated and reinforced my assessment. It gave me more confidence*". However, this participant changed their assessment to disagree with the AI and agree with the ground truth in Stage 4 and Stage 5 when shown detailed explanations from the AI about the annotations and overall assessment. The AI's incorrect assessment not only reassured participants who originally assessed incorrectly; it also incorrectly influenced participants who originally assessed correctly. In one case, when the AI predicted '*destroyed*' for an image that only presented '*minor damage*', the participant switched their correct assessment in Stage 1 to agree with the AI in Stage 2 noting, "*The AI made me reconsider, so I looked even harder for a destroyed building*". Interestingly, one participant that originally assessed the damage correctly and switched to agree with the AI's incorrect assessment then switched back to their original assessment because they disagreed with the reasoning of the AI's local explanations in Stage 4, stating, "*[I] disagree with the reasoning, but it convinced me that my initial assessment was correct*". The reasoning presented in the text-based annotations helped the participant understand the limitations of the AI. This example reflects a broader trend of strategy A (causal explanations), helping participants judge when *not* to agree with the AI in the *incorrect assessment* scenario, as shown in Figure 3.7.

3.5 Discussion

Through a sequence of two online studies, we identified explanation strategies that humans use to describe their decisions for a visual building damage assessment task and evaluated the impact of each explanation strategy on task accuracy and reliance on AI. Our findings offer insight into the types of explanation strategies humans employ when providing rationales for their decisions in the context of visual decision-making tasks. We introduce a new approach for exploring the impacts that prospective explainable AI techniques might have on human–AI decision-making by presenting participants with different types of human-generated explanations framed as AI explanations. Using this approach, we investigated the impacts of different human explanation strategies on human accuracy and reliance upon AI-based assessments.

Through a thematic analysis, we identified strategies humans use to explain their decision for a visual building damage assessment task. The most prevalent strategy humans used in their explanations was comparing the pre- and post-disaster images.

While this explanation strategy was the most prevalent strategy we observed in the responses, some caution is needed in interpreting this observation, as this pattern may be partially explained by the pre- and post-disaster images side by side.

3.5.1 The Mixed Effect of Explanations

Overall, we found that, compared to other explanation strategies, **causal explanations misled humans less and served to decrease over-reliance on AI damage assessments**. A wave of recent empirical results indicates that presenting AI explanations can often backfire, failing to improve or even *harming* human–AI decision-making in practice [16, 142, 165, 192, 194, 274]. For example, presenting explanations has sometimes been shown to promote over-reliance on AI recommendations, whether by lulling humans towards undeserved trust or by inducing cognitive overload [16, 194, 274]. In the context of this prior literature, our results help motivate the need to empirically investigate the impacts of different types of explanations on human–AI decision-making, across different real-world tasks and contexts. As others in the field have highlighted, AI explanations are not a monolith: there are many possible kinds of explanations, which may have different (potentially context-dependent) effects on human–AI decision-making [218, 232, 363].

In our study, we found that causal explanations had unique impacts among the explanation strategies we investigated: **in the context of erroneous AI damage assessments, causal explanations empowered humans to correctly second-guess the AI.** Our findings suggest that humans are better at calibrating their reliance on AI assessment by reasoning about causal arguments versus by assessing other kinds of explanations. This interpretation aligns with prior research on human causal cognition, which suggests that humans are predisposed to reason about the world in a causal rather than purely statistical or associative manner. In fact, some errors and fallacies that have been observed in human probabilistic or statistical reasoning can be understood as symptoms of this tendency: instances where humans attempt to use causal reasoning and assumptions in situations where these do not apply [15, 138, 193]. For these reasons, it may be that humans are better able to spot faulty reasoning in causal explanations versus other explanation strategies.

Although causal explanations helped humans identify erroneous AI assessments of the *extent* of damage, we found that **causal explanations led humans astray when the AI incorrectly identified the location of the damage**. Interestingly, our findings suggest that participants may have updated their own line of reasoning after seeing only the *visual annotation* components of causal explanations presented in Stage 3. The presentation of visual annotations did not have this effect on other explanation strategies, suggesting that the types of visual annotations that human explainers generate in the context of causal explanations are more persuasive than those associated with other strategies. It may be, for example, that the visual annotations associated with human-generated causal explanations tend to visually highlight potential causal

factors that influence human damage localization.

Interestingly, we found that in our context of building damage assessment from satellite imagery, causal explanations were most helpful in improving decision-making in cases where the AI's damage assessment was incorrect. On the contrary, **when the AI's assessment was correct, the type of explanation strategy presented did not significantly impact participants' accuracy or their reliance on AI assessments.** This suggests that causal explanations may be particularly valuable in settings where AI systems are likely to be highly imperfect, including high uncertainty settings or cases where an AI system is particularly vulnerable to blind spots [48, 180, 195].

3.5.2 Limitations

In this section, we briefly highlight key limitations across both of our studies:

3.5.2.1 "Who is the explanation for?". In Study 1, participants were asked to generate annotations and rationalizations for their damage assessment with the goal of convincing another person that their answer was correct. However, recent discussions in the human-centered explainable AI literature emphasize the importance of tailoring explanations to particular stakeholder groups (*i.e.*, model developers, business owners, frontline decision-makers, decision subjects, and regulatory bodies), who may have different use cases for AI explanations or different expertise through which to interpret explanations (*e.g.*, [165, 209]). In our study, we asked participants to convince "another person." However, it is possible that presenting a more specific prompt in Study 1 (*e.g.*, explicitly specifying that the user of the explanation would be another participant on the Prolific platform) would have yielded a different distribution of explanation strategies.

3.5.2.2 Potential Effects of Image Selection on the Identified Human Explanation Strategies. We filtered the xBD dataset to only use image pairs with one damaged structure or multiple damaged structures with the same damage assessment. This constraint limited the number of images we had represented in the dataset for each natural disaster. This also could have impacted the distribution of explanation strategies that we observed in our study. For instance, it is unclear how and to what extent the explanation strategies might differ when participants are presented with images that include several damaged regions with different damage assessments.

3.5.2.3 Instructed to annotate, not localize damage. The participants in Study 1 were not explicitly instructed to annotate the specific building they thought was damaged. They were only advised to mark evidence relevant to the level of damage for which they wanted to argue. Therefore, it is possible that some participants in Study 1 could have annotated evidence of damage to surrounding areas without annotating the specific building they believed was damaged. When selecting images for the *correct assessment, correct localization* scenario in Study 2, we only considered observations in which a damaged building was clearly marked. This limited the number of observations

we could choose from for the *correct assessment, correct localization* scenario.

3.5.2.4 Availability of Explanations from Study 1. We wanted to evaluate different explanation strategies and different variations of the explanation strategies. However, not all explanation strategies were represented equally for each image set[4]. It is possible that participants were naturally inclined to compare the pre- and post-disaster images due to the design of the interface showing the two images side by side. Participants in Study 1 were also limited by the tools we provided in the interface, which could have hindered the type of explanation strategies they would employ. In addition, to evaluate one explanation strategy at a time in Study 2, we were limited to observations from Study 1 that only used one explanation strategy. Due to these constraints, most image pairs within each scenario differ for each explanation strategy. This meant that assessing the impacts of particular explanation strategies independently of particular image pairs was not feasible. Thus, to control for baseline differences between image pairs (that is, differences in stage 1 before any AI outputs were shown to participants), our analyses compared *changes* in particular metrics (*e.g.*, accuracy and confidence) across stages, rather than comparing absolute values. Nonetheless, it is possible that we were still unable to observe certain effects due to differences across image pairs. For example, if participants' baseline accuracy for a given image pair was near 100% for a given scenario[5], then this may have masked explanation strategy effects that we would have otherwise observed (*i.e.*, due to a ceiling effect).

Some participants noted in the end-of-task survey that some of the images were low quality or hard to decipher due to the satellite images being taken at different times of the day or in different seasons. Unfortunately, the quality of the images is typical in this domain and a limitation of the open-source data set [126]. Participants struggling with these images could also be due to the fact that they are not domain experts and not used to analyzing lower-quality satellite imagery. Therefore, we acknowledge that using participants from an online crowdsourcing platform may impact the generalization of our findings to domain experts.

3.6 Future work

There are many opportunities for further work to build upon our bottom-up taxonomy of human explanation strategies. We plan to prepare and release an open-source dataset mapping explanation strategies to specific examples of human explanations that we collected in our study. This dataset will allow researchers to explore ways to generate explanations that are associated with a certain explanation strategy for image classification tasks. Generating these explanation strategies will allow researchers to

[4]See appendix in the full paper to see the distribution of explanation strategies from Study 1.
[5]See assessment accuracy for *correct assessment, correct localization* scenario in Section 3.4.2, Figure 3.7

evaluate the extent to which the explanation strategies impact a decision-maker's accuracy and reliance on AI on a larger scale.

Future work should consider running Study 2 on a larger scale. As discussed in our Limitations, in this study we were only able to evaluate one explanation strategy per image set. Future studies should consider evaluating all four explanation strategies on the same image, to better separate out the potential impacts of particular image sets versus explanation strategies. Furthermore, while this study focused on crowdworkers, researchers should consider recruiting subject matter experts to see how different explanation strategies impact their workflow and whether they have a preference for a particular strategy. In general, XAI for building damage assessment and disaster relief remains a critical yet underexplored research area. Future work is greatly needed to better understand how AI-based decision support can be designed effectively for this context.

Since the core explanation strategies we identified can be generalized to other visual decision-making domains, such as radiology, we encourage researchers to explore the impact that visual annotations paired with text-based annotations have on human accuracy and reliance on AI in a broader range of contexts. For example, in medical imaging, the majority of current explainable AI techniques focus on providing saliency maps [284]. The kinds of human-generated visual and text-based explanations presented to the participants in our study are much richer by comparison.

3.7 Conclusion

Explainable AI techniques are increasingly being evaluated to understand how they aid practitioners during the decision-making process. Findings from numerous user studies call for a more human-centered approach to explainable AI for human–AI decision-making. To address the need for human-centered explainable AI and understand its impact on the decision-making process, we conducted a series of two studies to understand the types of explanation strategies that humans use during visual decision-making tasks and to understand how presenting these explanation strategies (as if AI generated them) to human decision-makers impacts their accuracy and reliance on AI.

We identified four core explanation strategies for building damage assessment from satellite imagery: causal explanations (Strategy A), before-and-after comparisons (Strategy B), the proportion of structural damage (Strategy D), and the number of structures damaged (Strategy F). The most prevalent explanation strategy was comparing pre- and post-disaster images; however, this observation could be confounded by the interface design. Based on those four core explanations, we evaluate whether and how they impact humans' assessment accuracy and reliance on AI assessments. Our results show that causal explanations can help humans appropriately calibrate their reliance on AI damage assessments when the AI prediction is incorrect. However, causal explanations paired with incorrect AI localizations can also lead humans astray.

Chapter 4

The Role of Human Factors on Decision-Making with Imperfect XAI

This chapter is adapted from the following co-authored paper:

Morrison, K., Spitzer, P. (*co-first authors*), Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The impact of imperfect XAI on human–AI decision-making. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), 1-39.

Chapter 3 captures how humans can produce imperfect explanations for their assessments, which opens the door to acknowledging that an AI system could similarly produce imperfect explanations for its predictions. This chapter builds on Chapter 3 by capturing how a real AI system and AI explanations produced by that system impact how domain experts and non-experts make decisions when assisted by an imperfect AI system, where the XAI is designed to help users in verifying the AI's predictions.

4.1 Introduction

With the deployment of imperfect artificial intelligence (AI) in high-stakes decision-making scenarios, decision-makers struggle with knowing when they should and should not rely on AI advice, causing frustration and resulting in potentially harmful decisions. As a result, designing and developing human-centered explanations has become a core theme in computer-supported cooperative work (CSCW) and human–AI collaboration research [98, 209, 363]. Tangentially, recent work has proposed new explanation techniques that leverage machine learning models to explain the prediction of another machine learning model [17, 25, 137, 166, 187, 294, 321, 337, 346, 367, 383]. A subset of these studies propose to exploit language models to generate natural language explanations for image classifications with the rationalization that natural language is more "human-friendly" [137, 166, 337]. Aside from natural language explanations, another subset of recent work proposes advanced content-based image recognition techniques to generate example-based explanations [17, 294, 367]. These types of approaches to explainability introduce another level of uncertainty in the collaboration

between the decision-maker and the AI, as using another model to explain the original model can be imperfect.

CSCW, and more specifically human–AI collaboration, is prevalent across high-stakes scenarios [52, 216, 343, 361]. For instance, radiologists collaborate with AI to identify abnormalities in medical imagery [343]; conservationists use AI to help monitor biodiversity [24], and humanitarian aid workers use AI to help identify damaged buildings after natural disasters or armed conflicts from satellite imagery [238, 386]. While some of these human–AI collaboration scenarios require the human decision-maker to have several years of experience in the given domain, such as radiology, monitoring biodiversity with the help of AI doesn't necessarily require domain expertise [24, 269]. Platforms, such as iNaturalist [3], Merlin Bird App [5], and Wildbooks from WildMe.org [1], have allowed non-experts (*i.e.*, citizen scientists, hobbyists, or students) to partake in monitoring biodiversity alongside domain experts (*i.e.*, ornithologists and conservationists). While these platforms are valuable for non-experts to use, the AI models backing these platforms are imperfect: they do not always provide correct predictions [181].

Experts and non-experts interacting with the same imperfect AI and the same type of explanations in human–AI collaboration scenarios, such as decision-making [299] or learning systems [325], could result in some users misunderstanding or inappropriately relying on/overriding the AI advice. Experts may have more context outside of the AI's classification and confidence that a non-expert may not have. This might result in experts using their context information to appropriately rely on the AI when advice is provided, such as correctly overriding when wrong AI advice is presented and correctly using AI advice when it is correct. Non-experts, on the other hand, might not be able to judge the correctness of the AI advice appropriately, as they are missing this context information. For example, for bird species identification, ornithologists tend to be more aware of information related to the visual differences between the male and female birds for a given species, the bird's habitat, and migration patterns, whereas non-experts may not know some or all of that information. This same situation can arise in radiology, where residents ("non-experts") may initially be less familiar with certain diseases than an attending radiologist ("experts"). However, both experts and non-experts can struggle to identify certain instances. In this case, collaborating with AI can result in complementary team performance (CTP), leveraging the unique knowledge of both humans and AI, resulting in the human–AI team's task performance being better than the human or AI alone [136].

Inappropriate reliance can also occur in the presence of *imperfect XAI*, a term that we introduce to represent the phenomenon where an explanation reveals evidence that does not necessarily comply with the prediction. Papenmeier et al. [267] use the term 'explanation fidelity' while Kroeger et al. [187] use the term 'faithfulness' to measure how "truthful" an explanation is. We use imperfect XAI to align with existing terms, such as imperfect AI, in the CSCW and human-computer interaction (HCI) communities. Specifically, we define imperfect XAI as explanation techniques

that can potentially provide explanations that do not fit with the AI's predictions. We view explanation fidelity or faithfulness as a term that can be used under the umbrella term of imperfect XAI; we view explanation fidelity as referring to the continuum of explanation correctness, such as when explanations are partially correct. Imperfect XAI can exist regardless of whether the AI's advice is correct or not. AI explanations may oversimplify or improperly estimate complex models in order to make them more interpretable, deceiving and misleading the human decision-maker. As a result, non-experts may be more prone to under- or over-relying on AI advice in the presence of incorrect explanations. Within knowledge transfer scenarios, this could cause the non-expert to learn incorrect information about a given class. Furthermore, previous research shows that the language tone within natural language explanations can impact decision-makers [54]. However, the effect of language tone on humans' appropriate reliance when interacting with imperfect explanations is underexplored. Therefore, it is necessary to investigate how the communication style of explanations (*e.g.*, the language tone, the information provided, and their representations) impacts human–AI collaborations across levels of expertise [54, 179].

Collaborating with imperfect AI is not a new concept to the CSCW and HCI communities [136, 181, 216]. Despite numerous user studies over the years investigating human–AI collaborations and XAI, few have formally acknowledged the existence of imperfect XAI in their studies [136, 267]. By formally acknowledging the existence of imperfect XAI in human–AI collaboration, research has several new interesting dimensions to explore. Although numerous user studies seek to understand how humans align, perceive, and interact with different types of explanations in various human–AI collaboration scenarios (*e.g.*, [67, 177]), few studies explore the impact that incorrect or "noisy" explanations have on human–AI collaboration [136, 154, 267, 351]. Recent work has used technical approaches to mitigate "noisy" or incorrect natural language explanations [154], and Kroeger et al. [187] propose metrics to algorithmically evaluate the effectiveness of the generated post-hoc natural language explanations. However, to our knowledge, no studies investigate how the interaction between the correctness of explanations and the decision-maker's level of expertise impact appropriate reliance on AI, human–AI team performance, and the extent to which AI explanations deceive decision-makers.

With the growing use of machine learning models to explain other machine learning models in high-stakes decision-making scenarios, we argue that it is necessary to understand how humans interact with explanations that are incorrect, even when the AI's advice is correct. We also argue that it is important to understand the relationship that the level of expertise and the tone of explanations (*i.e.*, assertive, non-assertive, or neutral) have on a decision-maker's reliance on AI. Understanding these dimensions of human–AI collaboration will provide insight to XAI and CSCW researchers. With these topics under-explored in current literature, we present the following research questions:

RQ1: How does the correctness of explanations affect appropriate reliance on AI, and

to what extent do the decision-maker's level of expertise and the explanation's assertiveness moderate this effect?

RQ2: How is complementary team performance impacted by the correctness of explanations and the decision-maker's level of expertise?

RQ3: How do different types of explanations change the effect that the correctness of explanations has on appropriate reliance and complementary team performance?

RQ4: To what extent do incorrect and correct explanations deceive decision-makers with different levels of expertise?

To address our research questions, we employ an imperfect AI model for a bird species identification task [137]. We focus on bird species identification because the use of AI for wildlife conservation efforts among experts and non-experts is a rapidly growing field in research and practice [345]. Furthermore, it is less difficult to find people with varying levels of expertise in birding than in radiology who will have time to participate in our study.

Through a mixed-methods study, we answer our research questions by asking participants to classify bird images in two phases: without any advice from AI (phase 1) and then showing the AI's advice and explanation (phase 2). To answer **RQ** 1, we design a research model based on phenomena from relevant research in CSCW and conduct rigorous moderation analyses based on [134]. Our analyses leverage the appropriate reliance metrics defined by Schemmer et al. [301]. We design our study to be within-subjects for the correctness of the explanation and the assertiveness of explanations, allowing us to answer **RQ** 2 and between-subjects for the explanation modality, allowing us to answer **RQ** 3. Moreover, we calculate the magnitude of deception caused by incorrect explanations compared to correct explanations across both explanation modalities and levels of expertise to account for the impact of imperfect XAI on humans' decision-making behavior. This measurement gives us insight into **RQ** 4. Lastly, we conduct an inductive content analysis based on Gioia et al. [119] to assess the open-ended responses from participants to gain insight into designing for imperfect XAI in human–AI collaborations.

As a result of our study, we contribute the following to the CSCW community:

- *Research Model for Human–AI Collaboration with Imperfect XAI*: We propose a research model for the moderating roles of the decision-maker's level of expertise and the explanation's assertiveness on the effect of the correctness of explanations on appropriate reliance.
- *Novel Empirical Study*: To validate our proposed research model, we conduct the first empirical investigation that explores the moderation of assertiveness of explanations and the level of expertise on the impact that the correctness of explanations has on appropriate reliance. We do this on a human–AI collaboration scenario across two different modalities of explanations: natural language explanations and visual, example-based explanations. We also investigate the impact on complemen-

tary team performance. Our findings inform designers of human–AI collaboration systems on how to deploy imperfect XAI from a user-centric perspective.

- *Novel Metric for Impact of Imperfect XAI:* We contribute a novel metric to the human–AI decision-making field, accounting for the impact of incorrect explanations on humans' decision-making behavior when collaborating with AI. Specifically, we propose the Deception of Reliance (DoR) caused by imperfect XAI. With DoR, we investigate to what extent imperfect XAI deceives humans.
- *Qualitative Insights*: We provide insights on how decision-makers, regardless of expertise, prefer the tone of explanations to align with factors related to the AI's behavior and the impact on decision-makers. These insights can inform future works to conduct new evaluations.

4.2 Theoretical Development

The increasing use of explanations to reveal the rationale behind AI predictions has led to a rise in research examining the impact of explanations on decision-makers' behavior [82, 203, 299]. As imperfect AI is utilized more within high-stakes contexts, such as decision-making in the medical sector, research has focused on the impacts of potentially inaccurate AI advice on humans' decision-making [181, 202, 287]. However, there are very few works investigating how **imperfect XAI** impacts humans' appropriate reliance on AI advice [267].

Thus, in this work, we draw from the conceptualization of appropriate reliance that previous research has established [16, 299, 301, 303]. We specifically build on the conceptualization presented by Schemmer et al. [301] in Figure 4.1 by adding a new dimension to consider when investigating appropriate reliance in human–AI collaboration: The correctness of XAI advice. In our conceptualization in Figure 4.1, we simplified the correctness of XAI advice to be a binary case of correct or incorrect.

The introduction of this new dimension uncovers previously unexplored avenues within the realm of human–AI collaboration in CSCW, thereby offering a conceptual framework with a more profound comprehension of human decision-making behavior when collaborating with an AI system. As a result, researchers can calculate more specific metrics regarding human decisions after receiving the AI and XAI advice. For example, the ratio of under-reliance based on a correct prediction and incorrect explanation could be different than when based on a correct prediction and correct explanation. These types of scenarios should not be overlooked when investigating human–AI collaborations in work settings[1].

With this new dimension for analyzing human–AI collaborations, we investigate the effect that imperfect explanations have on humans' decision-making. Specifically, we investigate this relation in a sequential decision-making scenario. Based on the

[1]The Appendix of the published manuscript provides additional details of these newly introduced metrics.

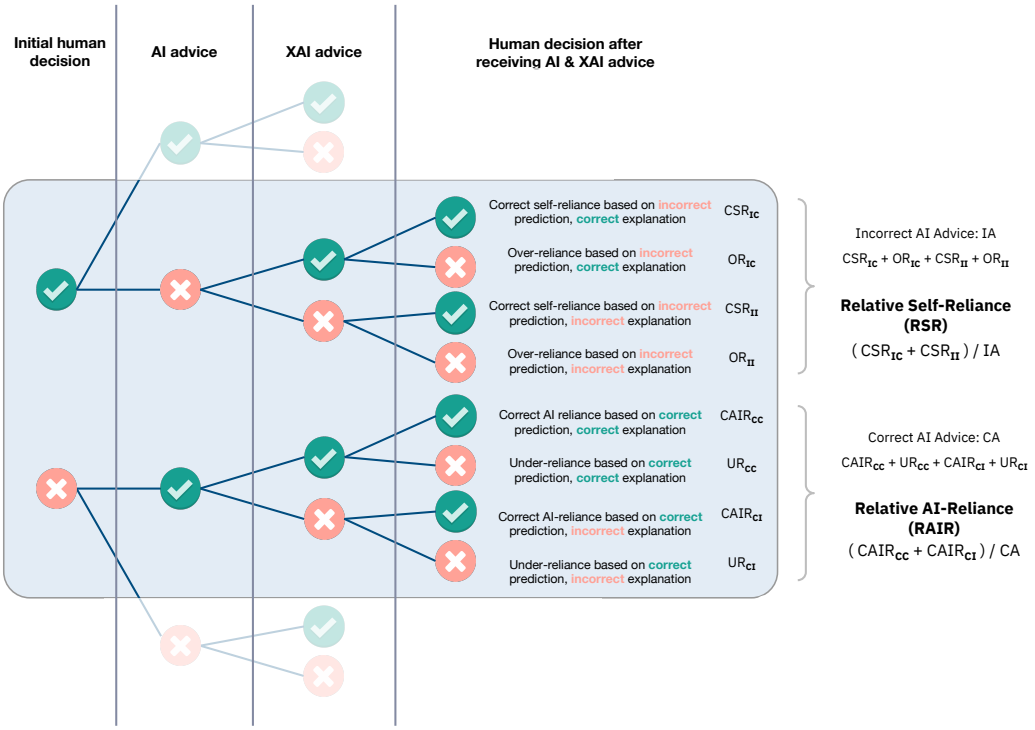


Fig. 4.1: Different paths that human decision-makers could follow based on receiving AI and XAI advice. This figure expands the model presented by Schemmer et al. [301] by adding an XAI advice dimension. Here, XAI advice is simplified into correct and incorrect explanations. Green check marks represent correct advice or decisions, while red x marks represent incorrect advice or decisions.

constructs of relative AI reliance (RAIR) and relative self-reliance (RSR), we account for the appropriateness of reliance [301][2]. RAIR comprises the cases in which the human corrects their initially incorrect decision by overriding it with the correct AI advice. On the other hand, RSR comprises all cases in which the human initially makes a correct decision, the AI system gives incorrect advice, and the human rightly dismisses this advice. Thus, we use appropriateness of reliance as the dependent variable in our research model (see Figure 4.2).

In the recent work of Schoeffer et al. [304], the authors investigate how explanations affect distributed fairness in AI-assisted decision-making. Their study shows that task-relevant explanations impact humans' reliance behavior in increasing stereotype-based errors. We apply these findings to our research model and assume that for the cases in which the AI provides correct advice, explanations will affect RAIR. Accordingly, we hypothesize:

Hypothesis 1: The correctness of explanations impacts humans' relative AI reliance in human–AI decision-making.

[2]Appropriateness of reliance is the quantitative measurement for appropriate reliance. These terms will be used interchangeably throughout the article.

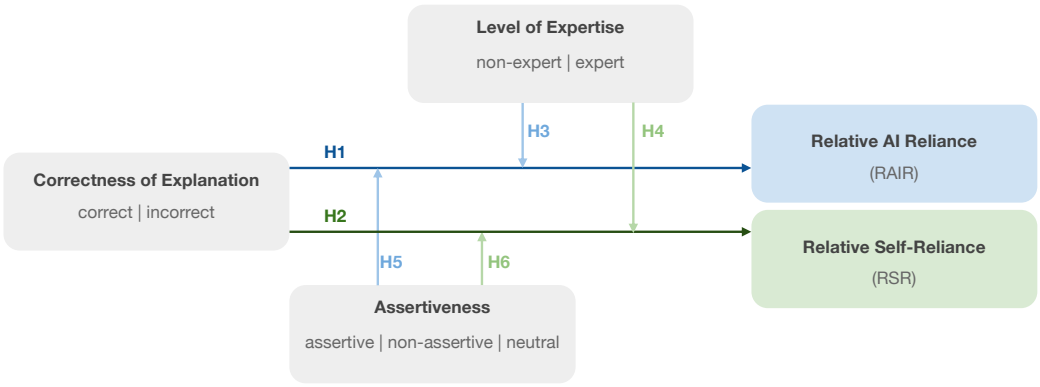


Fig. 4.2: Research model for collaborating with imperfect XAI systems. We analyze the moderation of the level of expertise and assertiveness on the effect of the correctness of explanation on RAIR and RSR.

Hypothesis 2: The correctness of explanations impacts humans' relative self-reliance in human–AI decision-making.

A human factor to consider in regard to the relationship between imperfect explanations and humans' appropriate reliance is the level of domain knowledge that humans possess. Previous work shows that humans' level of expertise can influence their decision-making [20, 54, 88]. For example, Dikmen and Burns [88] provide an imperfect AI and argue that higher domain knowledge leads to less trust in AI. With this impact of domain knowledge on human–AI decision-making, we intend to examine how the effect of imperfect explanations on appropriate reliance is influenced by humans' level of expertise. Humans with high domain-specific knowledge have demonstrated an enhanced ability to discriminate between erroneous explanations and accurate ones with greater rigor [205]. This discernment is facilitated by their extensive expertise, which empowers them to readily identify and discern false information [18]. Thus, in our study, we hypothesize:

Hypothesis 3: Humans' level of expertise moderates the effect of the correctness of explanations on RAIR.

Hypothesis 4: Humans' level of expertise moderates the effect of the correctness of explanations on RSR.

Humans' information processing is not only influenced by what they are provided but also by the way in which this information is provided. Previous research demonstrates that language style can impact humans' decision-making behavior [54, 147, 188]. For example, Huang et al. [147] show that the level of assertiveness in reviews affects humans' online review persuasion. Similarly, Kronrod et al. [188] investigate the effect of the level of assertiveness on the tone of language, and find that there is a relationship between the tone of language and its level of assertiveness. Recent research in HCI has examined how assertiveness in explanations affects humans' performance when interacting with AI [54], showing how the level of assertiveness does impact humans'

trust when collaborating with AI. With those findings and related research on the impact of assertiveness on humans' decision-making, it is reasonable to hypothesize that assertive explanations will impact humans' reliance on AI. Hence, we hypothesize:

Hypothesis 5: Explanations' level of assertiveness moderates the effect of the correctness of explanations on RAIR.

Hypothesis 6: Explanations' level of assertiveness moderates the effect of the correctness of explanations on RSR.

Our research model, shown in Figure 4.2, summarizes the hypotheses that we test in our mixed-methods study. Overall, with this research model, we test for the moderating effects of the level of expertise and level of assertiveness on the effect of the correctness of explanations on appropriate reliance.

4.3 Methodology

In this section, we describe the task of bird species identification, the experiment design, the recruitment process of participants, and the development of the explanations we use in the study. Finally, we end this section by outlining the data we select to use for the study and the metrics we use to analyze the results.

4.3.1 Task Domain: Bird Species Identification

Computer-supported cooperative work is becoming core to wildlife conservation efforts [38, 123, 345]. With mobile devices becoming increasingly powerful, non-experts and experts alike can use AI-powered applications like the Merlin Bird ID app [5] to identify bird species for monitoring biodiversity and learning about birds. The popularity of both birding and AI-based image classification techniques suggests that bird species identification would be a sensible domain to investigate our research techniques. Furthermore, this is a task for people with a wide range of expertise.

While identifying bird species from images may not be posed as a high-stakes task in our study, this task is imperative to conserving and managing species and biodiversity [11]. Furthermore, the task of fine-grained image classification, such as bird species identification, is comparable to higher-stakes tasks, such as identifying diseases from medical imagery [343]. For example, radiologists collaborating with an imperfect AI and imperfect XAI to diagnose diseases present in chest X-rays would go through a similar visual decision-making process as if they were trying to classify an image of a Bewick Wren in our study interface. Kayser et al. [166] and Hou et al. [146] propose imperfect natural language explanations for chest x-rays, similar to the explanation we implement in our study, which helps bridge our findings between bird species identification and higher-stakes tasks.

Previous studies that focused on human-centered XAI and human–AI collaboration also use the domain of bird species identification to better understand human–AI

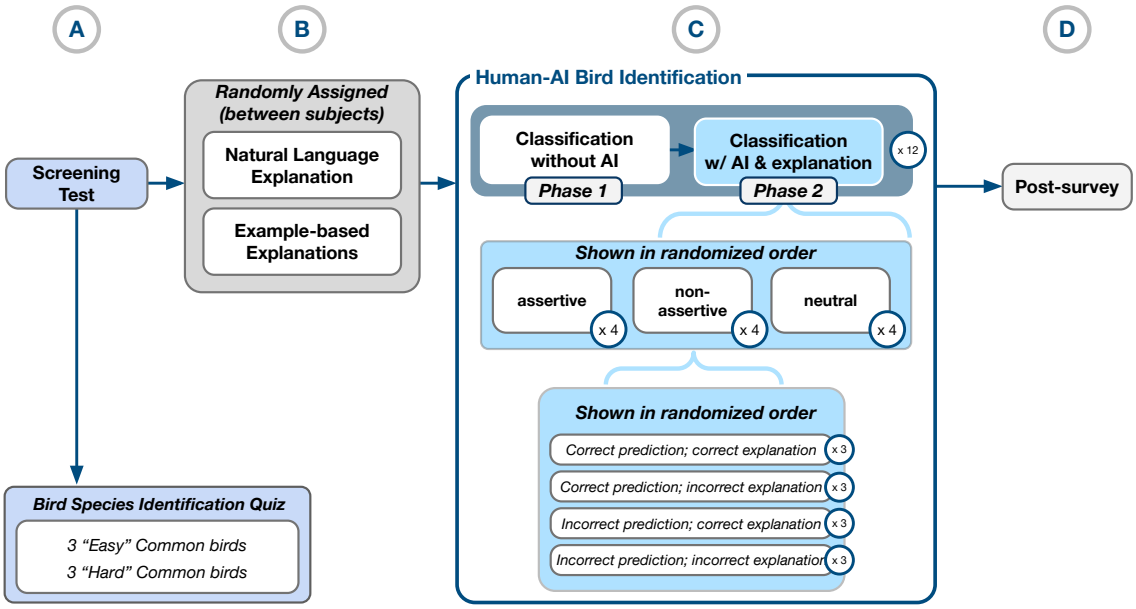


Fig. 4.3: We conduct a rigorous mixed-methods study leveraging a mixed design. Before participants start the task, they are shown a screening test (A). For the human–AI bird identification task, participants are assigned an explanation modality (B). During the task, participants are shown explanations with different levels of assertiveness and different scenarios of correctness (C). Lastly, participants complete a post-survey (D).

collaboration (*e.g.*, [49, 177, 253]). However, few previous works focus on human–AI collaboration for decision-making in the wildlife conservation domain overall. Yet, the field of AI for wildlife conservation is rapidly growing [345] and could benefit from research related to CSCW and HCI.

4.3.2 Study Design

To answer our research questions, we design a mixed between- and within-subjects study to examine various effects of explanations on appropriate reliance. Our study was carefully reviewed by two experienced birders: one experienced birder is a migration counter for a bird sanctuary, and the other experienced birder holds a graduate degree in environmental science, conducted research at a nature center, and worked at a nature conservancy. The procedure of the study follows the design outlined in Figure 4.3 and is divided into four different parts ($A - D$), which we explain in further detail below.

The study begins with part A in Figure 4.3: In a bird identification test, we assess participants' expertise in classifying six different species of birds[3]. We distinguish the six bird images based on their level of difficulty. This level of difficulty is derived

[3]Specific birds used for the bird identification test are reported in the Appendix of the published manuscript

from discussions with an experienced migration counter from a bird sanctuary. While previous work has collected participants' self-perception of expertise [177], this method is subjective, and participants may self-perceive their skills differently. Kazemitabaar et al. [168] measure participants "experience-level" through log data instead of subjective measures. Similarly, we try to avoid defining expertise subjectively. For the purpose of our analyses, we identify two different levels of expertise: *non-experts* and *experts*.

In the next section of the study (part B in Figure 4.3), participants are randomly assigned to one of two explanation types. Similar to previous studies [62, 286, 287, 335], the treatments differ in the explanation modality participants receive: *Natural language* explanations or *visual, example-based* explanations. We use natural language explanations because the AI model that we used for the study was specifically designed to generate natural language explanations based on fine-grained image classifications [137]. We choose to also look at example-based explanations as recent studies focus on this modality in human–AI collaboration [51, 67, 149, 177]. However, recent studies show that example-based explanations have potential benefits. Chen et al. [67] show that example-based explanations improve humans' performance, so much so that it leads to complementary team performance. With promising results from previous research and numerous clinical decision-support tools proposing to incorporate example-based explanations (*e.g.*, [17, 294]), we find it necessary to investigate the effect of example-based explanations on humans' appropriate reliance in the context of imperfect XAI.

The human–AI bird identification task (part C of Figure 4.3) consists of two phases. For each treatment condition, the participants are asked to initially identify the bird species from an image (phase 1 in Figure 4.4). After submitting an initial identification, they are shown the AI's prediction along with the explanation, and again, they have to submit an identification for the bird species in the image (phase 2 in Figure 4.4). The structure of phases one and two is corroborated with previous work [122]. Initially, participants must click on a button that shows "Show AI Explanation". Without revealing the explanation, the participant cannot proceed to the next question. This is one way for us to ensure that the participant acknowledges the presence of an explanation. Overall, participants do this process for twelve different random bird images.

As the AI that we are utilizing for identifying the bird species is not perfect [137], the predictions and explanations provided can be incorrect. In order to understand how this affects participants' appropriate reliance, we ensure that each participant is shown three samples of the following four categories in random order:

- **CC**: correct prediction and correct explanation
- **CI**: correct prediction and incorrect explanation
- **IC**: incorrect prediction and correct explanation
- **II**: incorrect prediction and incorrect explanation

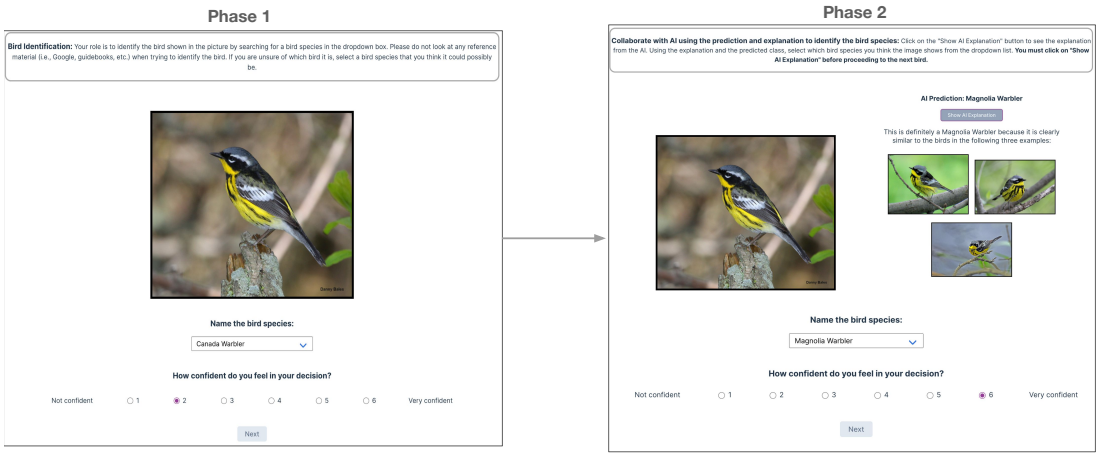


Fig. 4.4: Example of the two phases for a single bird image that a participant is shown in the study. This specifically shows a Magnolia Warbler (correct prediction, correct explanation), and this participant is assigned to the example-based explanations. For this bird, the participant is shown an assertive explanation.

Overall, participants are shown twelve different bird species. For each of the four categories we identified, we show three samples where each explanation is framed with a different level of assertiveness: *assertive*, *non-assertive*, or *neutral*[4]. As a result, each participant is shown one *assertive*, one *non-assertive*, and one *neutral* explanation for each category. We ensure that the order is randomized for each participant. Moreover, we also vary the samples shown, meaning that not every participant sees the same bird images. This is to ensure that our results are not dependent on the difficulty of the bird species.

After finishing the task, participants must fill out an additional questionnaire (part D of Figure 4.3). Here, we qualitatively assess participants' ability to properly rely on the AI based on the explanations that they were shown. Thus, we ask them: "*Under what circumstances would you prefer assertive (e.g., "definitely", "clearly") versus non-assertive (e.g., "might be", "appears to be") versus neutral explanations and why?*". Aside from this open-text question, we also ask participants about their occupations and the regions in North America that they are most familiar with in terms of bird species.

4.3.3 Data Selection

We create a dataset of bird images and explanations to show participants by manually curating bird images from the well-established CUB-200-2011 [358] dataset and explanations from the Generating Visual Explanations model [137]. The original dataset consists of 11,788 images of 200 different bird species and is split into 5,994

[4]Figure 4.6 provides examples of *assertive*, *non-assertive*, and *neutral* explanations.

training and 5,794 test images. Each bird species is represented with around 60 images of the respective bird class. When curating birds, we first filtered for bird families with several species in the CUB dataset. We specifically filtered out every bird class that is not a part of the Warblers, Wrens, Swallows, Sparrows, or Finches/Grosbeaks families. After applying this filter, we had $1,864$ out of $5,794$ images from the test set of CUB-200-2011. Of those $1,864$ images, $1,609$ were predicted correctly by the AI, and 255 images were predicted incorrectly by the model.

After filtering the bird species, multiple researchers on our team separately classified the natural language explanations and the visual, example-based explanations for a subset of the $1,864$ birds as incorrect or correct. Cases of doubt were discussed by a subset of the research team and excluded from consideration if an agreement was not met. In total, we identified ten examples for each category[5] and explanation type. In some cases, the example-based and natural language explanations for a single bird are used. As a result, the dataset represents 66 different images and 43 different bird species from the CUB-200-2011 dataset.

We define a correct natural language explanation to be when the explanation aligns with the description of the predicted bird class. We define an incorrect natural language explanation to misalign with all or part of the description of the predicted bird class. Thus, an incorrect natural language explanation contains a factual error. This type of incorrectness is present in different natural language techniques and is a focus of current research [221, 378]. We use descriptions from the Cornell Lab of Ornithology All About Birds Guide [2] to corroborate our classification for each explanation. Examples of incorrect and correct natural language explanations are provided in Figure 4.5.

For the example-based explanations, we define a correct explanation as the three most similar images belonging to the predicted class (as shown in phase 2 of Figure 4.4). We define an incorrect explanation to be most similar to at least one image that is not of the predicted class. This means that incorrect example-based explanations incorporate logical errors, as the examples shown are dissimilar from the predicted class. Moreover, such incorrect explanations can have an inconsistency, as the examples shown might differ in the classes shown. However, we only choose to show participants incorrect explanations that have at least two images that are not of the predicted class. For example, in Figure 4.5, the AI correctly predicts a Nashville Warbler; however, the three most similar examples are a Painted Bunting, a Yellow-Breasted Chat, and an American Redstart. In some cases where the advice is correct and the explanation is incorrect[6], the explanation may align with the ground truth class. For example, for a Tennessee Warbler, the AI predicts an Orange-crowned Warbler (incorrect advice since the wrong bird species is predicted), but the three most similar examples are all of Tennessee Warblers (incorrect explanation since the examples' bird species do not align with the prediction). It's possible that a model could be relying on spurious

[5]The four categories are identified in Section 4.3.2.

[6]The explanation does not align with the predicted class.

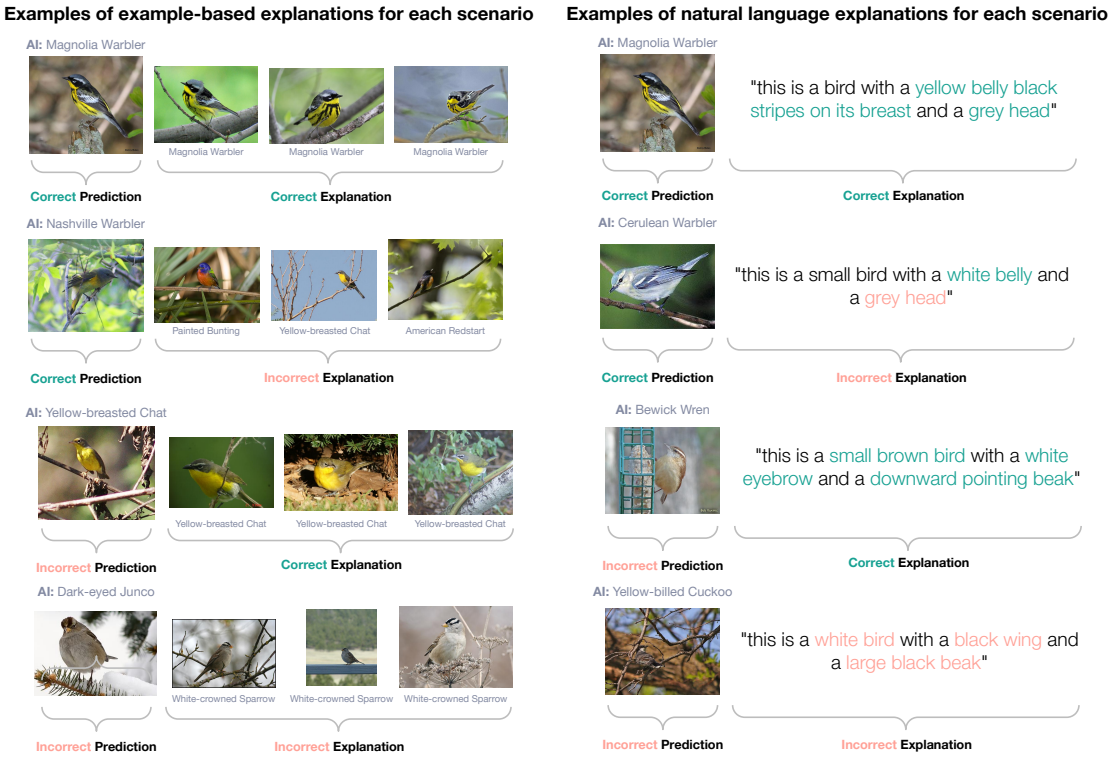


Fig. 4.5: Representative examples of the example-based and natural language explanations for each scenario: CC, CI, IC, and II. The class of the example-based images in the explanation is not shown to participants during the study. The red and green coloring on the natural language explanations was not shown during the study. This is only provided in the figure to guide the reader. The natural language explanation for the Cerulean Warbler is incorrect because this bird species does not have a grey head. The natural language explanation for the Yellow-billed Cuckoo is incorrect because this bird species is brown with a white belly, has a gold and black beak, and does not have black wings.

patterns to make classifications [273]. Since we are dealing with an imperfect AI, we do not choose to exclude such cases from our dataset.

4.3.4 Explanation Modalities

Natural Language Explanations. The natural language explanations were generated by the model proposed by Hendricks et al. [137]. We followed the PyTorch implementation [12] of Hendricks et al. [137] model to obtain the natural language explanations since the original model from Hendricks et al. [137] was unavailable. After running the test images through the model, a natural language explanation is generated for each classification. For example, the natural language explanation for the Magnolia Warbler in Figure 4.5 is: "`this is a bird with a yellow belly black stripes on its breast and a grey head`".

Example-Based Explanations. Previous work creates example-based explanations,

specifically normative explanations, by calculating the Euclidean distance between the given image and the images in the dataset [51]. Another study generates the example-based explanation by calculating the L_2 distance of the embedded features [252]. We generated the example-based explanations by following methods used in previous works [17, 343]. As done by Tschandl et al. [343] and Barata and Santiago [17], we calculate the cosine similarity between the extracted feature vector of the given image and the rest of the extracted feature vectors of the images in the training set. Unlike Barata and Santiago [17], we choose not to take the example's ground truth class into consideration. The extracted features from the images were provided by Hendricks et al. [137]. Because the model is not perfect, the example-based explanations are also not perfect. For example, even though the model correctly predicted an image of a Nashville Warbler in Figure 4.5, the three most similar images are of three different birds. For this study, we consider an example-based explanation to be incorrect if two of the three examples are of a different class than the predicted class. Inspired by Ford and Keane [111], we choose to show participants the three most similar examples.

Assertiveness of Explanations. Following previous studies [54, 264, 374], we define assertive explanations to include words and adjectives such as "definitely" and "clearly". We define the non-assertive explanations to include words and adjectives such as "might be" and "appears to be". For the neutral condition, we omit the adjectives to maintain the same structure of the information being presented. For the natural language explanations, we append the assertiveness to the beginning of the explanation generated by the model to read like a sentence. For the non-assertive and assertive conditions, we removed the text "this is a" from the generated explanation in order to incorporate it into the sentence structure we designed. The three versions of assertiveness for both explanation modalities are shown in Figure 4.6. To our knowledge, there is no literature to rationalize how to appropriately present assertiveness visually for

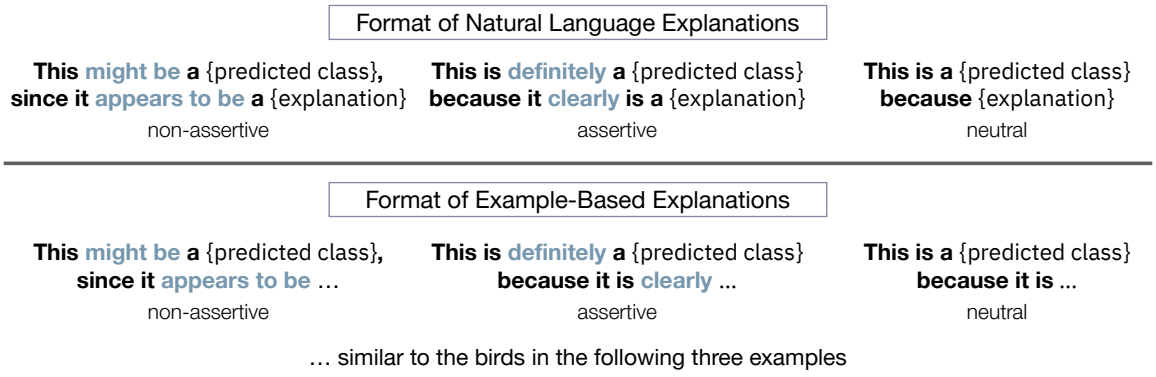


Fig. 4.6: The three different language tones that an explanation could have in terms of assertiveness for the natural language and example-based explanation modality. Non-assertive explanations included the words "might be" and "appears to be". Assertive explanations included the words "definitely" and "clearly". Neutral explanations did not include any additional adjectives.

example-based explanations, so we opt to use natural language in combination with the example-based explanations.

4.3.5 Recruitment

We recruit the participants through several communication channels that are related to the environment and conservation, such as the AI for Conservation Slack, Birding International Discord, Climate Change AI community forum, WildLabs.net community forum, and Audubon Society mailing lists. Additionally, we use Prolific as previous research has indicated that this platform is a reliable source of research data [265, 270]. We apply a custom filter on Prolific to target individuals who currently work in a field related to nature, science, the environment, or animals. Participants receive compensation that is above minimum wage.

Overall, we try to limit recruitment to only address people with prior knowledge of birding to minimize the prevalence of novices randomly guessing bird species identification. After excluding participants who provide incomplete and fake responses (*i.e.*, lorem ipsum response to our survey question), we have 136 people complete our study. In order to determine if a participant is familiar with birding, participants take a bird identification test (phase A in Figure 4.3). The participant's score on the bird identification test is used to determine whether they are considered a non-expert or an expert. Details related to clustering participants based on their test scores are provided in Section 4.4.

4.3.6 Quantitative and Qualitative Metrics

Quantitative Metrics We quantitatively calculate appropriate reliance across the four dimensions defined by Schemmer et al. [301]: correct AI reliance, correct self-reliance, under-reliance, and over-reliance. Accordingly, correct AI reliance measures the number of correct decisions when the human's initial decision is incorrect, and the human is rightly taking over the correct AI advice. Correct self-reliance is when the human initially makes the correct decision and does not overwrite their decision with the incorrect AI advice. On the other hand, under-reliance reflects the case in which the human initially makes an incorrect decision and does not adhere to the correct AI advice. On the other hand, over-reliance represents the scenario in which the human makes an initial correct decision but overrides their own decision with incorrect AI advice. Following the appropriate reliance metrics defined by Schemmer et al. [301], we calculate RSR and RAIR to account for the appropriateness of reliance.

With the new dimension for XAI advice, we can separately measure RAIR and RSR for correct and incorrect explanations and derive their impact on appropriate reliance. In order to measure this impact, we look at the Deception of Reliance (DoR) caused by imperfect XAI. For RAIR, we can apply the following:

$$DoR_{RAIR} = RAIR_C - RAIR_I. \tag{4.1}$$

In this equation, the subscript I represents incorrect explanations, whereas the subscript C represents correct explanations. We can compute the same for RSR:

$$DoR_{RSR} = RSR_C - RSR_I. \tag{4.2}$$

In order to measure the overall deception impact of explanations on humans' decision-making behavior, we compute the deception on appropriate reliance by calculating the Gaussian distance in the RAIR-RSR space between incorrect and correct explanations:

$$DoR(RAIR, RSR) = \sqrt{(RAIR_C - RAIR_I)^2 + (RSR_C - RSR_I)^2}. \tag{4.3}$$

According to Schemmer et al. [301]'s conceptualization of Appropriateness of Reliance, this results in the following:

$$DoR_{AoR} = AoR_C(RAIR, RSR) - AoR_I(RAIR, RSR). \tag{4.4}$$

This difference represents the deception between the correct and incorrect explanations. If the deception is a positive value, then incorrect explanations are more deceptive; if the difference is a negative value, then correct explanations are more deceptive.

Lastly, as defined by previous work (*e.g.*, [16, 122, 136]), we can calculate the human–AI team performance to determine if CTP exists. Following the constructs defined in those previous works, we determine if CTP exists by calculating the participants' performance in identifying the bird species **before** and **after** they see the AI advice and compare this to the performance of the model on the twelve bird images shown to the participant. We utilize accuracy as a performance metric. Since every participant is shown six birds that the AI correctly classifies and six that the AI incorrectly classifies, the model performance is 50%.

Qualitative Metrics We also conduct an inductive content analysis of the open-ended responses to better understand the participants' preferences regarding assertiveness. As a reminder, in the end-survey, we ask participants specifically "*Under what circumstances would you prefer assertive (e.g., "definitely", "clearly") versus non-assertive (e.g., "might be", "appears to be") versus neutral explanations and why?*". To analyze these responses, we follow the established procedure of Gioia et al. [119] and screen the answers in three coding workshops. In the first workshop, the first and second authors of this article initially screened the answers and applied open coding [143] to extract and aggregate core constructs of participants' answers. In this procedure, both authors discuss their findings and align their understanding of relevant constructs. Through a second coding workshop, we apply axial coding to derive subcategories of these constructs and align those with the data. In a final workshop, we distill those emerging themes and derive aggregated dimensions [376].

4.4 Results

To answer our research questions and confirm or reject our hypotheses, we conduct rigorous statistical and qualitative analyses. We measure appropriate reliance based on metrics defined in previous work: Relative AI reliance (RAIR) and relative self-reliance (RSR) [301]. By doing so, we answer **RQ** 1 and **RQ** 2 in Section 4.4.2. We also calculate the participant's task accuracy before and after receiving AI and XAI advice to answer **RQ** 3 in Section 4.4.3. This way, we can determine whether complementary team performance exists [136]. Additionally, based on the new metric, Deception of Reliance (DoR), we measure to what extent explanations deceive humans, answering **RQ** 4 in Section 4.4.4. Lastly, we qualitatively analyze open-ended responses through rigorous inductive content analysis in Section 4.4.5. For all of our research questions, we look at two different types of explanations: natural language explanations that are focused on specific features present in the image and visual, example-based explanations showing the top three most similar example images from the training set. While our analyses look at both modalities, we do not intend to compare them directly. Therefore, we do not conclude one modality is better or worse than the other.

4.4.1 Participant Statistics

On average, the study takes 24 minutes to complete. In order to distinguish experts from non-experts, we perform K-means clustering ($k = 2$) based on a principal component analysis with two components for four features from the bird species identification test (part A of Figure 4.3). These four features represent participants' scores in correctly identifying the family and species of the easy and the difficult bird images. By clustering the 136 participants into the expert and non-expert groups, we end up with 83 experts and 53 non-experts. With this clustering, the average bird identification test score (summing up all four scores in the identification test) for non-experts is $38.99\%(STD = 11.42\%)$ while the average test score for experts is $83.84\%(STD = 12.30\%)$[7]. Of the 83 experts, 42 see example-based explanations, and 41 see natural language explanations. Of the 53 non-experts, 25 see example-based explanations, and 28 see natural language explanations. In terms of the fields that the 136 participants represent, 45 participants have an occupation primarily related to biology, conservation, and/or the environment. 26 have an occupation primarily related to engineering and/or technology; 30 are either researchers, students, or affiliated with education in some other way; 24 have occupations in miscellaneous industries; and 11 are retired.

[7]Participants' performance on the bird identification test is shown in the Appendix of the full paper.

Coding of assertiveness		
assertiveness	Z1	Z2
neutral	0	0
non-assertive	1	0
assertive	0	1

	RAIR				RSR			
	Natural Language		Example-Based		Natural Language		Example-Based	
	coeff	p	coeff	p	coeff	p	coeff	p
const	1.26	.00	.43	.17	-17.16	.98	-3.60	.00
corr	.57	.29	1.02	.04	13.25	.98	-4.36	.74
exp	2.12	.00	-1.25	.00	15.89	.98	3.25	.00
Z1	-.46	.24	.26	.47	.14	.26	-.09	.83
Z2	.00	1.00	.00	1.00	-.31	.58	.09	.83
exp x corr	-1.00	**.05**	-1.04	**.03**	-13.33	.98	-.73	.57
Z1 x corr	.46	.44	-.03	.95	-.28	.71	-.45	.55
Z2 x corr	.19	.74	-.16	.77	.90	.23	-.43	.55

[1] *corr — correctness*; *exp — level of expertise*

Table 4.1: Moderation analyses of the correctness of natural language and example-based explanations on RAIR and RSR with the level of expertise and assertiveness as moderators. The coding of assertiveness used for the moderation analyses is provided.

4.4.2 Moderating Effects in Imperfect XAI Research Model

In order to test whether humans' level of expertise and the explanations' assertiveness moderate the relation of the correctness of explanations on humans' appropriate reliance, we conduct several moderation analyses utilizing the process macro model of Hayes [134]. An overview of the regression analyses is presented in Table 4.1.

4.4.2.1 Participants' level of expertise moderates the effect of the correctness of explanations on RAIR for natural-language explanations. As theoretically developed in Section 4.2, we model the correctness of explanations as an independent variable. Accordingly, we model RAIR as the dependent variable. To account for the moderation effect of the level of expertise and assertiveness, we examine each variable as a moderator and report the interaction effects with the correctness of explanations. The results of this moderation analysis are shown in Table 4.1.

The moderation analysis shows that the interaction of the level of expertise with the correctness of explanations is significant (coeff $= -1.00$, p-value $= .05$). We observe a negative coefficient. Accordingly, the moderation effect on the relation of correctness on RAIR is higher for non-experts than for experts. In other words, non-experts change their initially incorrect decision to align with the correct AI advice more often

than experts do when the natural language explanation is correct. However, there is no significant effect in the interaction of assertiveness and the correctness of explanations. Thus, we conduct a regression analysis with the moderators as independent variables to evaluate for a direct effect of assertiveness as recommended by Hayes [134] and Warner [368]. The results of the regression analysis show that there is no direct effect between assertiveness and RAIR (coeff = .04, p-value = .77). Thus, we confirm hypotheses 1 and 3 and reject hypothesis 5 for natural language explanations.

4.4.2.2 Participants' level of expertise moderates the effect of the correctness of explanations on RAIR for example-based explanations. Next, we present the moderation analysis of example-based explanations on RAIR. We set up the analysis for example-based explanations the same as the analyses for natural language explanations (see 4.4.2.1). As seen in Table 4.1, there is a significant moderation effect of the level of expertise (coeff = −1.04, p-value = .03). The negative coefficient signals that this moderation is higher for non-experts than for experts. The correctness of the example-based explanations has a positive coefficient (coeff = 1.02, p-value = .04), and thus, correct explanations have a positive impact on RAIR. Thus, if participants are provided with a correct explanation, they more often correctly follow the AI advice. Overall, correct explanations result in humans changing their initially incorrect decisions to align with the correct AI advice more often, and this is especially prevalent among non-experts.

Furthermore, the analysis reveals that assertiveness does not moderate the effect between correctness and RAIR. According to Hayes [134] and Warner [368], we drop the interaction term and conduct a regression analysis with assertiveness set as the independent variable. The result shows that there is no significant direct effect of assertiveness on RAIR (coeff = −.04, p-value = .79). Hence, we confirm hypotheses 1 and 3 and reject hypothesis 5 for example-based explanations.

4.4.2.3 Participant's level of expertise has a direct effect on RSR for natural language explanations. In addition to analyzing whether the level of expertise and assertiveness moderate the effect of explanations' correctness on RAIR, we conduct the same analyses for the effect of explanations' correctness on RSR. For RSR, we look at all cases in which the AI prediction is giving incorrect advice (*i.e.*, the prediction is wrong) and the initial human decision is correct [301].

The moderation analysis in Table 4.1 for the natural language explanation shows that there is no significant effect of correctness on RSR moderated by level of expertise (coeff = −13.33, p-value = .98) and assertiveness (Z1 x corr.: coeff = −.28, p-value = .71; Z2 x corr.: coeff = .90, p-value = .23).

Thus, we perform a regression analysis with the level of expertise and assertiveness as independent variables and drop the interaction terms. We observe that there is no significant effect of assertiveness on RSR (coeff = .10, p-value = .58). However, the level of expertise (coeff = 3.18, p-value = .00) has a significant effect on RSR. With a positive coefficient, this means that experts dismiss incorrect AI advice more than

non-experts when shown natural language explanations. This can also be seen in 4.8, which tells us that experts have a higher RSR than non-experts. Therefore, we reject hypothesis 2 and hypothesis 4, as the level of expertise does not have a moderating role but has a direct effect on RSR for natural language explanations. On top of that, we reject hypothesis 6.

4.4.2.4 The correctness of explanations and participants' level of expertise have a direct effect on RSR for example-based explanations. Lastly, we report the results of the moderation analysis for example-based explanations on RSR. The analysis is set up the same as it is in Section 4.4.2.3 but for the example-based explanations.

We can see in Table 4.1 that the level of expertise does not significantly moderate the effect of correctness on RSR (coeff = -.73, p-value = .57). Additionally, there is no significant moderation of assertiveness (Z1 x corr.: coeff = $-.45$, p-value = .55; Z2 x corr.: coeff = $-.43$, p-value = .55). Thus, we conduct a regression analysis without the interaction terms. The results of this analysis show no direct effect of assertiveness on RSR (coeff = $-.03$, p-value = .86). However, there is a significant direct effect of explanations' correctness on RSR (coeff = -1.40, p-value = .00) and a direct effect of level of expertise on RSR (coeff = 3.05, p-value = 0.00). This means that experts more often correctly override the wrong AI advice and stick to their correct initial decision compared to non-experts for example-based explanations. Additionally, when the explanations are correct, participants more often correctly override wrong AI advice and stick to their correct initial decision. Hence, experts have a higher RSR than non-experts, which can also be seen in 4.8. Moreover, as incorrect explanations have a higher impact on RSR, experts are able to identify false AI advice for incorrect explanations better. This also shows that experts are able to identify incorrect AI advice to a greater extent than non-experts; experts, in this case, rely more heavily on their own judgment.

Thus, we confirm hypothesis 2 and reject hypothesis 4 as the level of expertise does not have a moderating role but has a direct effect on RSR for example-based explanations. On top of that, we reject hypothesis 6.

Overall, the moderation analyses reveal that the level of expertise moderates the effect of the correctness of explanations on RAIR for both explanation modalities. Additionally, the analyses show that the level of expertise has a direct effect on RSR for both explanation modalities, and the correctness of explanations has a direct effect on RSR for example-based explanations.

4.4.3 Human–AI Team Performance

Hemmer et al. [136] argue that interpretability is a key component of human–AI complementarity. Several previous user studies have failed to show that incorporating XAI into AI systems can lead to CTP [110]. However, with a new dimension of XAI advice in Figure 4.1, we can contribute to the current literature by investigating how the correctness of explanations affects CTP. By calculating the participants'

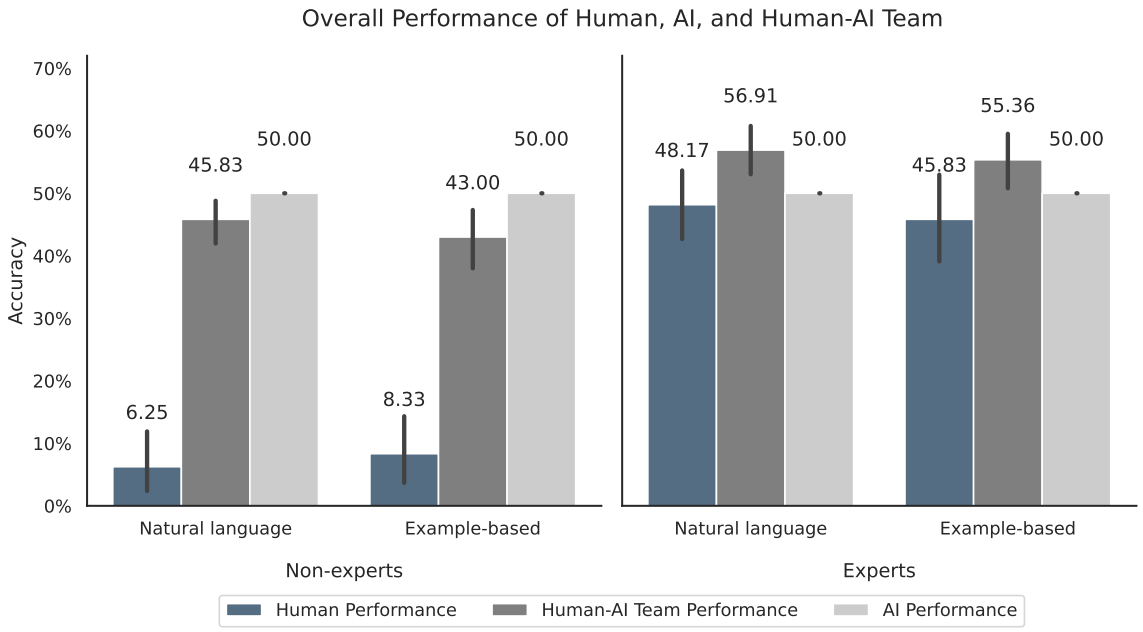


Fig. 4.7: The average overall performance of the human, AI, and human–AI teams for identifying 12 birds. The bar chart on the left shows the performance of the non-experts, while the bar chart on the right shows the performance of the experts.

performance before and after seeing the AI and XAI advice, we can determine whether CTP exists in the presence of imperfect XAI. As the analyses in Section 4.4.2 reveal, the level of expertise impacts appropriate reliance in terms of RSR and RAIR. Thus, in comparing the human–AI team performance, we distinguish by participants' level of expertise. We use accuracy as the performance metric. Figure 4.7 presents the performance of AI and humans for each treatment.

The AI's performance is always 50% because the study was designed to show participants six birds that the model correctly classified and six that the model incorrectly classified. In Figure 4.7, we see that when experts are paired with the AI, their performance improves by 8.74% for the natural language modality and 9.53% for the example-based modality. When experts are paired with AI, they perform 6.91% better than the AI alone for natural language explanations and 5.36% for example-based explanations.

While experts reach CTP, we do not see this for non-experts. However, we do see that the non-experts greatly improve their performance and nearly match the AI's performance when paired with the AI. Specifically, non-expert participants who see the natural language explanations improve their performance by 39.58% (task accuracy of 45.83%), while non-expert participants who see the example-based explanations improve their performance by 34.67% (task accuracy of 43.00%) when paired with the AI.

We can separate Figure 4.7 into correct and incorrect explanations. When we only

consider cases with correct explanations (additional figures provided in the Appendix of the published paper), the non-experts' task accuracy is approximately the same as the AI alone: 48.81% for natural language explanations and 49.33% for example-based explanations. Experts reach CTP in both modalities. When only considering incorrect explanations, we still see complementary team performance for the experts. However, the non-experts' task accuracy suffers more when shown incorrect explanations. Non-experts' task accuracy for natural language explanations is 42.86% and 36.67% for example-based explanations.

Additionally, we calculate two-sample t-tests to assess whether the trends in Figure 4.7 are significant. The team performance of experts and AI is significantly higher than the team performance of non-experts and AI in both explanation modalities (natural language: p-value = 0.00, example-based: p-value = 0.00). Furthermore, we also compare the performance for correct and incorrect explanations. Here, we see the same results: experts achieve a significantly higher team performance than non-experts (correct explanations — natural language: p-value = 0.00, example-based: p-value = 0.01; incorrect explanations — natural language: p-value = 0.00, example-based: p-value = 0.00).

4.4.4 Deception caused by Imperfect XAI

In Figure 4.8, we compare RAIR to RSR for both levels of expertise and the correctness of explanations. We show this comparison for example-based explanations (the graph on the left side of Figure 4.8) and natural language explanations (the graph on the right side of Figure 4.8). By measuring RAIR and RSR for incorrect and correct explanations separately, we can calculate the deception caused by imperfect XAI (refer to Equation 4.3 on p. 61). We do not visualize assertiveness since we do not see any significant direct or moderating effects.

The figure shows that experts have a higher RSR than non-experts for both incorrect and correct explanations across both explanation modalities, validating that experts rely more on their own initial decisions when AI advice is given. The most striking result that emerges from the data is that for example-based explanations, we observe that experts have a significantly higher RSR for incorrect explanations (RSR = 0.57) than correct explanations (RSR = 0.29), resulting in a negative DoR_{RSR} of -0.28 (p-value = 0.00). As a result, experts are falsely relying on the AI advice when provided with correct example-based explanations[8]. This means that experts are prone to being misled by correct explanations when the AI advice is incorrect. However, we do not see this trend for natural language explanations. Here, there is a positive DoR_{RSR} of

[8]Note that correct example-based explanations are consistent in showing three images of the predicted class. Incorrect example-based explanations represent three images that do not correspond to the predicted class of the AI. Moreover, the examples shown are not consistent with the bird species displayed in 90% of the *correct advice, incorrect explanation* cases and in 40% of the *incorrect advice, incorrect explanation* cases in our study.

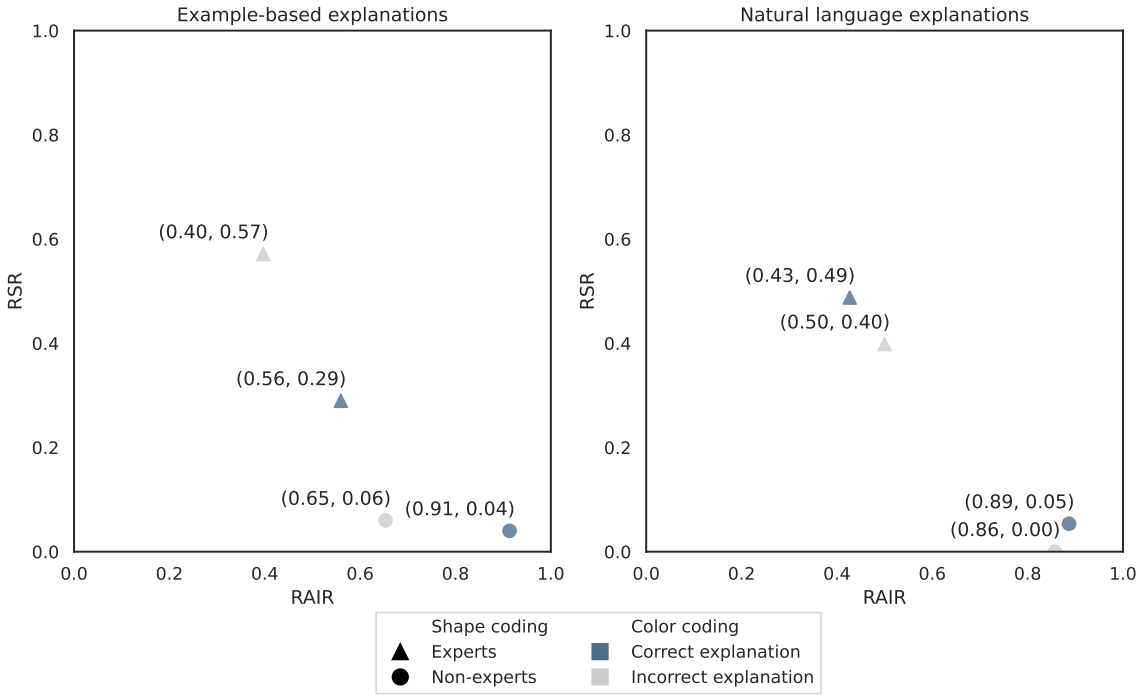


Fig. 4.8: Average observed RAIR (correct AI advice) and RSR (incorrect AI advice) for example-based explanations (on the left side) and natural language explanations (on the right side). We show the average RAIR and RSR for both levels of expertise, as well as correct and incorrect explanations.

0.09, which is not significant (p-value = 0.41). For example-based explanations, the DoR_{RAIR} is positive, meaning that experts rightly follow correct AI advice more often when provided with correct explanations than with incorrect explanations. The data shows a weak, significant positive deception of reliance for example-based explanations ($DoR_{RAIR} = 0.16$, p-value = 0.10). Similarly to the RSR cases, for the RAIR cases, the experts are provided with three consistent examples for correct explanations that represent the AI's correctly predicted bird species. The incorrectly provided explanations represent three images that can be inconsistent in the bird species. Thus, experts are deceived by such incorrect explanations even though the AI advice is correct.

Non-experts have, in both modalities, a similar DoR_{RSR}, indicating no significant difference in their RSR between correct and incorrect explanations. However, non-experts follow the correct AI advice for correct example-based explanations more often than for incorrect example-based explanations (significant with p-value = 0.01). For the latter, the three examples can show inconsistent bird specie(s) that are different from the ground truth of the shown image. Thus, the DoR_{RAIR} for non-experts is at 0.26. Interestingly, for natural language explanations, the incorrect explanations are not misleading as much ($DoR_{RAIR} = 0.03$, not significant, with a p-value = 0.68). This means that non-experts are not misled by incorrect explanations in natural

language as much as by visual, example-based explanations. In general, non-experts have a higher RAIR than experts.

Overall, participants have a higher $DoR(RAIR, RSR)$ for example-based explanations (experts: $DoR(RAIR, RSR) = 0.32$; non-experts: $DoR(RAIR, RSR) = 0.26$) than for natural language explanations(experts: $DoR(RAIR, RSR) = 0.11$; non-experts: $DoR(RAIR, RSR) = 0.06$). This means that especially the correctness of example-based explanations has an impact on humans' decision-making behavior.

4.4.5 Designing for Imperfect XAI

At the end of the bird identification task, we ask participants: "*Under what circumstances would you prefer assertive (e.g., "definitely", "clearly") versus non-assertive (e.g., "might be", "appears to be") versus neutral explanations and why?*". With imperfect XAI existing in human–AI collaborations, it is necessary not only to understand quantitatively how it affects decision-makers but also qualitatively. Even though we do not observe assertiveness having a direct or moderating effect on appropriate reliance, it is still valuable to analyze participants' preferences regarding the tone of explanations.

Through inductive content analysis of participants' responses to this question, we derive two dimensions that researchers and designers should consider when developing and evaluating imperfect XAI in human–AI collaborations: *AI Behaviors* and the *Impact on Human–AI Teams*. Each dimension is made up of four themes that are derived from concepts that emerge in the responses, shown in Figure 4.9. We highlight those themes in bold. 25% (34 participants) of the responses either do not provide reasoning for their opinion or do not answer the question such that it could be grouped into one of the eight themes we identify. We provide quotes from participants for each theme to structure the aggregated dimensions and shape our insights on designing for imperfect XAI.

4.4.5.1 Aggregated Dimension: AI Behaviors. 54 out of the 136 participants answered the survey question with comments relating to the first aggregated dimension: AI Behavior. Participants rationalize that the AI's behavior determines when they prefer assertive, non-assertive, and neutral explanations. Within the AI Behavior dimension, four themes emerge from the participants' comments, such as the model's overall performance, whether the model's prediction is correct or not, the model's confidence in individual predictions, and the correctness/quality of the explanation for a given prediction.

While only 6 out of the 136 participants make comments about the **model's performance**, it still provides interesting insight that should be considered. Instead of looking at the individual prediction level, these participants focus on the global performance of the model. One participant states that if developers find their model "*... to be 90% accurate in your testing, use more definite language, but if it's not there yet, consider making the tone more neutral and put more responsibility with the end*

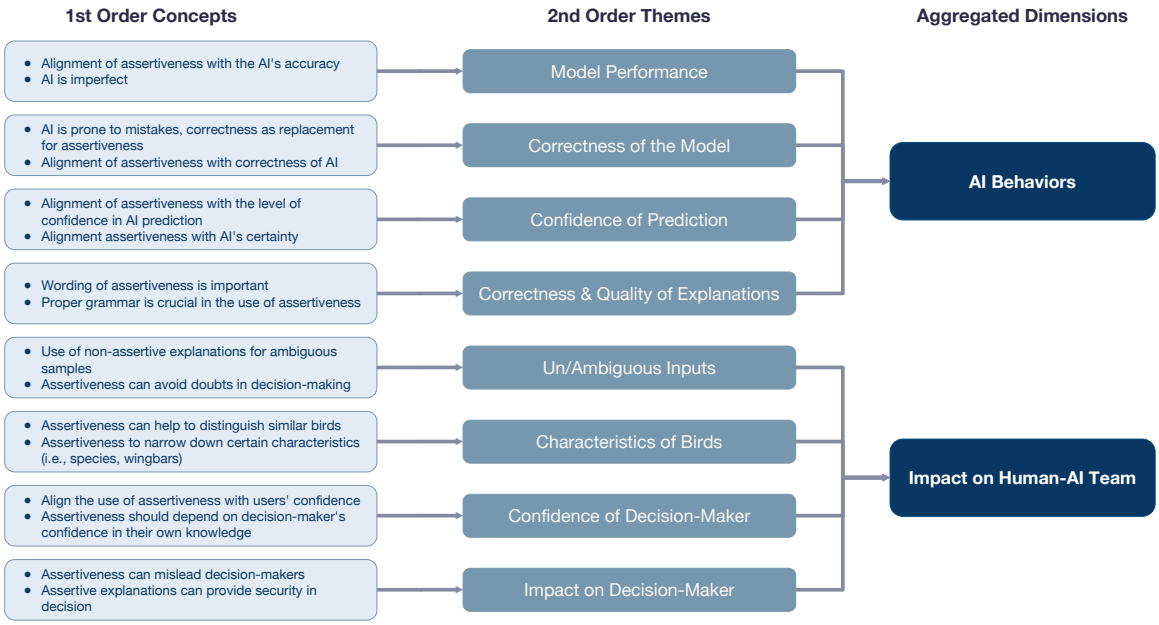


Fig. 4.9: Findings from the qualitative analysis of the survey question: "Under what circumstances would you prefer assertive (*e.g.*, "definitely", "clearly") versus non-assertive (*e.g.*, "might be", "appears to be") versus neutral explanations and why?". Factors that should be taken into consideration when designing explanations for imperfect XAI systems.

user to interpret the field marks ...".

Looking at the individual prediction level, 8 out of 136 participants commented on the **correctness of the AI's prediction**. For example, one participant says that "*... a non-assertive response would be preferred since the AI selections were incorrect ...*" while another participant says, "*I would prefer the assertive language to be accompanied by correct identifications*".

Considering individual predictions on a more granular level, many participants (31 out of 136) make comments related to **the confidence of the AI's prediction**. These participants comment on how this factor could be used to determine the tone of the explanation. Specifically, participants, "*... would prefer the level of assertiveness to depend on the level of confidence of the answer given by the AI*". One participant expands upon that sentiment by specifying when non-assertive versus assertive tones should be used: "*I would prefer assertive sentences when the probability of the AI model is very high, while I would prefer non-assertive when the probability is very close to other classes of the model*".

9 out of 136 participants make comments related to the **correctness and quality of the AI explanations**. One participant who sees example-based explanations comments on how some of the examples are incorrect and do not show the correct species. They use this specific situation to rationalize when they would prefer the tone of explanations to be assertive versus non-assertive: "*I would prefer assertive explanations if the*

'similar' photos were actually of the correct species and if the explanation confirmed this. Otherwise, non-assertive explanations are more helpful". Another participant who makes a comment related to this theme agrees that assertive language should be used, "*when all of the reference pictures match up and there are no other similar-looking species*".

One participant who sees the natural language explanations comments on the quality and detail of the explanation being a factor to use when determining the tone of the explanation, "*If the bird description [natural language explanation] is very generic, i.e., brown wings, gray body, or yellow beak (traits that correspond to many birds), I'd rather the AI appear more cautious in its judgment and use non-assertive explanations. However, if the bird has some standout characteristic that the AI correctly identifies [through the natural language explanation], i.e., bright yellow body or red-tipped wings, etc., then assertive explanations seem more convincing*".

4.4.5.2 Aggregated Dimension: Impact on Human–AI Team. 48 out of the 136 participants answered this survey question with comments related to the human–AI team, such as the confidence and knowledge of the decision-maker, impact on the decision-maker, unambiguous input data, and characteristics of the input data.

10 of 136 prefer the level of assertiveness to align with their **own confidence level** and knowledge of the domain. For example, one participant says, "*I would prefer more assertive explanations when I don't feel very sure about my choice*". Another participant adds that they would prefer assertive explanations if they "*... didn't know anything about the topic in question ...*". However, one participant says, "*I would prefer neutral explanations if I'm unsure of the species and assertive if I'm confident in my identification*".

15 out of the 136 make comments related to the **impacts on the decision-maker**. Some comments consist of concerns related to being misled and over- or under-relying on the AI, while other comments motivate the benefits of having assertive explanations. For example, one participant states that "*Assertive words create more security while changing your opinion or trying to gain knowledge*". Another participant who shares the same sentiment said, "*I would prefer assertive explanations because it would make me feel more secure about the answer*". However, some participants do not share the same sentiment about assertive explanations and pointed out the potential consequences of them: "*Assertive AI explanations were given for incorrect identifications, which would mislead users*". Given that potential to be misled, another participant rationalizes they "*... would prefer non-assertive explanations because I [they] do not fully trust AI with bird ID just yet*".

17 out of the 136 participants rationalize that the assertiveness of explanations should be based on how **ambiguous the input** is for a given prediction. For example, one participant brings up the quality of the input image and the difficulty of the bird ID as a way to determine whether explanations should be assertive or not: "*When it comes to less distinctive IDs like most sparrows, or harder to ID circumstances like winter or females or juveniles, or situations with weird lighting or harder angles it*

makes sense to use non-assertive.". On a similar line of thought, another participant says, "*I would prefer assertive explanations for birds that have distinctive traits over similar bird species, and non-assertive or neutral explanations for birds that are similar with characteristics that are more difficult to tell apart*".

6 out of the 136 participants commented on how the use of assertive explanations helped them realize various **characteristics of the birds**. For example, one participant values the assertive tone because it is "*... helpful for pointing out distinctive features that would help ID the bird.*". They also think that "*... the non-assertive language was helpful for species that share similar characteristics (aka clear, non-streaked breast) with other species that share that characteristic*".

4.5 Discussion

We investigate how imperfect XAI impacts humans' decision-making when collaborating with an AI. More precisely, we assess how imperfect explanations affect humans' reliance behavior and investigate the effects on the human–AI team performance. To answer RQ 1 and RQ 2, we assess the validity of our research model for two different types of explanations: natural language explanations and example-based explanations. Previous research emphasizes the need to consider imperfect AI when designing for human–AI collaboration [181]. With recent research looking into how humans and AI can achieve complementary team performance [16], Schemmer et al. [301] conceptualize the role of appropriate reliance in human–AI collaboration. We extend Schemmer et al. [301]'s framework by adding another dimension: XAI advice. Given that an explanation can be incorrect even if the AI advice is correct, it is crucial to understand the impact of incorrect XAI advice on decision-making. Furthermore, it is necessary to understand the impact of imperfect XAI for different types of explanations. Below, we discuss how our contributions are situated in current literature and the implications for CSCW.

In our study, we observe a significant moderation of humans' level of expertise on the effect of explanations' correctness on RAIR for both explanation modalities. However, we do not see this moderation for RSR. When humans are provided with wrong AI advice, their level of expertise does not moderate the impact of imperfect explanations on humans' RSR. We identify a direct effect of the level of expertise on RSR in both explanation modalities. Additionally, the correctness of explanations impacts RSR negatively for example-based explanations. Overall, our work synthesizes how humans' level of expertise impacts their reliance on AI when provided with imperfect explanations. Non-experts rely more on AI than experts, whereas experts rely more on their initial decisions. Especially for example-based explanations, imperfect XAI deceives experts' self-reliance and experts'/non-experts' relative AI reliance, fostering inappropriate reliance on the AI. Thus, this study sets a starting point to investigate the effect of imperfect XAI for different explanation modalities.

Our findings show that imperfect explanations impact human–AI decision-making. We observe that experts reach complementary team performance when imperfect explanations are provided. This holds true for natural language and example-based explanations. While non-experts do not reach complementary team performance, our analyses reveal that their performance can be improved to be similar to that of the AI performance. Moreover, there is a difference between reaching complementary team performance when the correctness, or fidelity, of the explanation changes (see the Appendix of the published manuscript). Previous research discusses the impact of explanations' fidelity on humans' reliance on AI and hypothesizes that fidelity has a positive impact on humans' reliance behavior on AI [136]. With our results, we confirm this hypothesis. Furthermore, Papenmeier et al. [267] observe that low-fidelity explanations (or incorrect explanations) impact user trust in AI when the global model performance is around 75% accurate, which helps validate our findings. We also observe that the lack of expertise among non-experts impacts their task performance when shown incorrect explanations, regardless of the AI advice being correct (see full paper). Similar to our findings, Nourani et al. [260] observe that non-experts tend to over-rely on AI advice, attributing this to their inability to identify when the AI is incorrect because of their lack of expertise. These findings contribute to a more integrated understanding of the impact of human–AI decision-making on different user groups in the presence of imperfect XAI. For example, this can inform managers on how to assign tasks to humans with different levels of expertise and provide them with explanations in different modalities. It could also lead to organizations modifying their human–AI collaboration workflows. From informal conversations with the product team of an AI decision-support tool[9] for biologists and conservationists to classify species and identify individuals from camera trap imagery, we learned that organizations using their tool have modified their workflow to incorporate "checks-and-balances". For example, intro-level biologists will collaborate with the AI to match individuals and then request a review of their "human–AI team" decision from a higher-up. In this unique human–human–AI collaboration scenario, the expert biologist could potentially correct situations when an intro-level biologist over-relies on AI advice because of an incorrect explanation.

Visual, example-based explanations are more deceptive than natural language explanations. To account for the impact of imperfect XAI on humans' appropriate reliance, we establish a novel metric **DoR**, (Deception of Reliance), to measure the difference in RAIR and RSR for correct and incorrect explanations. Our results indicate that people are more deceived by example-based explanations than by natural language explanations. In terms of RAIR (note that in RAIR cases, the AI advice is correct), experts and non-experts are deceived by incorrect explanations. In terms of RSR (note that in RSR cases, the AI advice is incorrect), experts are deceived by correct explanations. This is an interesting observation that may be explained by the consistency of the shown visual examples. For *correct advice, incorrect explanations*

[9]WildMe.org

cases, the XAI is providing three visual examples that show a different bird species than the bird species in the image to be classified (see Figure 4.6). Moreover, these three visual examples can belong to different bird species since the XAI is choosing the top three most similar bird images (in our study, this is in 90% of all *correct advice, incorrect explanation* cases). This inconsistency in example-based explanations might deceive experts and non-experts to not rely on the AI anymore when they identify visual differences in the images provided as explanations, disregarding the correct AI advice. We discover the same behavior for experts for *incorrect advice, correct explanation* cases. In those cases, the explanations consist of three images of the same bird species as the AI predicted. The incorrect explanations consist of three images that can be inconsistent in the bird species shown (in our study, this is 40% of all *incorrect advice, incorrect explanation* cases). Thus, this inconsistency in examples might deceive experts into no longer relying on themselves when they identify three consistent examples shown, disregarding the incorrect AI advice. Hence, the DoR of experts and non-experts is positive for RAIR cases as they are deceived by incorrect explanations, while experts additionally have a negative DoR and are deceived by correct explanations. Note that the overall RSR for experts is still higher than non-experts' RSR; the impact on deception caused by imperfect XAI is higher. As participants mentioned in the survey, it is more convincing that there is less uncertainty in the AI advice when three images that are similar to each other are shown than when three different images are shown. This corroborates our findings. This trend is not present in natural language explanations.

The language tone of explanations does not impact humans' decision-making behavior. Calisto et al. [54] conclude that the level of expertise influences whether the framing of the explanation should be assertive or non-assertive [54]. Based on their observations, they specifically suggest that natural language explanations should be designed such that the tone of the explanation is appropriate for the end user's level of expertise. Despite the numerous previous studies finding that the framing of the explanations has a significant impact on human–AI collaboration [54, 178, 179], our findings do not show an impact on appropriate reliance. Quantitatively, we do not find a direct effect of assertiveness on appropriate reliance. Qualitatively, we observe that the tone of the explanation should depend on certain situations, such as the AI's behaviors, instead of the human's level of domain expertise. We observe that 31% of the participants prefer assertiveness to align with the confidence of the model's prediction, while only 10% of participants prefer assertiveness to align with their confidence in the domain. This finding could be attributed to the fact that participants are collaborating with imperfect AI and are able to acknowledge that the AI advice was occasionally incorrect. However, given these qualitative suggestions and the potential for natural language explanations to present irrelevant or incorrect information [321], we encourage future work to explore various ways to alter the tone of an explanation based on the themes we identified in Figure 4.9.

Our findings can guide researchers and practitioners on how to assess and design

for imperfect XAI in human–AI collaborations. Regardless of the explanation modality, it is important to understand how humans interact with imperfect XAI. Visual explanations, such as example-based explanations and saliency maps, have been shown in the past to be of high educational value to the end-user (*e.g.*, [177, 226]), making it even more important to understand how to design for and mitigate imperfect XAI. This need is intensified with the role AI takes in organizational learning [325]. Especially in the workplace, AI can facilitate knowledge transfer and support organizations in retaining and distributing expert knowledge [155, 324, 372]. Similarly, it is also crucial to understand how imperfect XAI affects the learning of novices through AI-based learning systems [325] or through collaboration with AI [300]. Our findings can guide knowledge managers within organizations on how to make use of explanations for employees with different levels of domain knowledge. More precisely, knowledge managers should be aware of the impact of exposing humans with different levels of expertise to imperfect XAI. In addition, our findings contribute to a more integrated understanding of the impact that incorrect explanations can have on human–AI decision-making and inform different stakeholders in organizations. As non-experts are more affected by imperfect XAI, their performance drops more than experts' performance when incorrect explanations are provided in comparison to correct explanations. With this finding, knowledge managers can adjust their knowledge retention activities when training new employees; designers can adjust the development of human–AI collaboration systems to successfully facilitate explanations in decision-making and support humans in their work setting. Thus, we encourage practitioners designing human–AI collaboration systems to apply our findings to structure their design approach. This can aid organizations in laying out the strategic direction of human resource development by matching the use of explanations to humans' prior knowledge. Overall, these findings shed light on the ongoing discussion in CSCW on how to make use of explanations within work settings.

4.6 Limitations & Future Work

We elaborate on various limitations of our study, how they could impact the interpretation of our results, and identify opportunities for future work.

Lack of Information to Properly Identify Birds. Expert birders usually rely on more information than just the visual characteristics of a bird when determining the bird species, especially for ambiguous cases. For example, the location and habitat in which the bird was spotted can be imperative to determine the exact bird species within a family. It is unclear to what extent the lack of this information influences our results. Future technical work could consider using this information to help build more transparent bird classification models.

Correctness of Explanations versus Explanation Fidelity. Throughout our study, we consider explanations to be either incorrect or correct. However, as we mentioned, some

explanations that we classify as incorrect can contain evidence that is correct, making it difficult to only have a binary categorization for the correctness of explanations. While we use a binary scale for our analyses, explanation correctness, or fidelity, can be quantitatively measured on a continuous scale and categorized as low fidelity and high fidelity [267]. We encourage future work to explore explanation fidelity using multiple categories instead of two to understand the differences between low- and medium-fidelity explanations when it comes to task performance and appropriate reliance. This will provide insight into the impact that noisy explanations have on decision-making, such as when an explanation reveals some information that is aligned with the ground truth class and some information that is aligned with the predicted class.

Simplistic Natural Language Explanations. The natural language explanations are very limited and simplistic, although this is a fault of the model that we are using to generate those natural language explanations [137]. Compared to a field guide such as All About Birds [2], these natural language explanations could be correct for several species, given their lack of detail and their short length. This is potentially a side effect of the text descriptions in the CUB-200-2011 dataset being sourced from crowd workers instead of experts. We encourage technical researchers who are developing explanation methods that explain image classifications using natural language to pay careful attention to the text descriptions used for training the language model. For example, researchers could consider using text data from a field guide as their training data. We also encourage CSCW researchers to conduct more studies on how the explanation's length and level of detail impact experts' and non-experts' appropriate reliance.

Different Explanations Convey Different Information. While we do not intend to directly compare the two explanation modalities throughout our analyses, they are discussed in terms of similarities and differences throughout the article. Our main intention is to investigate how our research model holds across different modalities of explanations. While several previous studies compare multiple different types of explanation modalities qualitatively and quantitatively (*e.g.*, [67, 90, 174, 177, 335]), we encourage readers to avoid directly comparing the two explanations because they present different information. Previous research reveals that different explanation techniques can result in disagreements for the same dataset [292]. For example, the natural language explanations from Hendricks et al. [137] are feature-based, providing descriptions of features present in the image [137]. However, the example-based explanations present three similar images, which provide very different information from the natural language description of features. On top of that, the incorrectness in both explanation modalities is represented in different ways. While there are factual errors in natural language explanations that previous research addresses (*e.g.*, hallucination effects of natural language models [321]), there are logical errors (*e.g.*, inconsistencies) within example-based explanations. This opens avenues for future research to investigate how different human cognitive abilities (*i.e.*, cognitive styles)

impact the perception of these imperfect explanations in AI-assisted decision-making scenarios.

Visualizing Assertiveness for Example-Based Explanations. While there is previous research on how to visualize the confidence of a prediction for image classification [340], to the best of our knowledge, there is no work visualizing the assertiveness of example-based explanations. Given this, our example-based explanations use natural language to convey assertiveness. Visualizing the assertiveness for each example in an example-based explanation would provide the decision-maker with another cue about whether they should rely on the model.

AI Expertise versus Domain Expertise. Our analyses are based on the participants' expertise in bird species identification. We do not ask participants about their knowledge of AI. Thus, we do not analyze their expertise related to AI. It is unclear to what extent the participant's expertise with AI would impact the effect of imperfect XAI on appropriate reliance and task accuracy for bird species classification. Future work should consider looking at the effect that AI knowledge combined with domain expertise has on appropriate reliance, taking into account an imperfect XAI.

4.7 Conclusion

This article sets out a research model to investigate the effect of imperfect XAI on human–AI decision-making. Thus far, human-computer interaction and CSCW literature fail to thoroughly scrutinize how explanations' correctness affects humans' decision-making and their reliance behavior on AI. Hence, through a human study with 136 participants, we empirically analyze humans' decision-making and specifically assess whether their level of expertise and explanations' assertiveness moderate the effect of imperfect XAI on appropriate reliance. Furthermore, we explore the extent to which incorrect explanations deceive decision-makers' reliance on AI. With our findings, we make several contributions: First, we propose a research model to investigate the moderation of assertiveness and humans' level of expertise on imperfect XAI in decision-making tasks. We thereby extend the existing conceptualization of appropriate reliance by a new dimension of XAI advice. Second, through an empirical study, we reveal that imperfect explanations and participants' level of expertise affect human–AI decision-making for two different explanation modalities. In addition, we show the effect on complementary team performance and provide guidance for future studies on how to investigate imperfect XAI in the context of human–AI decision-making. Third, we propose a novel metric called Deception of Reliance (DoR), which allows us to measure the impact of incorrect explanations on decision-makers' reliance. Our results inform designers of human–AI collaboration systems and provide guidelines for their development. Fourth, we reveal which role the language tone in explanations plays and outline important dimensions that should be considered when designing for XAI advice.

Part II

Emerging Uses of AI Explanations

Chapter 5

Navigating Appropriations of XAI in a Deployed AI System

This chapter is adapted from the following co-authored paper:

Turri, V., **Morrison, K.** (*co-first authors*), Robinson, K. M., Abidi, C., Perer, A., Forlizzi, J., & Dzombak, R. (2024, June). Transparency in the Wild: Navigating Transparency in a Deployed AI System to Broaden Need-Finding Approaches. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1494-1514).

Part I reveals the impact of imperfect XAI on human–AI decision-making across two empirical studies in two different domains with real AI explanations and AI explanations provided by crowdworkers. Across both studies, Part I captures the mixed effect of output-oriented XAI on human–AI collaboration. These studies contribute to the growing body of work that similarly finds mixed effects of XAI in human–AI collaboration. This could lead to the conclusion that XAI is just not suitable for human–AI collaboration. However, these mixed results and limitations are primarily prevalent among output-oriented XAI designs. This raises the question of whether AI explanations have value in human–AI collaboration if used for different purposes. Therefore, Part II takes a step back to look at the emerging uses of XAI beyond verifying or interpreting AI outputs.

5.1 Introduction

As the field of AI matures, calls for transparency in AI systems from the responsible AI (RAI) research and practitioner community have become more prevalent. To date, research on delivering transparency for end-users has typically focused on explainable AI (XAI) methods. XAI aims to provide explanations for how models arrive at outputs for individual or classes of inputs. While XAI methods can be helpful for tasks such as discovering spurious patterns [316], recent work in XAI literature has shown that XAI methods lack actionability [99], may not advance model understanding [13], or are potentially harmful during high-stakes decisions [94, 330]. Furthermore, recent research within the FAccT community has argued that transparency is much more

than explainability [73, 108], or at least the current approaches to explainability [233]. Therefore, we align our work with the perspective from Liao and Vaughan [210], which states: "*Transparency is fundamentally about supporting appropriate human understanding, and this understanding is sought by different stakeholders with different goals in different contexts.*"

While transparency has become an important and debated topic in RAI research, AI practitioners have limited guidance on selecting, implementing, and maintaining transparency mechanisms for real-world systems. In practice, alignment between user needs and computational techniques for generating explanations is not always clear [144, 211, 213]. Although numerous studies suggest how to take a human-centered design approach to transparency need-finding (*e.g.*, [47, 66, 71, 100, 101, 128, 158, 212, 236, 306]), few leverage methods outside of the scope of interviews or user studies of the AI system's end-users. Triangulating varied data sources, however, has been shown to be a crucial aspect of building a comprehensive understanding of a problem space [58].

To explore gaps between transparency research and practice, we conducted an ongoing case study to implement transparency mechanisms for a deployed AI decision-support tool. We partnered with Wild Me, a wildlife conservation non-profit, which offers wildlife researchers AI-enabled decision support tools for identifying individual animals from uploaded images (*e.g.*, [24, 35]). Wild Me has offered a variety of transparency mechanisms within its Wildbook platforms over the past few years (*e.g.*, saliency maps visualizing the regions of the image that contributed the most to the prediction [307]). Other mechanisms outside of conventional XAI include match scores and model documentation. Wild Me introduced these transparency mechanisms to follow RAI practices and to give end-users methods for understanding and collaborating with their models. While the Wild Me team felt these mechanisms had been helpful to end-users, they wanted to reassess how they deliver transparency as they shift from interpretable models to deep learning models. Therefore, our case study goal was to identify the uses and challenges of the current transparency mechanisms and uncover stakeholders' transparency needs. Our need-finding indicates that users have unanswered questions about model behaviors and would like to provide transparency to additional stakeholders. To the best of our knowledge, this case study is the first to retrospectively analyze how transparency mechanisms have held up in a real-world deployment over several years using broader need-finding methods.

Our case study focuses on incorporating various data sources and connecting with a broad ecosystem of stakeholders to support our analysis. Specifically, we expand upon current need-finding approaches by analyzing Wildbook's community forum and user-made training documentation to provide unique insights into transparency challenges and needs. We also conduct in-depth need-finding with the ML and product teams, not just the end-users, and run experiments on Wildbook's MiewID model. By employing these additional methods, we identify tensions and overlaps across stakeholder needs, learn about technical limitations, and evaluate the efficacy of

currently deployed transparency mechanisms. We conclude by discussing how insights from our case study can shape new approaches to transparency need-finding. The lessons we learned during this case study are presented to provoke discussion within the FAccT community about how to identify and expand both the problem and solution spaces for AI transparency in decision-support tools. We present an overview of our methods in Section 5.2, analyses from interviews, experiments, and forum posts in Section 5.3, and discussion geared towards the FAccT community in Section 5.4. We summarize our work and contributions in Section 5.5.

5.2 Case Study Overview

We detail methods used during our 5-month case study below. Our methods are designed to understand end-users through a variety of data sources in addition to interviews, such as community forum posts and training documents. Additionally, we met repeatedly with the ML and product teams to capture their experiences with existing transparency mechanisms, ML models, and/or unmet end-user transparency needs. Figure 5.1 summarizes Wild Me users' workflow using Wildbooks, and Figure 5.2 summarizes our need-finding methods.

5.2.1 Wild Me's AI Decision-Support Tool: Wildbooks

Wild Me, founded in 2008, offers AI decision-support platforms called Wildbooks to assist end-users in identifying individual animals from input images. Each Wildbook supports a different species or group of species [24]. Wildbooks employ a pipeline involving object detection and matching algorithms to identify individuals from images. First, uploaded images are sent through an object detection model, which draws bounding boxes around localized animals (A1 in Figure 5.1). Users may accompany their image upload with metadata such as location, time, and custom keywords (A2 in Figure 5.1). Users can add or remove bounding boxes manually to ensure their accuracy. Next, users input the annotated images through one or more computer vision algorithms (A1 in Figure 5.1). Users are then presented with a list of the top-k closest individuals and image matches (part B in Figure 5.1).

Users can click on each of the suggested IDs within the ranked list to view their uploaded image alongside potential matches. To assist users in confirming matches, users are shown "match scores" for each ranked potential match (B2 in Figure 5.1) and given the option to view a saliency map of their input image and potential match image (B1 in Figure 5.1). These saliency maps, presented as heatmaps, highlight areas of importance in the model's matching process. When users want to view the saliency map for a potential match, the saliency maps appear in a new tab window. Users confirm all matches themselves. According to the documentation provided by Wild Me, "the 'score' indicates how similar the uploaded image is to potential matches within the Wildbook database". Match scores are computed differently depending

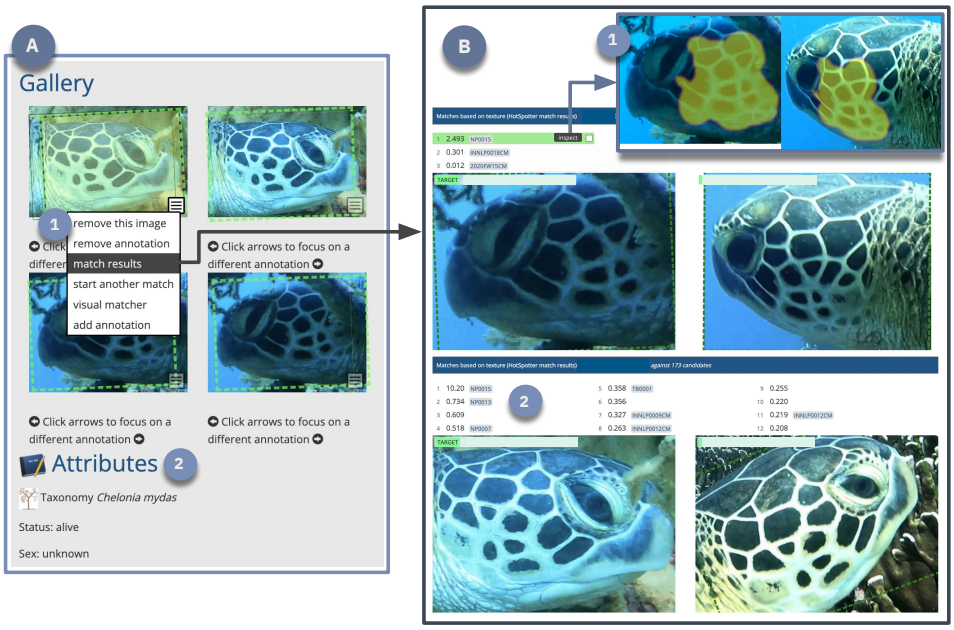


Fig. 5.1: Workflow and features represented from the Internet of Turtles Wildbook. For every encounter ("*a sighting of a single animal at a specific location and time*"), users are shown the image/s and can add metadata, such as the sex. Users can attempt to match the unidentified individual (A1) to another already in the database. Matched results appear in a new tab (B) and will show separate results for each unique image in the encounter and/or each unique model available for that species. (B1) Users can inspect a match result, which reveals a saliency map to rationalize the model's result. (B2) Users can also rely on the similarity "score" to determine the best match result.

on the model, with models employing a range of metrics such as cosine similarity to identify potential matches from within WildBook's private or public databases.

Wildbooks support several different computer vision models, such as Hotspotter (an interpretable model) [76] and MiewID (a deep learning model) [371]. Users can communicate with other users or the product or ML team about workflows, bugs, and feature requests via their online community forum. Moving forward, Wild Me plans to develop and deploy multi-species MiewID deep learning models for more species and communities. As part of this transition, the Wild Me team is interested in reflecting on existing transparency mechanisms within Wildbooks and considering ways to replicate the interpretability afforded by models like Hotspotter as they rely on more advanced deep learning methods.

5.2.2 Need-Finding Methods

Meetings with ML Team [MT]. We attended weekly hour-long meetings (a total of nine meetings) with the Wild Me ML team as participant observers to understand the technical underpinnings of the Wild Me models and systems. Meetings were not recorded; however, three members of the author team took notes throughout

the meetings. Meeting insights were discussed with the entire author team to derive implications and themes. During the meetings, members of the ML team would walk through experimental results, such as comparisons of a MiewID model's performance (*i.e.*, rank-x and mAP scores) for each species within a multi-species dataset. These meetings were attended sporadically by additional researchers at the intersection of ML and conservation work. Our interactions with the ML team led us to perform our own training and testing of the MiewID model to further explore model behaviors.

Investigation of the Tool's Machine Learning Models [ML]. To build a deeper understanding of Wild Me's deep learning models, we trained and tested a MiewID model for beluga whales[1] on the "Where's Whale-Do?" beluga whale dataset[2]. Our investigation gave us firsthand insight into the training, testing, and deployment pipeline used by Wild Me, as well as how the current transparency mechanisms are implemented. We chose only to explore the MiewID model because WildMe intends to use this model as their primary model moving forward.

Meetings with the Product Team [PT]. As suggested by Liao et al. [212], we incorporated the designers and product team early on in the need-finding process. To learn about Wild Me's perspectives on end-user behaviors and needs, we held six one-hour-long meetings with the Wild Me product team, which consists of Wild Me's lead product manager, customer experience engineer, and director. Meetings were not recorded; three members of the author team took notes during these meetings. Meeting notes were discussed with the rest of the author team afterwards. At the beginning of our case study, the product team onboarded us on a Wildbook platform and gave us an overview of the system. After our onboarding, these meetings shifted to learning about existing transparency mechanisms and rationalizations behind current design choices. We also asked the product team about their ideas for improving the system based on their knowledge of the users' experiences. Through these meetings, we were informed of training materials developed by Wildbook end-users, which we reviewed as part of our need-finding.

Analysis of Community Forum Posts [CF]. We scraped and analyzed a year's worth of posts (from October 2022 through October 2023) from Wild Me's community forum. The community forum is a public online resource for users to post and ask questions. While many topics are discussed on the forum, for the purpose of this study, we limited the scope to posts related to the "Feature Requests" and "Bug Reports" tags to uncover users' transparency needs. After this filter, we scraped 362 posts. We conducted open and axial coding analysis to analyze transparency uses, challenges, and needs [373]. More details about the coding process, including codes and example quotes, can be found in the Appendix of the published manuscript.

Interviews with End-Users [EU]. As suggested by previous literature (*e.g.*, [212, 355]), we conducted semi-structured interviews with experienced Wildbook end-users.

[1]We trained the MiewID model using code provided in Wild Me's MiewID GitHub repository: https://github.com/WildMeOrg/wildbook-ia

[2]https://lila.science/datasets/beluga-id-2022/

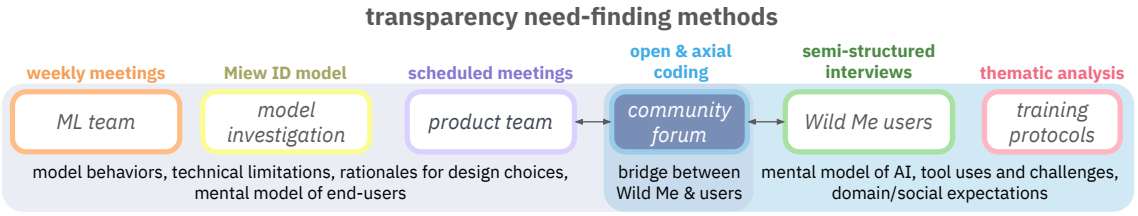


Fig. 5.2: Our need-finding methods consisted of meeting the ML and product teams and experimenting with the MiewID model. These three sources provided information related to model behaviors, technical limitations, and rationales for technical and design choices. We interviewed end-users and analyzed their training protocols to understand their mental models, use cases, and expectations for Wildbooks. Lastly, we analyzed the community forum posts as a window into interactions across and within stakeholder groups.

During the interviews, participants were asked to detail their motivation for using Wildbooks and their experiences and perspectives on current transparency mechanisms. We leveraged a contextual design inquiry approach by having users walk us through their workflow with Wildbook, from collecting data to confirming potential matches [27]. We also asked modified questions from Liao et al. [212] to elicit users' transparency challenges and limitations. We recruited from a list of participants among the top 100 most active profiles on the community forum leaderboard with activity in the past year or among the top active profiles for the past couple of weeks. We also made a community forum post with our recruitment call and sent a recruitment call in the monthly Wild Me newsletter. Participation was restricted to those in the United States who were at least 18 years old. Interviews lasted for one hour over Zoom. Interview transcripts were analyzed through affinity clustering.

Analysis of Training Documents [TD]. Inspired by Cai et al. [52]'s work identifying the onboarding needs of clinicians using AI systems, we analyzed user-made training documents from four different communities. As part of the onboarding process, research teams share training documentation via Google Drive. Wild Me's official training documentation is available on their site[3] and on their YouTube channel [4]. Examples of training materials include video walk-throughs for using the system and written protocols for image collection or matching. By going through the training material, we identified all potentially relevant quotes related to model use and transparency mechanisms. Common themes were identified by clustering the quotes in an affinity diagram.

5.3 Findings

Our approach to need-finding enabled us to surface nuanced perspectives, opportunities, and limitations related to Wild Me's existing transparency mechanisms. To validate

[3] https://docs.wildme.org/product-docs/en/wildbook/getting-started-with-wildbook/
[4] https://www.youtube.com/@wildme3451

Table 5.1: A summary of the five end-users interviewed, each using the tool for a different animal: sharks (Sharkbook), whales (Flukebook), sea bass (Giant Sea Bass Wildbook), sea turtles (Internet of Turtles), and snow leopards (WhiskerBook). We summarize their main role in their organization and motivation to use the Wildbook.

Wildbook	**User Role**	**Years of Use**	**Use Case**
Sharkbook	Citizen Scientist	9 years	Maintain a record of locations that seven-gill sharks visited and how frequently they visited those locations to have
Flukebook	Marine Mammal Researcher	10 years	Create an archival catalog of sperm whales that includes a mapping between photo ID and genetic ID
Giant Sea Bass	Graduate Student	7 years	Working with the local government to produce a population density estimation to help maintain endangered species management
Internet of Turtles	Teaching Professor	4 years	Identify the impact of increased civilization in coastal regions by working with local sea turtle hospitals
WhiskerBook	Conservation Researcher	3 years	Calculating population growth over time to determine the effect of conservation policies on the snow leopard population

the need to broaden approaches to transparency need-finding, we present how each need-finding method helps us understand transparency solutions within the context of Wildbooks. Instead of highlighting findings from each method separately, we group relevant findings from each source into distinct themes: uses, challenges, and needs. First, we present meta-analyses of our need-finding methods. Then, we highlight how different stakeholders use existing transparency mechanisms in subsection 5.3.1. In subsection 5.3.2, we highlight specific challenges and limitations with the tool's existing transparency mechanisms. Lastly, in subsection 5.3.3, we connect insights from the need-finding methods to end-user needs not currently addressed by existing transparency mechanisms.

Need-finding Meta-Analyses. We interviewed five end-users: a citizen scientist who uses the Sharkbook, a Marine Mammal researcher who uses the Flukebook, a conservation researcher using the WhiskerBook, a teaching professor using the Internet

of Turtles book, and a graduate student using the Giant Sea Bass Wildbook. As seen in Table 5.1, some users leveraged Wildbooks solely to maintain a catalog of the species. By contrast, others used Wildbooks to understand or influence conservation policies. All of these users have been working with Wildbooks for at least three years and have been manually identifying species in their domain for even longer.

Regarding the community forum posts, we ended up with 71 codes after an open and axial coding of the 362 posts. Most of the forum posts were not related to the transparency of the models but instead were about feature requests (33% of posts) and confusion about the system's design/workflow or how-tos (41% of posts). The remaining posts were about errors (15% of posts), transparency of the system (less than 5% of posts), or miscellaneous topics (6%). The resulting codebook can be found in the Appendix of the published manuscript.

The four training materials created by end-users were produced for different audiences with a variety of formats. For example, the Internet of Turtles interviewee's training material is geared towards undergraduates collecting field data and uploading images to the Wildbook. This single-document training protocol does not go into detail about the process of matching individuals because undergraduate students are tasked solely with collecting data for upload. However, the training materials from the Giant Sea Bass Wildbook interviewee, the WhiskerBook Interviewee, and an African Carnivore Wildlife end-user all target users who are collecting field data and identifying individual matches. These detailed training protocols, consisting of numerous documents and videos, include content about interpreting individual match scores and determining which models to use. The training protocols for these three Wildbooks also detail how to annotate images prior to input into a matching algorithm.

5.3.1 Usage of Existing Transparency Mechanisms

Transparency mechanisms have been integrated within Wildbooks for several years, and saliency maps have been implemented for at least three years. As a result, we were given the unique opportunity to learn how well popular transparency mechanisms have worked in practice. Our findings indicate that both mechanisms (saliency maps and match scores) can provide value to end-users and the Wild Me team. These mechanisms are informative to developers and help the product team support end-users' transparency needs. We detail specific uses observed using each need-finding method for both transparency mechanisms below; Table 5.2 provides a high-level summary of uses.

5.3.1.1 Saliency Map Uses. ML Team [MT], [CF]. Our discussions with the ML team taught us that saliency maps are an essential part of their model training and evaluation workflow. The ML team uses saliency maps to help convey model performance issues during their meetings. For example, during one session, one of the ML engineers walked through a graphic depicting examples of correct and incorrect

Table 5.2: Wild Me's tool currently offers two transparency mechanisms: saliency maps and match scores. This table provides a link between the need-finding methods and the key uses of and challenges with the transparency mechanisms. Insights are mapped to the community forum analyses [CF], end-users' training documents [TD], end-user interviews [EU], meetings with the product team [PT], meetings with the ML team [MT], and model experiments [ML].

Technique	Uses (Section 5.3.1)	Challenges (Section 5.3.2)
Saliency Maps	• [MT] Identify model behaviors, such as blindspots or spurious patterns • [MT], [CF] Communicate unexpected model behaviors • [PT] Incorporates saliency maps in the tool to help build end-users' trust • [EU], [TD] Consult while confirming a match as a second source • [EU], [TD] Communicate with external stakeholders and new users	• [ML], [MT] Saliency maps for some models are ambiguous or uninterpretable • [MT] Need to prioritize computationally efficient saliency maps • [EU] The discoverability of the saliency maps feature is low • [EU] Need more clarity on the goal of saliency maps • [EU], [CF], [PT] Need more guidance on where and when to find the saliency maps • [EU], [CF], [TD] Saliency maps can be wrong or convincingly misleading
Match Scores	• [PT] Incorporates match scores in the tool for power users to consult while reviewing match results • [EU], [CF], [TD] Consult to reason about the recommended matches "confidence", "accuracy", or "quality"	• [MT], [ML] Score ranges and meanings not standardized across models • [PT] Presenting how the match scores in a user-friendly way • [CF] Unclear on how to interpret and compare match scores

model predictions with their accompanying saliency maps to communicate aspects of a model's performance to the rest of the ML team. During another meeting, an ML engineer showed saliency maps with ambiguous trends from one of the newer models still under development to discuss how to better explain that model's predictions. ML engineers also leveraged saliency maps within the community forum to address a user's comment about errors in image segmentation. Overall, the ML engineers rely on the saliency maps as a diagnostic tool for understanding model behavior and as a way to communicate these behaviors to others.

Product Team [PT]. According to the lead product manager and customer expe-

rience engineer, saliency maps are used by their team and within their products to build trust among end-users. The product team also discussed prior user research involving saliency maps, which indicated that maps are important for cultivating trust and adoption. The team also shared Figma prototypes of new designs for Wildbook's interface, which indicated that saliency maps will be used moving forward.

End-users [EU], [TD]. Through interviews with end-users, we learned that end-users view the system as a way to narrow the search space for potential matches. However, users still manually match individuals by visually analyzing their unique patterns, markings, or spots (Figure 5.1, part B). Matching methods vary by species and research team based on pre-AI workflows for identifying individuals. Many users identify and validate key points across images without using saliency maps at all, instead zooming in on the image for closer manual inspection. For example, the Internet of Turtles interviewee said, "*Sometimes I will zoom in on the picture...so I haven't used the [saliency maps] that much.*" While many users are inclined to match individuals manually, some users, such as the WhiskerBook interviewee, leverage saliency maps as a starting point for identifying shared features across individuals. Figure 5.1 highlights how the saliency maps from two images cover the same scales on a turtle, which the user can then use to confirm the match. Conversely, the Giant Sea Bass Wildbook interviewee uses platform-specific visualizations that highlight specific points instead of regions.

While end-users may not rely heavily on saliency maps during the matching process, maps can be used to communicate information about how Wildbooks work to external stakeholders and newer users. For example, the WhiskerBook and Internet of Turtles interviewees have both leveraged saliency maps when sharing results with outside stakeholders (*e.g.*, local policymakers, government officials, and research audiences) to justify their use of the system and the validity of their research. Our Internet of Turtles interviewee also imagined how they could use saliency maps as part of the onboarding process to demonstrate how the models work.

5.3.1.2 Match Score Uses. ML Team [MT]. Match scores derived from Wildbook models are used to determine the top-k closest matches to an input image. While match scores could hypothetically be analyzed during the training process as an indicator of model calibration, our meetings with the ML team indicated that scores are not consulted within their current workflow.

Product Team [PT]. While the lead product manager and customer experience engineer feel that match scores can be confusing for end-users to interpret, the product team also shared that many end-users develop their own intuition for interpreting match scores to assist in their matching process. The product team has considered removing the match scores to encourage users to rely solely on the rankings instead of the score. However, they've opted to keep match scores because they have learned from prior user research that scores have been helpful to many power users.

End-users [EU], [CF], [TD]. Many experienced users have developed their own systems for interpreting match scores. For example, when the Internet of Turtles

interviewee "*...sees a [match score] above 1...there's a pretty good likelihood...that's a match.*" Alternatively, if the interviewee sees a score below 1: "*I'll look at it the [potential match], but I know that there's not really anything there.*" Across training materials produced by end-users, we saw that higher match scores were described as indicating a *higher likelihood* of a match. In the training material for the Giant Sea Bass Wildbook, for example, users are advised that when scores are "*...greater than 200 [it's] a very high probability of a match... 2 is the threshold for potential matches.*" Similarly, the training video from the WhiskerBook Interviewee discusses the match score as "*...the ranking of how well the images matched to each other.*" Like saliency maps, match scores are just one factor in navigating potential matches, as all interviewees stated they still manually match the result just to be sure. For example, the Internet of Turtles interviewee described the match scores as "*...a very generic starting point...*". Questions about match scores on the community forum further corroborate that they are used to some extent during the matching process.

5.3.2 Challenges with the Existing Transparency Mechanisms

Aside from use cases, we also learned about current challenges related to transparency mechanisms. Due to model and system constraints, these challenges are often connected to design and technical choices made by the Wild Me team. We saw that transparency mechanisms can unintentionally introduce confusion to end-users within the matching process. We detail specific challenges observed from each stakeholder for both transparency mechanisms below; Table 5.2 provides a summary of challenges.

5.3.2.1 Challenges with Saliency Maps. ML Team [MT], [ML]. We learned from the ML team that while saliency maps for interpretable models produce visualizations that highlight matching features across the input and potential match images, saliency maps for their deep learning models can produce unusual gridlike patterns that are hard to understand. During our own experiments with the MiewID model, we struggled to produce reasonable saliency maps using a different visualization approach. As a result of the misalignment between saliency methods and the MiewID model, saliency maps have become less usable over time in this context. Additionally, the ML team voiced that generating saliency maps can be computationally inefficient, which only adds to the already long processing time needed to identify matches. These challenges suggest that alternative transparency solutions to saliency maps may need to be explored moving forward.

Through our experience using the MiewID code, we learned that the current system utilizes GradCAM++ to generate saliency maps [64]. Discussions with the ML team revealed that the team found this visualization method to be one of the faster techniques. However, challenges in applying traditional saliency map techniques may arise as the Wildbook models are for information retrieval rather than classification, which means these models produce embeddings rather than classification scores.

Product Team [PT], [CF]. The product team faces challenges in informing end-users

of the saliency map capability. While saliency maps are included in a number of demos and across documentation related to Wildbooks, the product team has seen on the community forum and within their user research that end-users are often unaware of how to access these features. Further, the tab for accessing the saliency map is labeled the "inspect tool", which can lead to confusion about what the tool is and the information it provides. Discussions on the community forum between the customer experience engineer and end-users revealed that saliency maps are discarded 1-2 weeks after user uploads are matched, at which point the button for inspecting the saliency map disappears. This decision was made as a result of technical constraints related to storage. The product team has had to field numerous questions and complaints related to the unavailability of saliency maps in the community forum.

End-Users [EU], [CF], [TD]. While the WhiskerBook and Internet of Turtles interviewees were aware of and used the saliency maps feature before, community forum posts and discussions with the Flukebook and Sharkbook interviewees reveal that many users do not share the same experiences. Many users are unaware of the existence of saliency maps, struggle to access them, and/or rely on alternate visualizations. For instance, the Sharkbook interviewee was unaware of the feature and uncertain of its meaning after being shown the feature during an interview. Examples of platform-specific alternatives to saliency maps used by interviewees include "trailing edge" visualizations for whale flukes and "spot-based" visualizations for giant sea bass.

Users who are aware of saliency maps mentioned pain points related to their use, including the expiration of saliency maps. To quote a post from a community forum user, "*I'd like to see the inspection results, but I'm getting this error 'inspection image unavailable (likely outdated)' on every attempt.*" Saliency maps can also inaccurately highlight mismatched elements across images, as warned by the WhiskerBook training materials, making saliency maps difficult to use or potentially misleading during manual matching. Per the WhiskerBook interviewee's experience: "*Sometimes it is wrong, and it's not useful.. sometimes the match is not correct, and it will match on a pattern that is not the same.*"

5.3.2.2 Challenges with Match Scores. ML Team [MT], [ML]. According to the ML team, the meaning of match scores varies across algorithms. For example, the MiewID algorithm uses cosine distances between embeddings to identify the top-k matches. The scores shown to end-users are the cosine distances between the target embeddings and (pre-computed) match candidate embeddings. Through our close investigation of the MiewID model, we learned that the embeddings are generated by training an EfficientNet model [184] with an ElasticArcFace module. During evaluation, the ElasticArcFace module is removed, and embeddings are extracted from the final layer of the EfficientNet. This differs from the Hotspotter algorithm, which uses a variation on Local Naive Bayes Nearest Neighbor (LNBNN) to calculate potential matches. The ranges of match scores and the information they communicate will therefore vary across algorithms and trained model instances.

Product Team [PT], [CF]. The product team frequently responds to questions

from end-users about how to interpret match scores. Similar questions were seen in the community forum, where users often ask what match scores represent. Because match scores can mean very different things, conveying their meaning to end-users can be challenging. The product team responds to user questions about match scores on the community forum with responses akin to, "*All three algorithms have relatively inscrutable scores. They're not probabilities and not especially meaningful but are rather distance metrics of one form or another and will vary from the trained instance of the algorithm to another trained instance.*" Another definition in a similar community forum post reads: "*The more visual texture similarity, the higher the score.*" Despite the product team adding a statement about this caveat in their own training documentation, they continue to encounter questions about match scores on the forum.

End-Users [CF]. Based on our analysis of the community forum, end-users are hesitant about interpreting match scores. One user on the community forum asked, "*Regarding 'Match score' in Wild Me Documentation, it is stated that, 'The match score represents the numeric value returned from the algorithm.' But what exactly are these numeric values? What do they represent: percentage, a score from 0 to 1, or 0 to 10? Or something else?*" Many end-users incorrectly interpret the match scores as probabilities. On the community forum, users ask questions such as "*Is the highest score 1?" and "The CurvRank v2 has 2 results [with a score] around 5. If the highest rank [score is] 10, [do] the results have a confidence of 50 percent?*" and "*FinFindR has 2 results around 0.2. What is the maximum score? What is the confidence of the matching [Whale X]?*" Ultimately, users struggle to interpret match scores and are unable to compare scores across algorithms since these scores may fall within different ranges.

5.3.3 Broader End-User Needs

Our analyses of end-user interviews, community forum posts, and training documents helped us form an understanding of end-users' mental models of AI and their workflows throughout the entire Wildbook pipeline. Talking to the ML and product team helped us situate those needs within the limitations of the system design and models. Based on our analyses, we learned that while current transparency mechanisms are targeted toward the match exploration stage, users have transparency needs across stages of the pipeline leading up to and following the match process. This is also seen in related works such as Corti et al. [75].

5.3.3.1 Distribution of Acceptable Model Inputs [PT], [MT], [CF], [TD]. We identified unmet transparency needs related to the image annotation process throughout our data sources. Before users can view potential match results, they must oversee the upload and annotation process. While the annotation process is automated using object detection models for some Wildbooks, users sometimes need to manually add or correct annotations due to model failures. During the annotation process, users

also have to assign a viewpoint of the animal to the image (*e.g.*, top, left).

Discussions in the community forum reveal that users grapple with uncertainties about which attributes of model inputs are likely to cause failures during the annotation process. Numerous users have expressed frustration after trying to bulk upload a set of images only to learn that the type, quality, and/or size of their images were not accepted by the system. For example, one community forum post discusses the unexpected importance of crop ratio, "*The Hotspotter algorithm does not seem to be able to match against images other than with the same crop aspect ratio. So if there are no images with the exact same crop aspect ratio, there are no matches.*" Similarly, discussions on the community forum suggest that researchers expect viewpoints to be heavily factored into the AI-enabled matching process, as the viewpoint is critical to their manual matching processes. Conversely, the WhiskerBook training documents indicate that users believe the algorithm matches against all images, not just those with the same labeled viewpoints, so annotation issues will not impact matching. Interestingly, concerns about model inputs and annotation failures were not brought up during the interviews with end-users, although they were prevalent on the community forum.

The process of curating acceptable model inputs is further complicated by end-users' pre-AI notions of image quality. Through the product team, we learned that wildlife conservation researchers grade the quality of their images for individual identification using image quality and individual distinctiveness scores [113]. According to the ML team, some users include these scores within the metadata for their images, but their systems for rating images vary widely across users and domains. The concept of distinctiveness scores may influence end-users' mental model of high- versus low-quality images and may lead to misaligned definitions of image quality between the end-user and system specifications. End-users are left to make informed guesses about how their annotations impact their matches.

5.3.3.2 Explore and Compare Algorithmic Performance [EU], [TD]. Performance metrics for Wildbook models are not generally available to end-users. As a result, users are left to fill in the blanks on model performance through their own analyses and/or develop protocols about which algorithms to use based on their experiences. For example, a group of WhiskerBook users conducted quantitative analyses of model performance for different k-values and published these findings for the snow leopard research community [37]. This information can then be used to inform user protocols for reviewing potential matches. The WhiskerBook training material, for example, advises WhiskerBook users to consult multiple models. By contrast, the Flukebook interviewee consults Wildbook models consecutively, analyzing results from the first model and looking at the second model only if a match is not already found. Alternatively, the Giant Sea Bass Wildbook interviewee has determined differences between the two models from "trial and error": "*...I3S [one model] I feel like is more precise, but they give less options...maybe there's a potential chance for like 5% error where Groth [another model] I feel like is probably like a 97% chance of having it within the 20*

matches." These findings showcase how users have compensated for not being provided information about model performance (*i.e.*, through trial and error or through their own comparison experiments). Users are willing to go to great lengths to conduct these analyses because they are essential to their work.

5.3.3.3 Human–Human–AI Collaboration Workflows [PT], [CF]. Although experienced users typically review and finalize matches on their own, many research teams employ an offline system for validating matches made by newer users or handling challenging matches. Some Wildbooks, such as the Giant Sea Bass Wildbook and Sharkbook, invite the community to take part in uploading their own images and suggest potential matches for the experts to approve. These human–human–AI collaboration workflows, in which a user reviews AI suggestions and collaborates with other users to make a final decision, often take place behind the scenes of the system's workflow. This process aims to maintain the quality of the animal ID database and prevent errors from propagating, such as an individual animal incorrectly being identified as two separate entities throughout the database. According to the product team and "Feature Requests" on the community forum, users have asked for the Wildbooks to offer features that accommodate their collaborations. On the community forum, users have requested a new kind of user role with limited access privileges "*...allowing them to record an encounter, consult the encounters, launch a match, and see the results but not to validate this match or to assign an individual to a photo.*" To finalize the match, "*...[we] would then let only selected and more experienced users perform this critical last assignment step.*" Alternatively, users on the community forum have also asked for "peer review" functionality, in which case "*matches validated by the contributors are in an 'unapproved' state by default, and there is then a validation step when someone to whom we have assigned this role can edit it in an 'approved' state.*" While the interviews with end-users indicated that users participate in human–human–AI collaboration for difficult matches, they did not reveal details to the extent found on the community forum.

5.4 Reflections & Insights on Navigating Transparency in Deployed AI Systems

Throughout our 5-month case study with Wild Me, we utilized diverse data sources to paint a varied picture of current transparency mechanisms. We uncovered from interviews with end-users and analyzing user-created training materials that user workflows often involve manual review of AI outputs and complex review structures. We learned that the ML team uses saliency maps to investigate model behaviors and communicate those behaviors to others. However, through our experiments with MiewID, we saw firsthand how these maps can become challenging for deep learning models. By contrast, the product team uses existing mechanisms to build trust among end-users, although analyzing the community forum posts expose how users can also

struggle to interpret or find these mechanisms.

Our methods enabled us to identify tensions between stakeholders, which we can mitigate by prioritizing functional needs and finding dual-use solutions. For example, the ML team's challenge of providing computational efficiency ultimately needed to take precedence over other stakeholder concerns and constrained the solution space. However, investing in methods for interpreting match scores, for instance, could have positive implications for the ML developer and end-user workflows while remaining computationally inexpensive.

While interpreting our reflections, it should be taken into consideration that our research setting may be unique in that we have direct access to a variety of stakeholders and data sources, such as the community forum. Additionally, we are working with a deep learning model that performs an information retrieval task, and the nature of the system and user workflows may be somewhat specific to our case study. However, we do our best to provide generalizable reflections for the broader FAccT community without overextending their implications. We situate our findings within related literature and the FAccT community by reflecting on our approach below.

Need-finding with the ML team and the deployed models provides necessary insight into technical limitations and challenges. While previous transparency need-finding frameworks have included model developers in their process (*e.g.*, [212]), the model developers are leveraged primarily to implement transparency solutions. In our work, we positioned technical considerations to be front and center by considering the ML teams' perspectives to be on the same level as end-users. Being hands-on with the machine learning model used within the system, as opposed to only hearing from the ML teams about how the system works, exposed us to additional model behaviors and technical limitations related to transparency that we would not have otherwise discovered. For example, using ML models for information retrieval within the system means that saliency maps need to be generated in a way that differs from classification contexts. Overall, our meetings with the ML team taught us about model behaviors that end-users may be unaware of.

While this method exposes valuable and relevant insights, gaining access to the ML team and models may not always be possible, given the researchers' networks or data-sharing policies. Additionally, the insights that arise can mislead researchers if this method is prioritized more than an end-user-centric method. We encourage transparency researchers to conduct need-finding closely with the ML team and the deployed models early on in the transparency need-finding process to gather the necessary information to fuel transparency solutions, such as behavior descriptions of the model [49] and onboarding modules [53]. Ultimately, some form of internal support from the ML or development team should be a requirement for deploying transparency mechanisms.

Community forums for AI tools can expose valuable crowdsourced information about transparency needs and challenges. Although analysis of large-scale crowdsourced data can be time-consuming, and a small portion of the data may

be specific to transparency, we found the community forum extremely valuable in identifying needs across the wider user base. Previous work has also found end-users to share model blindspots and failures they encounter online, such as on Twitter [207]. The value of leveraging end-users' crowdsourced insights for improving model transparency and fairness has been emphasized by several recent workshops and studies (*e.g.*, [48, 84, 85, 196]). As shown by our findings and previous transparency research, analyzing end-user content beyond semi-structured interviews is vital for identifying and understanding transparency needs. As such, we recommend that AI practitioners consider building community forums for their users. We also recommend that transparency researchers value insights from community platforms and other social media, as these can be useful for triangulating user data [58].

Exploring how end-users compensate for unmet transparency needs helps identify potential transparency solutions. A vital step in the need-finding process is the observation stage, allowing researchers to understand users' current workflows and identify their needs without explicitly asking users what those needs are. Very few existing processes for transparency need-finding suggest leveraging observational or ethnographic studies to understand users' current workflow with the AI system [47]. Furthermore, none has suggested analyzing community forum posts or user-designed training documentation. While interviewing end-users, we discovered that many users were creating their own training documentation on how to collaborate with the system and interpret the predicted match results. Cai et al. [52] observed clinicians having a similar need for users to be onboarded to the AI system, corroborating our findings. We also learned that users rely on the community forum as a way to expose the transparency of the models they work with. From understanding model blindspots in the community forum to comparing the models in training documentation, end-users have created numerous means to compensate for unmet transparency needs. Outside of the matching process, users have questions about model failure modes, image quality requirements, and algorithmic performance. The information end-users are seeking by their own means can be addressed by building on ideas suggested from previous FAccT contributions, such as model cards [77, 235] and data cards [276]. Furthermore, a user study with non-expert analysts highlights the significant value added by providing an interactive tool that offers information about a model's uses and blind spots [77]. While we identified users' needs to align with various transparency solutions, such as model cards (*e.g.*, [235]) or onboarding methods (*e.g.*, [53, 242, 243]), many transparency taxonomies fall short of suggesting solutions beyond local explanations [322]. We encourage the FAccT community to consider incorporating contextual inquiry-inspired approaches within transparency need-finding to broaden the transparency solution space. Ultimately, incorporating contextual inquiry or observational studies in the transparency need-finding process can help researchers and practitioners uncover how end-users compensate for currently unmet transparency needs. In the absence of training materials, researchers may benefit from working with end-users and the developer team to create documentation to see how transparency needs or uses arise.

Transparency solutions must be adaptable to complex human–human–AI collaboration workflows. As revealed by many of our interviewees, end-users often work within a larger team with members with varied levels of experience, matching species within their domain. End-users may hand more challenging matches to senior members of their team for review. This is not the only place where human–human–AI collaboration workflows exist. For example, radiology students and residents may collaborate with AI to help them localize and/or grade tumors and then pass the patient case to their attending for validation or further investigation [175]. With the exception of works such as McNeese et al. [231] and Munyaka et al. [244], few papers investigating human–AI teaming explicitly acknowledge human–human–AI collaborations. To our knowledge, no previous works recognize the need to adapt transparency mechanisms to such workflows. From our analysis of the community forum and discussion with the product team, we identified the need for transparency solutions to be adaptable to complex human–human–AI collaboration workflows within Wildbooks. Future research is needed to explore transparency mechanisms in the context of human–human–AI collaboration scenarios. One such solution could be to consider collaborative saliency maps, where users can annotate regions of the map they believe are convincing or misleading for their human collaborators to review. Another solution could be to consider offering users the option to explicitly identify which model they used to identify a match and a short rationalization for their decision. In terms of broadening transparency need-finding approaches, we recommend that researchers identify not only the end-users' workflow but also the processes that exist before and after the end-user interacts with the data or model.

5.5 Conclusion

Current transparency need-finding approaches often focus primarily on end-users and use this information to determine XAI solutions. Although conversations with end-users are an essential component of understanding user needs, users may have challenges that are missed or overlooked without a broader context and cannot be addressed by XAI solutions alone. This work sets out to broaden transparency need-finding approaches within FAccT and the larger RAI research community by demonstrating what multi-stakeholder engagement looks like in the context of a 5-month case study centered around a widely-deployed AI-decision support tool. By leveraging a diverse set of need-finding methods (*i.e.*, meetings with the ML and product teams, experiments with an ML model, analysis of community forum posts, interviews with end-users, analysis of user-created training documents), we investigate the utility of currently deployed transparency mechanisms and provide insights about end-users' unique transparency needs. Our discussion highlights recommendations for transparency researchers to consider when approaching need-finding in practice. We intend to experiment with a variety of transparency solutions moving forward such as

reframing match scores, providing algorithmic performance information to end-users, or building explanations for human–human–AI interactions. While our work does not cover all possible approaches for identifying transparency needs, we invite researchers from FAccT and beyond to leverage such approaches as a stepping stone to adopting more varied need-finding methods in their work.

Chapter 6

Exploring Emerging Uses of XAI in Clinical Decision-Making

This chapter is adapted from the following paper:

Morrison, K., Li, Z., Kim, M., Hong, C., Kraisangka, J., Correa-Jaque, P., Fauvel, C., Sahay, S., Vanderpool, R., Everett, A., Lin, S., Kanwar, M., Benza, R., & Perer, A. (2026, July). It's Not Just for Trust: Designing for Emerging Uses of Explainable AI in Clinical Decision-Making. *Under Major Revisions* at ACM HEALTH 2026.

If explanations were only useful for verifying or interpreting AI outputs during human–AI decision-making, we wouldn't have seen end-users adapting the saliency maps for uses like those observed in Chapter 5. But this is just one instance. So we set out to identify their emerging uses beyond this one instance in another domain where AI-assisted decision-making is prevalent.

6.1 Introduction

Explainable AI (XAI) has largely been designed and evaluated through the lens of four traditional uses: to foster trust, calibrate reliance, enable algorithmic recourse, and support understanding of model behavior [49, 112, 163, 200, 217, 275, 318, 364]. While these uses of XAI are core to developing safe, transparent, and responsible AI, they constrain the design space for how else AI explanations can be used in human–AI collaborations. Furthermore, mixed results and critiques on XAI's ability to actually achieve those four traditional uses [315, 347, 348] raise the question of whether we are designing XAI with the right uses in mind. For example, fostering trust and calibrating reliance on AI are proving to be more complex constructs that an AI explanation alone cannot influence [297]. Research has captured a variety of factors, such as task difficulty, domain expertise, and consistency of AI performance [56], that contribute to shaping trust and reliance behaviors during human–AI collaboration [92]. Within high-stakes decision-making contexts such as healthcare, factors beyond AI explanations, including clinical trial validation and institutional support, are key to gaining physicians' acceptance of AI-powered clinical decision-support systems (AI-CDSS) [46, 317].

Other critiques of XAI have highlighted XAI's lack of actionability during AI-assisted decision-making [315]. Singh et al. [315] state that an explanation is actionable if it provides information that helps people figure out what actions need to be taken in order to change the AI's prediction. Consequently, a subset of XAI research has focused on designing actionable AI explanations to be used for algorithmic recourse in human–AI decision-making contexts [29, 315, 366]. Beyond algorithmic recourse, emerging uses of XAI in empirical studies remain limited. Discussions in recent literature have alluded to XAI being used as a mediator in human–human interactions [75, 158, 177, 198] or as a tool to facilitate knowledge discovery [326]. Yet HCI work on XAI in human–AI decision-making has underexplored designs that focus on emerging uses.

Despite evidence that AI explanations often fall short of achieving those four traditional uses, research continues to pit different explanation configurations (*i.e.*, method, modality, content, and interactivity [26, 60, 90, 149, 220, 278]) against one another to identify design guidelines that will lead us closer to the "best" XAI configuration for fostering trust [364], enabling algorithmic recourse [347], supporting model understanding [364], or calibrating reliance [150, 200]. This highlights a gap in the literature where little research exists that seeks to design for and understand the emerging uses of XAI in human–AI collaboration. Although these traditional uses are important to the development of responsible AI, they do not reflect the complexity of real workflows, where many tasks are interdependent and extend beyond a single decision point. Furthermore, these uses do not account for the collaborative, communicative, and exploratory needs and challenges that arise within complex workflows, such as clinical workflows. Yet XAI methods offer mechanisms that can fundamentally support these broader uses when designed with them in mind. We argue that designing for emerging uses can help address longstanding challenges across human–AI collaboration and computer-supported cooperative work contexts. For example, in healthcare, AI explanations can support communication grounding in human–human interactions, such as when a physician uses AI to identify and discuss appropriate treatment options or justify clinical reasoning to a patient. However, with the current empirical XAI literature in clinical contexts focusing primarily on traditional uses of XAI and for a single task in isolation [275, 318], these broader needs presented by clinical workflows remain underexplored.

To investigate these critical gaps in the design for and understanding of XAI's emerging uses in a real context, we focus on an AI clinical decision-support tool for Pulmonary Arterial Hypertension (PAH). PAH is a rare and incurable disease that predominantly affects women [78]. This progressive disease demands intricate, shared decision-making across all stages of care, from medical history evaluation to long-term treatment planning and symptom monitoring. Although there are algorithmic tools for PAH that offer mortality risk calculations based on the patient's lab results and demographics to aid in treatment planning [74], these tools are often limited to risk prediction for the current visit without XAI, neglecting nuanced communication and reflective practices critical to patient care. These factors present PAH as a

compelling context for examining emerging uses of XAI across the clinical workflow. Our investigation into these emerging uses is guided by the following questions: (*RQ1*) *How can existing XAI techniques be designed to better enable emerging uses by clinicians treating patients with PAH?* and (*RQ2*) *What emerging uses of XAI do clinicians treating PAH envision across their evolving workflow?*

Recognizing that the design for emerging uses of XAI in clinical workflows remains underexplored, we adopted an iterative prototyping approach with clinical stakeholders as a method of inquiry to narrow down and refine designs of existing XAI methods (Phase 1). We chose prototyping for Phase 1 because prototyping methods have been shown as appropriate ways to design for and understand XAI needs [158, 198]. These co-design sessions informed the design of our high-fidelity prototype in Phase 2, which we used as a design probe with seven PAH physicians to further explore the emerging uses of XAI in clinical decision-making. We use our final prototype as a design probe in Phase 2 because design probes have been previously shown to be useful in eliciting rich data within clinical decision-making [52]. To our knowledge, no prior work has used participatory design to intentionally design for or understand the emerging uses of XAI across the full clinical workflow.

Our study presents four emerging uses of XAI specific to clinical decision-making through a combination of interactive and exploratory XAI designs: (1) collaboratively exploring treatment options with patients, (2) generating hypotheses to identify or reflect on treatment plans, (3) communicating with other stakeholders, and (4) supporting education efforts about the disease. These findings emphasize that limiting XAI designs to traditional uses constrains its design space, limiting the opportunity to identify emerging uses of XAI. Our work contributes to the ongoing discourse on XAI by encouraging researchers to explore how existing XAI designs, metrics of success, and evaluations may limit how XAI is used in human–AI and human–human–AI decision-making.

6.2 Study Context: Pulmonary Arterial Hypertension

We recognize the importance of establishing the study context and mapping the journey of key stakeholders involved in PAH patient care. We present the clinical journey of managing a single patient with PAH, from primary care referral to monitoring treatment outcomes, as shown in Figure 6.1. Our journey map has been validated by our clinical collaborators. We introduce specific challenges with existing PAH risk assessment tools (examples shown in Figure 6.2). This journey serves as a foundation for identifying and designing for emerging uses of XAI explored in our study. Each phase of the journey is detailed below.

Review Stage. After a patient is referred to a PAH care center by their primary care physician, the PAH physician will review the patient's existing medical history.

Exam Stage. After reviewing the patient's medical history, the physician requests

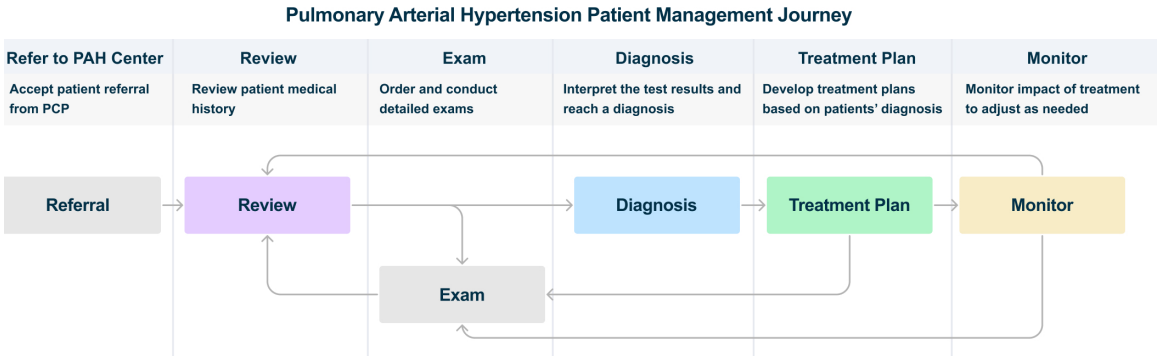


Fig. 6.1: Single PAH patient management journey. Journey validated by PAH physicians.

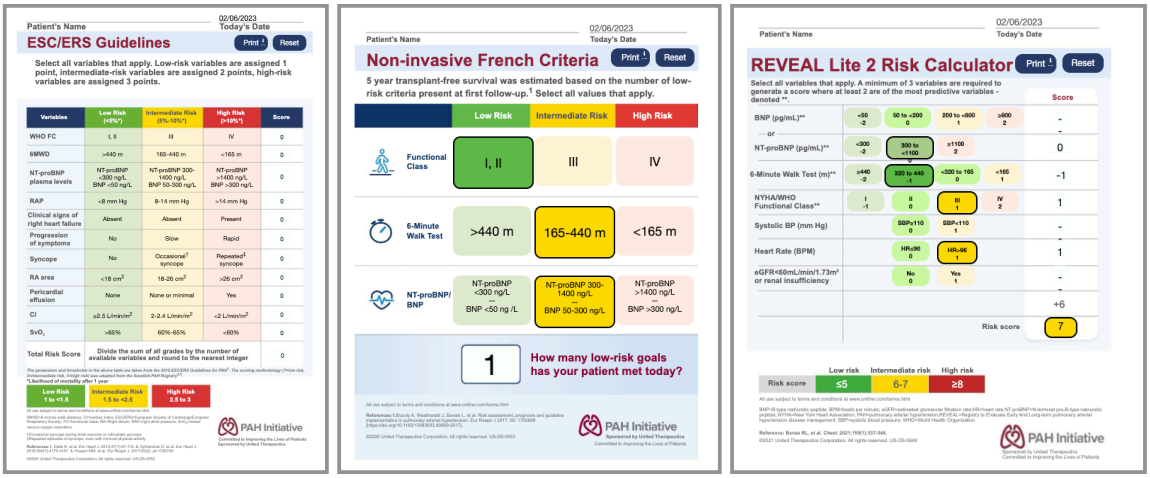


Fig. 6.2: Examples of existing PAH risk calculators. (Left) Interface for the ESC/ERS Guidelines risk assessment tool. (Middle) Non-Invasive French Criteria risk assessment tool. (Right) Reveal Lite 2 risk assessment tool. PAH physicians have referred us to these interfaces, explaining how they aid them in PAH diagnosis and treatment planning. These interfaces are offered by the PAH Initiative [74].

exams if they need additional data before diagnosing the patient. For example, the patient may need an echocardiogram or right heart catheterization so the physician can determine any abnormalities in the heart's function. After the exams are completed, the physician will return to the review stage to consider any additional factors that may influence the diagnosis.

Diagnosis Stage. After the patient completes additional tests, the physician will calculate the patient's risk of mortality at 6 months, 1 year, or 5 years to help identify appropriate treatment interventions. This typically occurs just before the physician's visit with the patient. There are a variety of algorithm-based risk assessment calculators (*e.g.*, REVEAL [22], COMPERA [139], and ESC/ERS Guidelines [148]) to aid in diagnosis and treatment planning. For example, the calculators shown in Figure 6.2 operate on a green, yellow, and red color scale, which helps to easily communicate the patient's risk level. Some institutions have integrated these calculators into their

electronic health record systems, while others have posters of them on the walls in the patient rooms or nothing at all, such as in community clinics.

Treatment Planning Stage. After calculating the patient's risk, the physician will identify an appropriate treatment plan, considering any existing treatment plans and co-morbidities the patient may have. Numerous PAH physicians leverage the ESC/ERS 2022 guidelines [148] when determining an appropriate treatment plan for PAH patients. However, selecting a treatment is a shared decision-making process between the patient and the physician, as certain treatments can require significant lifestyle modifications. This is where being able to visualize different scenarios of the patient's condition can help inform the next steps in treatment planning and encourage reflective practices.

Monitor Stage. After a treatment plan has been established, physicians monitor their patient's reactions to the treatment through follow-up visits, which typically occur every three to six months. The frequency of these visits depends on factors such as the patient's diagnosis, their response to the treatment plan, and external constraints, such as pharmacy or insurance bottlenecks. During these follow-ups, physicians reassess the patient's progress, often recalculating their risk based on new information — whether from exams, unexpected hospitalizations, or regular check-ups — to determine if modifications to the treatment plan are needed.

6.3 Phase 1: Iterative Prototyping of XAI Designs

Our iterative prototyping phase sets out to address our first research question: (*RQ1*) *How can existing XAI techniques be designed to better enable emerging uses by clinicians treating patients with PAH?* During this phase, we engage in low-fidelity prototyping of design concepts for AI explanations while gathering feedback from key stakeholders in PAH.

6.3.1 Methods

6.3.1.1 Inherited Prototype. Before we started iterative prototyping, we inherited an already existing AI-CDSS prototype built by a subset of the clinical team without human–computer interaction expertise (shown in Figure 6.3). We used this inherited prototype as a starting point for our visualizations because it emulates a core interaction and visualization that the clinical team wanted to further explore in the AI-CDSS: manipulating feature values to view and compare different scenarios. This scenario-creation interaction is facilitated by XAI methods, such as counterfactuals, which can be automated [241] or exploratory [366]. However, because this prototype was designed with minimal feedback from clinicians outside the team, we only treat this inherited prototype as a starting point for the possible types of interactions to offer. We describe how the low-fidelity prototype that we developed in Figma across two

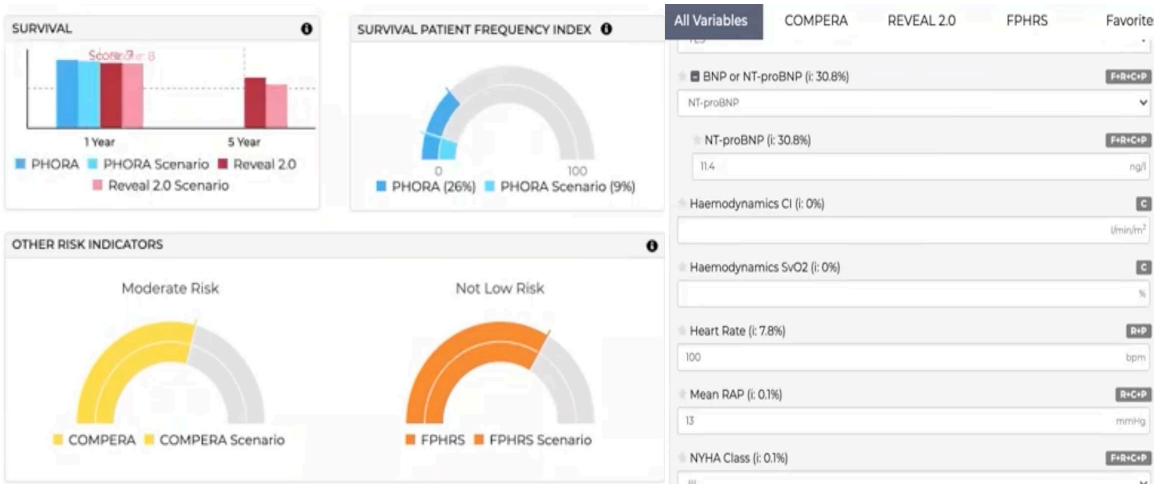


Fig. 6.3: Inherited AI-CDSS design. (Left) Visualizations designed by a subset of the clinical team to visualize how models would predict risk differently when values of the features were modified. (Right) A navigation tab for different risk prediction models and the features for each model. Each feature's value can be overwritten to create a new scenario.

rounds builds on this scenario-creation interaction while expanding to other XAI designs below.

6.3.1.2 Round 1 Prototype. In the first round, we developed an interactive Figma prototype showcasing an initial set of design concepts for explanations: subgroup insights, understanding driving factors of risk, and "what-if" hypothesis generation capabilities. We conducted 30-minute semi-structured interviews with five PAH clinicians (***P1–P5***) using the prototype as a probe to gather feedback on the design concepts and elicit potential emerging uses for AI explanations across the PAH patient care journey.

6.3.1.3 Round 2 Prototype. After talking to those five PAH clinicians, we updated the low-fidelity Figma prototype to reflect the feedback and conducted another set of 30-minute semi-structured interviews with 13 PAH clinicians (***P6–P18***) using the updated prototype as a probe. These sessions further refined our understanding of the roles that explanations could play throughout the PAH care journey, laying the groundwork for the final design of the high-fidelity prototype. As a result, we selected a subset of the design concepts to further develop in the high-fidelity prototype. Descriptions of the design concepts used throughout prototyping are provided in Section 6.3.2.

6.3.1.4 Participant Recruitment. The five participants for the first round of prototyping consisted of four physicians and one family nurse practitioner. The Round 1 interview sessions lasted for an average of 28.56 minutes. The second-round prototype was shown to 12 physicians and one behavioral psychologist for PAH patients (average interview time: 18.58 minutes). Participants were informed of all our IRB-approved

Total Participants	Role Breakdown	Experience Breakdown
18	16 PAH physicians 1 Nurse Practitioner 1 Behavioral Psychologist	1 with 1–5 years 5 with 10–19 years 12 with 20–40 years

Table 6.1: Participant demographics for Phase 1.

study procedures and voluntarily participated in our study after signing a consent form. We aimed to recruit a group of stakeholders with diverse years of experience with PAH care, as shown in Table 6.1. However, we did not intentionally aim to recruit from a diverse set of roles, which is why we only have one nurse practitioner and one behavioral psychologist.

6.3.1.5 Analysis Approach. Our primary analysis method is a thematic analysis. Two thematic analyses were done: one was to capture the design feedback while prototyping the low-fidelity prototype, and the other was to capture any insights mentioned about emerging uses of XAI as our designs evolved. To do the thematic analyses, two authors identified relevant quotes from the interview transcripts and collaboratively grouped the quotes directly in Figma based on recurring ideas and patterns observed. Each group of quotes was assigned a title to represent the theme. The findings from these thematic analyses are presented in Section 6.3.2.

6.3.2 Findings: XAI Design Concepts

Our initial design concepts were informed by existing features in the inherited early AI-CDSS prototype, insights from relevant literature on AI-CDSS systems, and informal discussions with PAH physicians. Based on this foundation, we developed a set of XAI design concepts that have been used for calibrating trust and reliance, but can also afford emerging purposes. These include providing insights into patient subgroups, understanding the factors driving risk predictions, and supporting effective communication with different stakeholders. Each design concept was iterated on after each round; we rationalize final designs with interview quotes from Phase 1 participants below.

6.3.2.1 Subgroup Insights. Recent XAI work in human–AI collaboration has suggested that showing explanations based on model behaviors for different subgroups can improve the appropriate use of the model [49]. Similarly, having an understanding of the percentage of data similar to your instance that was present in the training set [133] or understanding the error boundaries of the AI model [242] can improve the users' appropriate use of the model.

Given that the inherited AI-CDSS designs already offered an instance frequency visualization, and recent literature has shown that it benefits human–AI teams to have a sense of when the model may perform poorly, we include a design communicating an

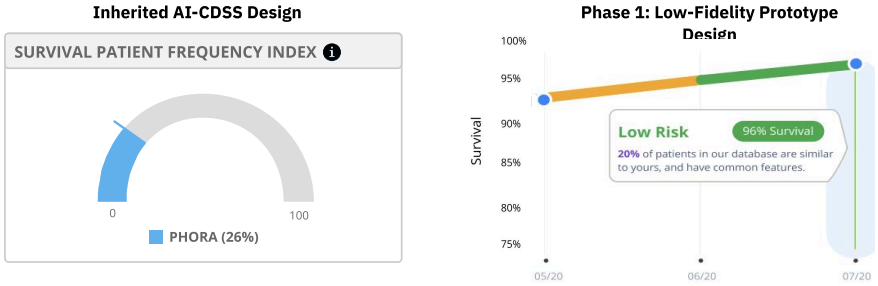


Fig. 6.4: Evolution of the Subgroup Insights Design. (Left) Inherited AI-CDSS design with donut chart visualizations showing the 'survival patient frequency index'. (Right) An iteration of Phase 1 designs showing text callouts for the instance frequency is integrated into the patient's historical risk trend line chart.

aggregate statistic about the instance related to the training data in the low-fidelity prototype. For example, in Figure 6.4, when a risk prediction is made, the AI-CDSS shows an aggregate statistic for the percent of training data that is similar to the patient instance. However, many participants in Phase 1 struggled to see the value this design intended to provide. ***P5*** questioned the utility of the feature, asking, "*What is the value of knowing that 20% of patients in the database are similar to yours*?" and misunderstood its purpose, remarking, "*I actually thought that that was another level of risk*." Similarly, ***P1***, ***P2***, and ***P3*** could not envision how this explanation would add value to their workflow. ***P4***, the nurse practitioner, critiqued the terminology, suggesting that while "frequency" might be misleading, and alternatives like "similarity" or "comparison" would not alleviate the underlying confusion. Given that participants found this design concept more confusing than valuable, we chose not to keep this feature in the Phase 2 high-fidelity prototype design.

6.3.2.2 Understanding Driving Factors of Risk Prediction. Previous work has found empirical evidence of understanding factors that drive predictions to be important in understanding AI and increasing actionability [29]. While the inherited AI-CDSS did not offer this design, we included a feature attribution style design in the low-fidelity prototype. The design is shown as blue dots increasing in quantity based on feature importance, as shown in Figure 6.5 (left) to represent important features driving the prediction. Although the participants from Phase 1 stressed the importance of understanding "*what's the driving variable behind all of this...*" (***P1***), the blue dot visual design in the Phase 1 prototype was confusing to some of the participants.

Initially, our Phase 1 low-fidelity prototype design enabled what-if explanations by overwriting the variables' values in a text box, shown in Figure 6.5 (left), as inspired by the inherited AI-CDSS. However, it was not integrated with feature importance. Interactive what-if explanations combined with feature importance visuals have been shown to help users understand driving risk factors and provide actionable insights [29]. During the interviews for the second round of the low-fidelity prototype in Phase

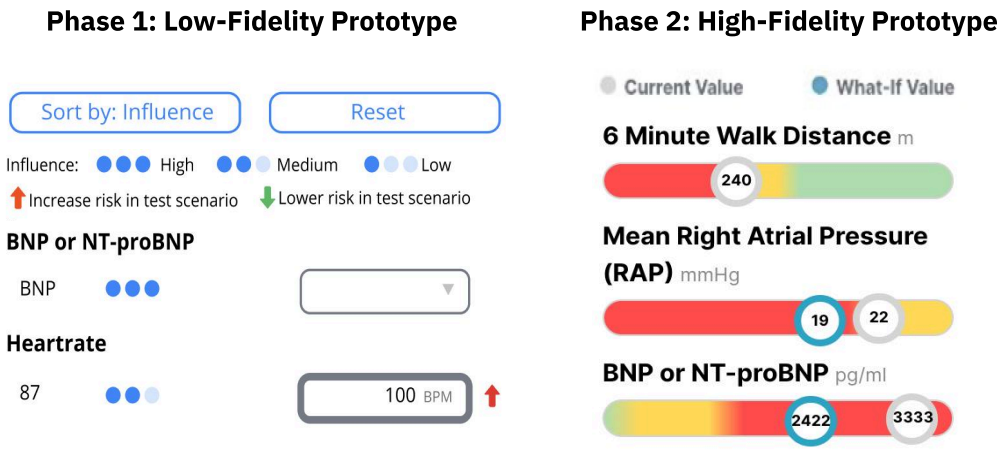


Fig. 6.5: Evolution of Designs for Understanding Driving Factors of Risk Prediction. (Left) One iteration of the Phase 1 low-fidelity prototype design used a three-dot system to represent the influence of a feature on the prediction. (Right) Phase 2 high-fidelity prototype uses partial dependence bars to visualize the impact of features on each other and the prediction simultaneously.

1, ***P6***, the behavioral psychologist, mentioned that "*It might be easier for people to drag a line than to like [input text in the variable field] 'let me try 550'... 'let me try 575'*" and that the current design "*...doesn't tell you how much of that [scenario risk score] is age, and how much of that is walk distance*" when you make complex "what-if" scenarios. As a behavioral psychologist, **P6** viewed the "what-if" explanation scenarios as a way to potentially assist patients in "*therapeutic transitions*" and help explain to them "*why we want you in rehab...*".

To combine information provided by feature importance visualizations with interactive "what-if" explanations, our final design (seen on the right side of Figure 6.5) in the Phase 2 high-fidelity prototype takes inspiration from Krause et al. [185] and Beretta et al. [23]. The XAI design for this concept draws on classic partial dependence plots (PDPs), but modifies them to be interactive and feature-dependent: the bars visualize the effect of different feature values on the predicted risk score as bar toggles are interacted with [185, 308]. This minimizes the feature-independence assumptions of classic PDPs and enables dynamic "what-if" reasoning that existing XAI algorithms do not support. To illustrate, consider a PAH patient presenting with an elevated heart rate (HR) and reduced six-minute walk distance (6MWD), whom the model classifies as high risk. A traditional XAI technique like SHAP would decompose this prediction into additive feature contributions (*e.g.*, indicating that HR contributed +15% and 6MWD contributed +10% to the elevated risk score relative to a baseline). While this answers why the model produced its current prediction, it does not address the question more relevant to clinical action: what would happen to this patient's risk if treatment successfully reduced their HR? Our interactive "what-if" explanations directly support this reasoning. A clinician can adjust the HR to a target

value and immediately observe how the predicted risk changes for all other features simultaneously. This "what-if" interactivity reveals how changing one feature may affect the patient's other current values, according to the model. Our interviews with PAH clinicians suggested this intervention-oriented framing aligns more naturally with clinical decision-making than traditional XAI's feature attribution.

The grey circle (bar toggle) and value in Figure 6.5 (right) represent the patient's current value, while the blue circle (bar toggle) and value represent the user's selection for creating a "what-if" scenario. The color gradients show the categorization of risk predictions for every possible combination of values. Visually, the color gradients are red–yellow–green, as this aligns with the existing culture that PAH physicians have, noting that "*...a lot of us have just gotten used to seeing the risks that way*" (***P1***). These gradients are updated simultaneously for every feature as the what-if values are added and modified.

6.3.2.3 Hypothesis Testing Capabilities. Ludwig and Mullainathan [223] underscore the central role of hypothesis generation throughout the clinical workflow, while Zhang et al. [387] qualitatively observed clinicians' interest in using counterfactuals during treatment planning. However, the previous works are limited in their exploration of XAI designs that are adaptable to the evolving clinical workflow and enable emerging uses. Hypothesis testing was already incorporated in the inherited AI-CDSS design and has similarly been suggested by previous works [29]. ***P8*** motivates PAH clinicians' interest in hypothesis testing, saying, "*...it would be nice to be able to see and to share with patients like, you know, hey, if this happened, this is where your prognosis would go*". ***P6***, the behavioral psychologist, noted the importance of the variable included in making "what-if" scenarios, saying: "*I want to separate out the behavioral aspect of it from the part that they have no control over, which is their aging*". This aligns with related XAI research in AI-CDSS, which states that AI-CDSS should only offer clinical variables in what-if visualizations that can be realistically modified [29]. Based on this, we gathered feedback from PAH clinicians on which variables would make sense to include and in which order they should be presented.

As shown in Figure 6.6, the inherited AI-CDSS featured hypothesis testing through the use of side-by-side bar charts (Figure 6.6, left). However, we chose to use line charts to better incorporate with a historical trends plot for side-by-side comparison in our low-fidelity prototype (Figure 6.6, middle). Interestingly, feedback on the low-fidelity prototype in Phase 1 highlighted more areas for improvement. For instance, ***P2*** found the scenario presentation designs confusing and suggested separating the plots entirely from the original patient data. Inspired by this, we refined the Phase 2 high-fidelity prototype to completely separate scenario generation and exploration from the data and visualizations of the current visit (Figure 6.6, right). This modification ensures clarity by isolating the current risk score from scenario-based risk comparisons, thereby supporting clinicians' needs for clear and actionable insights during hypothesis testing. Inspired by related work on interactive "what-if" explanations [366], we design for multiple comparisons, saving scenarios, and summarizing variables changed.

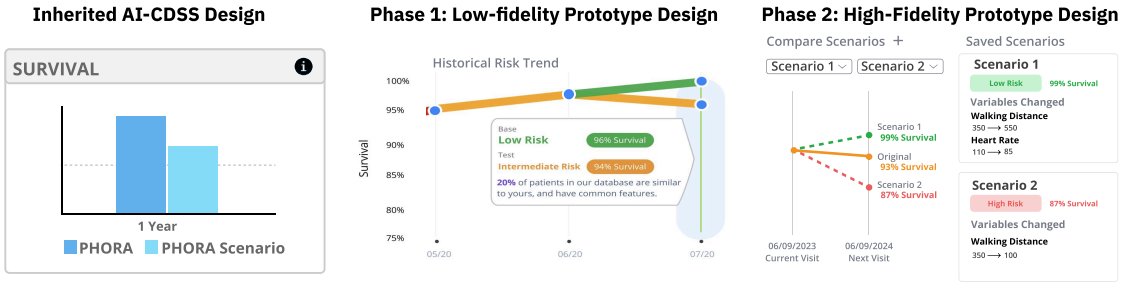


Fig. 6.6: Evolution of the Hypothesis Testing Capabilities Design. (Left) The inherited AI-CDSS design shows bar charts to compare against what-if explanations. (Middle) The Phase 1 low-fidelity prototype design adapted line charts as a historical trend to compare against what-if explanations. (Right) The Phase 2 high-fidelity prototype separates the what-if explanations from the patient's historical trend visualization, enabling emerging uses by offering a comparison of scenarios and a summary of variables modified.

Phase 1 Insight	**Source (P#)**	**Design Decision in Phase 2 Prototype**
Subgroup insights visualization was confusing and unclear on how it supports workflows	**P1–P5**	Removed the subgroup insights entirely from the high-fidelity AI-CDSS prototype
Feature importance can enrich "what-if" explorations	**P6**	Mapped the feature importance on the clinical variables in the "what-if" exploration
Hypothetical and real data can be easily confused	**P2, P4**	Separated scenario exploration into its own pane, distinct from historical risk trends and replaced value overwriting with slider-based controls
XAI has the potential to support patient or trainee communication	**P1, P3, P7, P8**	Connected XAI component to a comparative graph and treatment guidelines

Table 6.2: Mapping Phase 1 insights to design decisions in the Phase 2 AI-CDSS.

6.3.2.4 Summary of Phase 1 Insights. We provide a clear mapping of Phase 1 insights to Phase 2 design decisions in Table 6.2. The four key insights include removing subgroup insights, integrating feature importance into the "what-if" exploration, separating "what-if" exploration from real data, and designing to support human-to-human communication. Collectively, these insights help address *RQ1* by suggesting that clinicians gravitate more towards interactive and exploratory techniques that can more easily integrate with their decision-making processes rather than static, feature attribution methods.

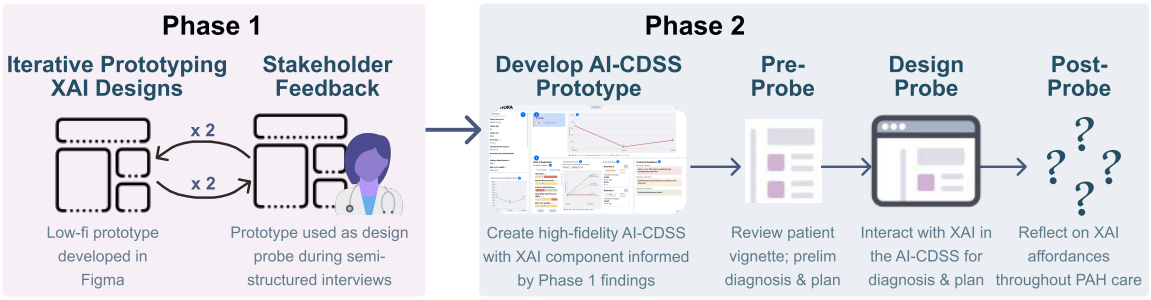


Fig. 6.7: Summary of Phase 1 and Phase 2 methods. Phase 1 consisted of iterative prototyping with feedback from key stakeholders. Informed by developed insights, we develop a high-fidelity AI-CDSS and use it as a design probe for Phase 2. Phase 2 consisted of semi-structured interviews that included a pre-probe, a probe, and a post-probe component.

6.4 Phase 2: Exploring Emerging Uses of XAI

Figure 6.7 summarizes the methods used for Phase 1 and how they inform the methods in Phase 2. After iterating on the XAI design concepts (Phase 1) described above, we used the resulting insights to inform the final high-fidelity prototype for Phase 2. We summarize how insights from Phase 1 inform the design of the high-fidelity prototype that we used as a design probe in the next subsection.

6.4.1 High-Fidelity AI-CDSS Design & Development

Inspired by designs from the inherited AI-CDSS, our prototype in Figure 6.8 presents the (1) patient's clinical variables, (1a) historical trends of each clinical variable, (2) a PAH risk prediction model called PHORA [161], (2a) a historical trend of predicted risk, (3) a "what-if" exploration feature (our final XAI design component), and (3d) a treatment guidelines component. This high-fidelity prototype presents an interactive, conditional explanation system that differs from classical partial dependence plots (PDPs) because it avoids the feature independence assumption. Our findings from Phase 1 confirmed that clinicians found the hypothesis testing functionality useful. They also imagined emerging uses for it, such as augmenting conversations with patients. Our findings also revealed the value that feature importance could have when integrated within "what-if" visualizations to guide comparative insights. Lastly, the findings motivate our choice to exclude the subgroup insights due to confusion and an unclear purpose.

We deployed the high-fidelity prototype online and used it as a design probe in 45-minute semi-structured interview sessions with seven PAH physicians to answer *RQ2*: *What emerging uses of XAI do clinicians treating PAH envision across their evolving workflow?* We detail the interview protocol, recruitment process, and analysis approach in the following subsections.

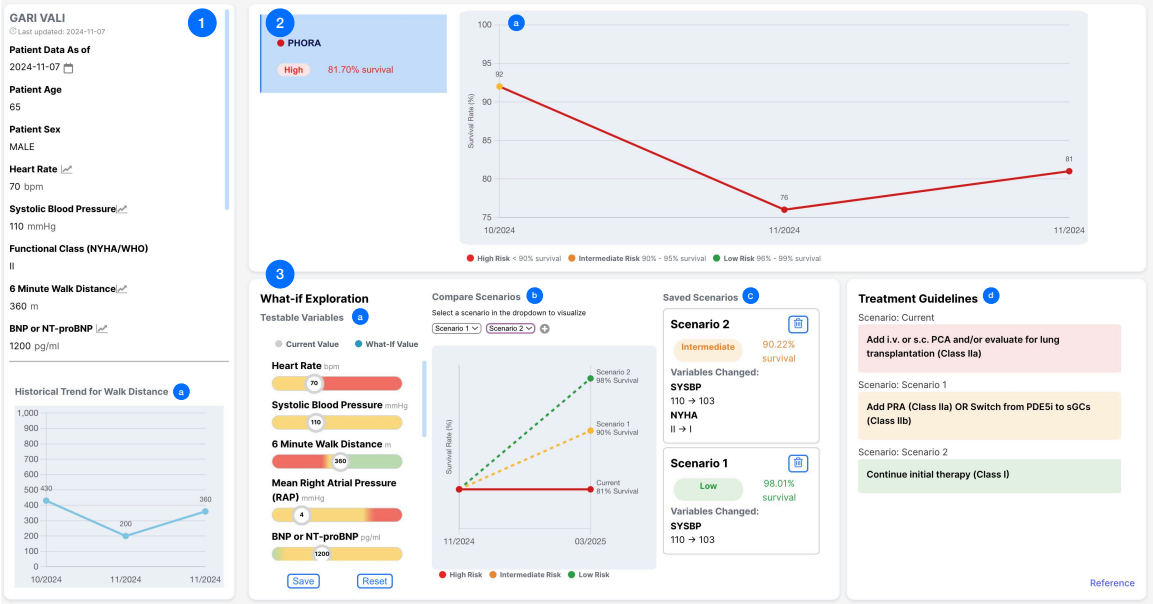


Fig. 6.8: Final Phase 2 high-fidelity AI-CDSS prototype. (1) List of clinical variables used in the risk model. (1a) Historical trends graph for each quantitative clinical variable. (2) Risk model output and (2a) historical trend of predicted risk. (3) Main XAI component informed by Phase 1 findings.

6.4.2 Method

We conduct semi-structured interviews with seven additional PAH physicians (**P19**–**P25**). It is important to note that the semi-structured interview was not intended to quantitatively evaluate the model's performance or physicians' appropriate reliance on the AI. These sessions were designed to be generative and help us better elicit the potential uses of the XAI design throughout the physician's journey. This simultaneously provides insight into whether our XAI designs can help enable those uses. We describe the structure of the interviews below.

6.4.2.1 Pre-probe. Before interacting with the dashboard, participants were presented with a hypothetical clinical scenario in which they diagnosed an anonymous patient from a physician-validated vignette, identified possible treatment plans, and communicated their decisions to the hypothetical patient (the interviewer). The physician-validated vignette was designed to be a complex intermediate-high risk patient case where the PHORA [161] risk assessment model similarly predicted high risk. The pre-probe was designed to get the physician thinking about the patient's case and how they would diagnose the patient before seeing an AI recommendation with explanations.

6.4.2.2 Probe. After participants communicated how they would diagnose and treat the patient, they were directed to the dashboard and provided a brief overview of the different features: model prediction, patient data, historical trends, "what-if"

Total Participants	Role Breakdown	Experience Breakdown
7	7 PAH physicians	3 with 2–6 years 3 with 10–20 years 1 with >30 years

Table 6.3: Participant demographics for Phase 2.

exploration, and treatment guidelines. Participants were then directed to finalize their diagnosis and treatment plan using the dashboard and communicate the diagnosis and treatment plan to the interviewer as if they were the patient.

6.4.2.3 Post-probe. After participants interacted with the dashboard, they were given the opportunity to reflect on the XAI designs and how they might use them in practice. If not mentioned, they were asked explicitly how they could see XAI being used across the different phases of the PAH care journey identified in Figure 6.1.

6.4.3 Participant Recruitment & Analysis Approach

Representing five additional institutions and various experience levels, as shown in Table 6.3, seven PAH physicians (**P19**–**P25**) were recruited with the help of our clinical collaborators. We intentionally aimed to recruit only physicians for this phase as they are the ones reviewing patient history, calculating risk scores, and determining treatment options (the tasks simulated in this phase). Participants were informed of all our IRB-approved study procedures and voluntarily participated in our study after signing a consent form. Interview sessions lasted an average of 43.5 minutes.

Following the same analysis process as described for Phase 1 in Section 6.3.1.5, we conduct a thematic analysis of the relevant quotes from the interviews to capture how participants imagined using the XAI designs to help them diagnose and treat the sample patient. A qualitative approach has been shown to be appropriate to elicit rich insights through the use of semi-structured interviews paired with a design probe [36]. Two authors collaboratively grouped quotes and assigned themes to each group. Any disagreements regarding groupings were discussed to reach a consensus. Themes were then mapped to the PAH journey in order to capture the emerging uses of XAI.

6.5 Findings on Emerging Uses of XAI

This section focuses on scenarios in which PAH clinicians thought they could use XAI, addressing *RQ2*. We include quotes from both the Phase 1 as well as Phase 2 because emerging uses were brought up in both phases. We group emerging uses of XAI based on a single PAH patient management journey in Table 6.4 with a list of which participants from both phases mentioned which uses.

Phase	**XAI Use**	**Phase 1** (***P1–P18***)	**Phase 2** (***P19–25***)
Review	Educating physicians in-training	P2, P13	P19, P21
Diagnosis	Help patients understand the diagnosis	—	P20–22
	Iterative hypothesis exploration	P1, P8–P10, P12, P17	P20–25
Treatment	Celebrate & motivate patients	P2, P6, P8, P13, P16	P20, P21, P25
	Engage patients in treatment planning	P5, P6, P8	P19–P21
	Help physicians gain patients' trust	P15	P20
Monitor	Discussions with care staff	P2	P20, P25
	Reflection on disease management	P10–P12	P19–21, P23

Table 6.4: Participants from both phases are mapped to the emerging use of XAI they imagined across the different phases in a PAH physician's journey of managing a single patient.

6.5.1 XAI Use in the Diagnosis Workflow

While the diagnosis is not the start of the workflow, it is one of the most significant parts because it helps the physician determine what treatment options they will discuss with the patient. We observed two primary ways that physicians across Phase 1 and Phase 2 felt they would use the XAI features during their diagnosis workflow: (1) help patients understand the diagnosis and (2) iteratively explore hypotheses.

6.5.1.1 Help patients understand the diagnosis. Three of the seven physicians from Phase 2 mentioned that they could use the tool to help them communicate better with their patients about their diagnosis and risk score. For example, ***P21*** shared how it can be difficult to have, "*...a higher level abstract conversation about how BNP levels relate to you dying in a year...so maybe this [the XAI designs] would help bridge that gap a little bit*". Similarly, ***P20*** stated that the what-if explanation design can "*...**helps patients understand why I'm so concerned, even if they're unfamiliar with their disease states...***". ***P22*** further imagined how they could use the what-if explanations to help them communicate with their patient:

> "*If I tell them your mean right atrial pressure is 16, and we want to get it down to 9. They're gonna look at you like, what the heck are you talking about? But, if I show them how on that little graph thing ["what-if" comparison] if getting it down to 8 changes it from red to green, that, they understand.*"

However, ***P24***, specializing in pediatric PAH, where the patient usually is not old enough to participate in treatment discussions, raised the concern that the design of the "what-if" explanations component could distract caregivers (*e.g.*, parents) from

focusing on the right information. Aside from patient communication, ***P20*** pointed out that the partial dependence bars design can bring "*...attention to those factors that are playing the biggest role in the ultimate risk category...*", which they imagined would be helpful.

6.5.1.2 Iterative Hypothesis Exploration. Six of the 18 clinicians from Phase 1 commented on how, with improvements, the "what-if" XAI design could be used for iterative hypothesis testing to identify the best treatment options. For example, ***P12*** said "*If a patient is going the wrong direction, you want to see where I can intervene...that's going to make the biggest difference*". Six of the seven physicians from Phase 2 used the "what-if" explanations as a way to validate their hypotheses or determine what needs to change in the patient's regimen and condition so that their risk decreases. For example, ***P24*** started off by experimenting with the 6-minute walk test variable to see how much the patient's risk would decrease. ***P22*** modified the testable variables until they were able to get the patient's risk lower. Their interaction led them to finalize a treatment plan by concluding, "*...now all she needs is a second drug*". ***P23*** specifically had used the XAI tool to help them: (1) "*...figure out when the risk becomes medium*" and (2) "*...[identify] which parameter affected the [risk score] the most...*" by observing how the partial dependence bar plots change. ***P21*** created what-if scenarios with the intention to validate how they expect a treatment would affect the patient's variables: "*making these [what-if scenarios] you know, this is what, realistically, I probably hope for by adding another agent.*"

6.5.2 XAI Use in the Treatment Planning Workflow

The treatment planning workflow follows the diagnosis, where the risk score calculation and medical history strongly influence how physicians determine which treatment path to pursue. One of the five participants from Phase 1 Round 1 and four of the 13 participants from Phase 1 Round 2 imagined how, with improvements, the what-if explanations could be used to have celebratory or motivational discussions with patients, while one participant from Phase 1 Round 1 and three from Round 2 mentioned its potential in being used to engage patients during treatment planning. One participant from Phase 1 also commented on its potential to help physicians gain patients' trust. Four of the seven physicians from Phase 2 imagined how the scenarios could augment their treatment planning workflow by better enabling them to (1) celebrate or motivate the patients, (2) engage the patients in treatment planning discussions, and (3) gain patients' trust.

6.5.2.1 Celebrate or motivate the patients. PAH treatments can be tough regimes, significantly altering the patient's lifestyle. These lifestyle changes can be difficult for patients to navigate and difficult for physicians to explain why and how the patient's condition could improve if the treatment plan is properly followed. For example, ***P8*** felt like with design improvements to the Phase 1 prototype, the what-if explanations

could be used to help them rationalize difficult therapies:

> "*If we were going to introduce a new therapy and said, you know, your BNP has been climbing, it's looking worse. Let's say we started this new therapy. And our goal is that it'll do X, Y, and Z. Now we can look and see where your risk would be if we made our goals. I can see that meaning a lot to patients too. Just getting them to buy in on the therapy. Sometimes they're tough therapies.*"

Similarly, ***P16*** imagined how an improved version of the "what-if" explanations could help them frame their conversation with patients: "*I think [it] would hold a lot of weight with a patient like saying, you know, if we can change factor x by this, look what it does to your outcome, and again,* ***just helping frame that conversation***". ***P20*** validated the design updates to enable this emerging use by similarly imagining how the "what-if" explanations component could be used to help them "*...****emphasize to them [patients] that we're unlikely to get where we need to go if we just stay where we are right now***".

As important as it is to motivate patients to adhere to difficult treatment plans, it is also important to celebrate their progress. ***P6*** (the behavioral psychologist) and ***P13*** imagined using "what-if" explanation visualizations to "*...reinforce, to encourage them that they're improving. Keep taking your meds...*" (***P13***). For example, ***P20*** imagined seeing patients that were doing really well and showing them the partial dependence bars' "*...trends going the other way and the colors moving more towards green [as] something they can celebrate*" instead of "*...telling them a whole bunch of mumbo jumbo about echoes and so forth*". ***P25*** similarly felt like this feature could be used by physicians to "*...help them [patients] be able to gauge how well they're doing and* ***reinforce positive behaviors***".

6.5.2.2 Engage patients in treatment planning. Traditionally, physicians would converse with patients without using visual aids when discussing treatment plans. However, integrating visual and data-driven tools into these discussions has the potential to better engage patients in the decision-making process [247, 248]. One participant from Phase 1 Round 1 and two from Round 2 mentioned this as a potential use. ***P6***, the behavioral psychologist, imagined how they could use the "what-if" explanations component *"'...to help people figure out ways to be more physically active"*. ***P8*** further suggests how the what-if explanation design could be used to "*...shape a discussion about treatment...we can dive into the details of how this would come about. And then this* ***may help explain why I'm not likely to add more medication onto your regimen at this time...***". Three of the seven physicians from Phase 2 validated that the modified designs for the "what-if" explanation functionality can indeed empower them to engage patients in the treatment planning process. For example, ***P20*** said "***We need ways to include patients in their own decision-making using data. And I think that this does that very, very well.***" ***P19*** imagined how if they could "*...pull this up in the exam room, this is more information to show the patient or talk to the patient about how, what therapies we feel, at the current time are beneficial, unnecessary*".

P21 imagined how this could aid in difficult conversations: "*I often run into like, you know, patients not doing well on like oral therapy. And we're having the conversation about this. I think this could help there...*".

6.5.2.3 Help the physicians gain patients' trust. Physicians need to build a strong relationship with their patients so that the patients trust and value their impressions and treatment suggestions. P15 felt like the "what-if" explanations component "***...gives a visual justification***" to their treatment decisions. However, ***P20*** brought up how it can be difficult initially to engage with new patients because "*They don't know my experience level*", but having "*...concrete examples of where we are and where we would like to be...could engender confidence so early on in a relationship...so that they don't think I'm just pulling this stuff out of nowhere.*"

6.5.3 XAI Uses in other parts of the workflow

While physicians primarily calculate risk scores during the diagnosis and treatment planning phases of their workflow, there are instances where calculating a risk score during the review and monitoring phases is necessary or beneficial. Only four participants across all phases mentioned using the "what-if" explanations component to support the education of physicians-in-training during the review phase, and nine participants mentioned emerging uses of the "what-if" explanations component during the monitoring phase.

6.5.3.1 Educating Physicians in-training. Medicine strongly depends on its apprenticeship model throughout education and training. Four of the 25 participants across the two phases mentioned using the "what-if" explanations component to help them educate and train their residents and fellows. This can occur when physicians review a patient's case with their interns and residents. For example, ***P2*** imagined how they could use this "*...for teaching and sort of education purposes, modifying the variables and...looking to see how that changes the risk*". ***P13*** similarly imagined a scenario where they'd, "*...play with it when I teach others about risk stratification. And this is a good teaching tool even outside of clinical care*". Beyond using it to help train others about PAH risk score calculations, ***P21*** imagined using the "what-if" explanations while teaching their residents how to select and modify treatment plans:

> "*[I would] take them through like, okay, this is the risk score right now before we're even starting therapy. We're gonna start X, Y plus Z. This is what we're hoping for when we see them in the clinic in 2 months. And then I'll do that What-if tool where you could see this is where we realistically could see that in a few months...*".

6.5.3.2 Reflection on Disease Management. An important process of PAH care is the monitoring phase, where physicians reflect on their patients' treatment plans and decide if they need to make any changes. Three participants from Phase 1 Round 2 and four of the seven physicians from Phase 2 saw the "what-if" explanations component's

potential to encourage and support reflection on disease management. For example, ***P22*** said while they are reviewing a patient's chart, the CDSS could bring to their attention that, "*...maybe my gestalt is not as good as I thought...maybe I need to get more aggressive with this patient's treatment*". ***P19*** said:

> "*I don't always have the luxury of time when I'm in the exam room, and sometimes data is missing, right? If there's an echo done outside the institution, that may take time to get that report back. Or maybe there's a functional class adjustment...So, you know, during my admin time, I could potentially see using this calculator when I get additional data*".

6.5.3.3 Discussions with Care Staff. Physicians frequently interact with clinical staff, who also interact with the patient. Three participants out of the 25 imagined how the "what-if" explanations component could aid in discussions with other stakeholders. For example, ***P2*** brought up how, with improvement, they could use the XAI feature to present to their colleagues how they can make a patient's condition better: "*...we all work with our palliative care colleagues a lot and to be able to graphically show our palliative care colleagues like this patient is getting sicker, and I can make them better...*". ***P25*** confirms that the new designs can enable such an application, imagining that "*...it would probably be good for them [nurses] to see how their [patients] risk is changed and why it's changed...*". ***P20*** said they could imagine how this could be valuable in situations when they talk with their nurses or get a call from a referring provider about their patient's condition.

6.5.4 Perspectives on Trusting & Relying on the AI-CDSS

As previous research has already uncovered, calibrating clinicians' trust in AI is not as simple as just showing an explanation to validate a prediction [46, 317]. While our study did not have trust and reliance as a core research goal, we still probed the physicians during Phase 2 for their perspectives on trusting and relying on this AI-CDSS for diagnosis and treatment planning. While we did not get noteworthy reactions from five of the seven physicians, two physicians had very interesting perspectives. ***P21*** based their perceived trust on the fact that the tool was in line with what they are used to doing. Aside from ***P20*** mentioning that trust for them comes over time, they also mentioned how they are basing their trust on the AI-CDSS due to the fact that they knew and admired some of the physicians who developed the model that was in the AI-CDSS:

> "*trust for me comes over time...this is the first time I'm playing around with this, so I don't know If I can say I trust it. It seems to make a lot of sense, and I know the people behind the creation of this...so I have maybe an insight that many people don't...but I know [Physician 1] and [Physician 2] well, and you know I've admired their work. So, yeah, there's trust based on kind of me knowing who all is behind this...*"

6.6 Discussion & Limitations

We use iterative prototyping to co-design XAI capabilities that align with clinicians' collaborative, communicative, and exploratory needs. We implement those designs in a high-fidelity prototype and use it to probe for emerging uses of XAI while simultaneously identifying designs that best enable those emerging uses throughout the evolving Pulmonary Arterial Hypertension (PAH) care journey of a single patient. Our findings demonstrate that by combining interactive and dynamic partial dependence bar plots with "what-if" explanation generation and exploration capabilities, the uses of XAI in an AI-CDSS can extend well beyond fostering trust, calibrating reliance, enabling algorithmic recourse, and supporting model understanding. Throughout the PAH care journey, PAH clinicians envisioned how XAI can enhance clinical conversations, facilitate planning and reflective practices, and support educational efforts. We discuss the implications of designing XAI for emerging uses in clinical decision-making by drawing from theories and literature on sensemaking [272, 293], grounding [72], clinical reasoning [258], and design [89]. We further situate our findings within related work to ground these implications, and conclude by outlining the limitations of our study design and suggesting future directions to explore.

6.6.1 Designing Process-Oriented XAI to Support Emerging Uses of XAI in Clinical Decision-Making

While many empirical studies tend to pit two XAI methods against one another to find which is the "best" to use (*e.g.*, [150, 364]), our iterative prototyping led us to design for emerging uses of XAI without limits. With our design, we capture how PAH clinicians felt subgroup insights were confusing and how feature importance visualizations integrated into "what-if" explorations with a separate scenario-comparison feature could be used to collaboratively explore treatment options with patients, identify and reflect on treatment plans, communicate with other stakeholders, and support health education of trainees, nurses, and patients. These findings can result from designing process-oriented [389] XAI rather than designing XAI that focuses on explaining "why" an AI behaved a certain way (output-oriented). We discuss these emerging uses in detail and their implications across a series of dimensions below: communication grounding (Section 6.6.1.1); planning, reflecting, and sensemaking (Section 6.6.1.2), and educational interactions (Sections 6.6.1.3). We further connect the emerging uses of XAI to relevant theories and present example outcomes of using XAI in different ways in Table 6.5.

6.6.1.1 ***Designing XAI to Facilitate Grounding in Conversations.*** Our research highlights how clinicians imagined "what-if" explanations' role in their clinical conversations: providing them additional insights and visuals to better communicate with their patients, which is fundamental to patient care [10, 106] and a key determinant of

Table 6.5: Four emerging uses of XAI that expand traditional uses, grounded in theory: multi-stakeholder collaboration, hypothesis exploration and reflection, trust and motivation building, and bridging stakeholder knowledge gaps.

Emerging Use	**Theory**	**Example Outcome**	**Sec.**
Support multi-stakeholder collaboration and coordination	Grounding [72], Design [89]	Stakeholders reach shared understanding and alignment on next steps through grounding processes	6.5.2.2, 6.5.3.3
Enable hypothesis exploration, sensemaking, and reflection	Sensemaking [272, 293], Design [89], Reasoning [258]	User reasons forward by exploring scenarios, identifying patterns and reflecting on decisions rather than trying to judge the correctness of a recommendation by reasoning backward	6.5.1.2, 6.5.3.2
Build trust and motivation between stakeholders	Grounding [72]	Expert builds credibility; stakeholders remain engaged and committed	6.5.2.1, 6.5.2.3
Bridge knowledge gaps across stakeholders	Reasoning [258]	Stakeholders understand complex information appropriate to their expertise level	6.5.1.1, 6.5.3.1

a physician's credibility [33, 310]. We also observed how our "what-if" generation and exploration design in our high-fidelity prototype can give physicians the freedom to generate visual aids to use in celebratory conversations, which is something that Szymanski et al. [336] found to be valued by patients. Beyond celebratory conversations, we observed clinicians' interest in using explanations as a way to motivate patients, corroborating Kim et al. [176]'s findings. The behavioral psychologist who participated in our iterative prototyping sessions approached these designs as a way to connect with the patient, helping them understand on another level why their physician is advising them to pursue a different treatment route, such as rehabilitation therapy. This is similar to what we heard from physicians in phase 2. The nurse practitioner is underrepresented in our discussion, as their comments primarily focused on interface-level design feedback during iterative prototyping rather than on emerging uses. Physicians also expressed interest in using XAI to communicate with other stakeholders, such as nurses and palliative care staff, hinting at XAI's critical role in providing physicians with powerful visuals to better facilitate grounding. Ultimately, these emerging uses, overlapping with findings from previous works [29, 75, 158], represent broader uses that either support multi-stakeholder collaboration and coordination or build trust and motivation among stakeholders (Table 6.5).

As argued by Miller [232], XAI methods should be grounded in social and cognitive psychology. Rohlfing et al. [289] expand upon this with a conceptual framework for explanations as an interactive social process that involves an explainer and explainee. Recent studies have similarly emphasized the importance of drawing from commu-

nication theories to design human-centered XAI [170, 234]. Through the lens of social science and the concept of grounding in communication [72], AI explanations can be seen as a communication guide to help all parties reach a point of mutual understanding or common ground regarding an outcome or concept. In this sense, our findings suggest how our XAI designs can support these grounding processes within clinical decision-making. Corti et al. [75] similarly conclude that XAI may have value in multi-stakeholder conversations across clinical workflows. In contrast to our interactive design to enable this grounding, Spillner et al. [323] propose an XAI method that focuses on explaining the points of uncommon ground between the human and AI. However, the intentional design of XAI to better facilitate grounding introduces interesting design challenges, as it will be necessary to align XAI with human explanation strategies [232, 238].

6.6.1.2 ***Designing Interactions to Facilitate Planning, Reflection, & Sensemaking.*** Another use that we captured throughout our prototyping and interview sessions was being able to explore hypotheses, engage in sensemaking, and reflect on treatment decisions (Table 6.5). We observed how our XAI designs can encourage reflective practices by enabling clinicians to revisit and evaluate their initial line of reasoning through interacting with the scenarios. Similarly, physicians' interactions with the XAI component suggest how it can enable hypothesis exploration and sensemaking behaviors [272, 293]. Related work has also acknowledged that XAI has properties that can help facilitate the sensemaking process users undergo when understanding AI systems [114]. A series of studies have found how XAI can be used to help people plan tasks by providing actionable insights [198], enable hypothesis generation during decision-making [387], and support hypothesis generation and testing during the model development lifecycle [50]. Zhang and Reicherts [389] comment that AI systems can enable these emerging uses when designed with a process-oriented approach or by designing for the system to be appropriated by its users [89]. Expanding from this line of work, we suggest that our process-oriented XAI designs may be able to help clinicians reason forward by enabling them to consider the decision context, rather than trying to make sense of why an AI recommendation fits a given input. Alternatively, Naiseh [246] discuss how the concept of social XAI can encourage users to reflect on their own decisions, potentially reducing biases within human–AI teams.

Ultimately, designing interactions alongside XAI insights that facilitate planning and reflection should involve holistically understanding the information needs and stakeholder conversations that occur before and after a diagnosis is made. For example, our "what-if" explanation design is tied to risk prediction and simultaneously displays recommended treatments from the treatment guidelines [148]. By doing this, the physicians recognized the value our XAI designs could bring during their administrative time, as they review patients' conditions and treatments. These findings underscore the value of explanations that not only support immediate decision-making but also enhance ongoing reflection and reasoning within clinical workflows.

6.6.1.3 ***Designing Mechanisms to Enable Clinical Educational Interactions.*** The clinicians in our study highlighted the value that our interactive "what-if" explanations design could bring to educating others about PAH. For example, they recognized how they could leverage the XAI visuals while teaching trainees about treatment impact, nurses about a patient's PAH condition, and patients about the disease itself. This captures the broader use of XAI for bridging the knowledge gaps between stakeholders, whether that be physician and nurse, physician and trainee, or physician and patient (Table 6.5). As we observed from the feedback and interactions with the design probe, as well as seen in prior literature, XAI has the potential to support skill and knowledge development [198, 246, 326]. On a related thread, Kim et al. [177] similarly discovered end-users' interest in using XAI to help them improve their skills. However, it is important to understand that not all XAI designs may successfully support educational interactions. XAI designs that intend to support the education of medical trainees will need to adapt to pedagogies used within medicine, such as learning through differentials and variations [258]. For example, designs could include mechanisms that allow users to easily compare multiple scenarios to enable clinical reasoning, as shown in our designs. By designing XAI systems that integrate learning and educational capabilities tailored to the reasoning skills needed for decision-making, we can create tools that not only support clinicians in their current workflows but also contribute to the apprenticeship model seen throughout clinical contexts.

6.6.2 XAI for AI-CDSS Ethical Implications & Considerations

As we observed physicians simulate diagnosing and treating a hypothetical patient with our high-fidelity prototype, we observed how the participants were overlooking the accuracy and confidence of the underlying model. This behavior made it apparent how physicians can fall to confirmation bias or be misled by imperfect XAI systems, especially if they are anchoring on an incorrect piece of information or lacking enough expertise to spot inaccuracies. Therefore, it is imperative to focus on mechanisms and designs that minimize the misuse of "what-if" explanations in AI-CDSS. Giving the end-user so much agency over crafting the explanation to use in clinical conversations can be consequential, as explanations have been known to disagree with each other [186] and present misleading or confusing information [167, 239]. This could misinform junior physicians about the severity of their patients' conditions or lead them to misjudge the impact of treatments on their patients' conditions. Similarly, when being used for educational purposes, unidentified XAI inaccuracies can unintentionally de-skill users [327]. When using XAI in conversations with other stakeholders to justify decisions or identify potential next steps, XAI designs must consider the potential inaccuracies and disagreements that may arise. As designers of XAI for high-stakes contexts where users appropriate XAI beyond calibrating trust and reliance, it is becoming an important challenge for the field to not only design for appropriately trusting and relying on the AI predictions but also on the information and insights

being derived by the XAI and the context where those insights are shared.

6.6.3 Limitations & Future Work

We initially narrowed our recruitment to clinicians for this study to explore the clinical validity of the designs and emerging uses, as well as to ensure that the designs were rooted in the clinical workflow. Although our iterative prototyping method actively engaged the PAH clinicians in understanding how XAI can bring value to their workflow, our study raises new questions that need to be addressed in future research. Specifically, we address the lack of including patients and caregivers in our interviews.

6.6.3.1 Lack of Patient Perspective. Our study lacks insight into the evolving needs for and uses of explanations from the patients' and caregivers' perspective. Given that many emerging uses of XAI were situated in contexts involving patients, it would be ideal to continue future design iterations with patients and physicians simultaneously. As such, our ongoing current research is focused on the patients' and caregivers' perspectives to validate the clinicians' perspectives and design a patient-facing tool. Patients' and caregivers' perspectives can help validate the potential of XAI to augment patient-physician communication. It can also help uncover emerging uses of XAI in the context of at-home disease management. For example, Wong et al. [377] found that exploring counterfactuals helps patients better understand their diagnosis, evaluate trade-offs between different treatment options, and actively participate in discussions. Conducting studies with patients can also help identify contexts in which patients are uncomfortable using XAI for an emerging use. For example, Kim et al. [173] found that patients may only value such functionality in low-stakes situations as they fear that they would miss testing the right scenarios in high-risk situations [173]. Carmichael [57] also allude to different perceptions of explanations as the associated risk of the decision changes. Most importantly, future work with patients on XAI designs in AI-CDSS can help ensure that the designs are not "*increasing the distance between clinicians and the patient*" [288]. Such a metric could be used in future studies to validate the designs of XAI in AI-CDSS.

6.6.3.2 Focus on Physicians' Perspectives. A majority of the perspectives we capture for design feedback are from physicians. As such, the emerging themes that we identified reflect a PAH physician's workflow and perspectives. Not only does this limit the generalizability of our findings, but it also limits the potential additional XAI designs and emerging uses that we could have identified. We encourage future works to explore how designs may differ depending on the context and target stakeholder.

6.6.3.3 Lack of Comparison to other XAI Techniques. Another limitation of this work is the lack of a larger qualitative or quantitative comparison to existing XAI techniques. This would further provide evidence that existing feature attribution methods may limit how XAI is used in emerging ways. Future work could consider running a within-

or between-subjects study that compares an existing XAI technique, such as SHAP, to a more forward-looking, interactive technique, such as the one we presented. Such a study design would further validate the designs that enable emerging uses and identify the contexts in which they are most effective. Beyond lacking comparison to other XAI techniques, we do not explore the implications of imperfect XAI. It is necessary to design safeguards to ensure users apply insights derived from XAI appropriately, especially when XAI presents inaccurate or misleading information. Future work is encouraged to further explore XAI designs that investigate safeguarding users from inappropriate reliance on derived insights.

6.7 Conclusion

Inspired by the numerous potential emerging uses of XAI within and outside of clinical decision-making, our work uses iterative prototyping and the use of a design probe to identify designs that can enable emerging uses. Across the two phases, with feedback and insights from 25 PAH clinicians representing 23 different institutions, we present detailed scenarios on how PAH clinicians would utilize interactive XAI visuals that combine information about feature importance and facilitate interactive, forward-looking explorations. Our study highlights numerous emerging uses of XAI in clinical decision-making: from collaboratively exploring treatment options and identifying/reflecting on treatment plans to communicating with stakeholders and supporting health education. Despite still knowing little about how clinicians will utilize XAI in human–human–AI decision-making, we hope this work encourages discussions about new designs for XAI systems and evaluation metrics for XAI's impact on human–human–AI collaborations, extending beyond trust, reliance, understanding, and algorithmic recourse.

Concluding Part II of this thesis, we take away that transparency solutions can be appropriated in unexpected ways and less restrictive XAI designs open the doors to emerging uses: supporting sensemaking, augmenting communication, and facilitating knowledge transfer. These two studies emphasize the importance of designing XAI systems that are adaptable to complex human–human–AI decision-making processes. Part III seeks to formally expand the design space of AI explanations to address key gaps in literature.

Part III

Expanding the Design Space with Process-Oriented XAI

Chapter 7

Exploring the Explanatory Potential of Multimodal Generative AI (GenAI) in Radiology Workflows

---

This chapter is adapted from the following paper:

Morrison, K., Mathur, A., Bradshaw, A., Wartmann, T., Lundi, S., Zandifar, A., Dai, W., Batmanghelich, K., Eslami, M., & Perer, A. (2025, July). Opportunities & Challenges in Designing Text-to-Image Generative AI for Medical Education, Training, & Practice. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. Vol. 8. No. 2. 2025.

A key takeaway from Part II is that output-oriented XAI could be adapted for emerging uses, such as supporting sensemaking, multi-stakeholder communication, and knowledge transfer. However, these emerging uses are not well supported by output-oriented XAI designs, which were initially designed to explain AI outputs for verification, interpretation, or justification. With the continued improvement of GenAI, recent work has highlighted its potential to support emerging uses, motivating us to investigate its explanatory power.

Part III expands the design space for AI explanations in human–AI collaboration by reframing how we use them and interact with them. One analogy that motivates this reframing is based on the design of elevators. Elevators are inherently a black box. The inner workings, such as whether cables are worn or if the motors are rusty, are unknown. Yet, millions of people rely on them every single day in hospitals, residential buildings, and commercial properties. In the 1950s, Otis Elevators developed the first glass elevator, showing grand views of San Diego from a luxury condominium [251]. Today, glass elevators have become extravagant (and pricey) experiences people pay for. Although glass elevators have been built at businesses around the world, it is not because businesses are afraid people would not trust or invest in them if they had a "black box" elevator. Businesses build glass elevators to acknowledge the great value transparency in an elevator ride affords: aesthetics, breathtaking views, and a memorable experience.

7.1 Introduction

The rapid integration of Generative Artificial Intelligence (GenAI) across various workplace and educational settings, including healthcare, has simultaneously sparked excitement and concern. GenAI holds great promise for augmenting workflows, improving productivity, and enhancing training across a wide range of industries [61, 262]. However, the rapid development of domain-specific GenAI without understanding the needs of and challenges faced by domain stakeholders raises the risk of GenAI being misused. Within healthcare, there is a growing amount of literature on new architectures for text-to-image (T2I) GenAI capable of generating complex medical imagery from text prompts [68, 129, 227, 380]. However, much of the research has focused on technical advancements, such as novel architectures, instead of being rooted in the needs and challenges of medical stakeholders.

Traditionally, synthetic image generation has been used to address pivotal technical medical imaging tasks [320], such as image reconstruction or correction [204]. Now that foundation T2I models are becoming widely accessible, the allure of simulating rare conditions, creating patient vignettes, or augmenting training datasets is attracting significant interest [271, 342, 359, 360]. However, much of this enthusiasm is driven by technological evaluations [28, 189, 215, 256] rather than human-centered ones, leading to a *disconnect* from key stakeholders' real needs and challenges. As recently emphasized by human-centered AI researchers, evaluations need to extend beyond algorithmic metrics to understand how the AI can address stakeholders' real needs and the challenges that arise in doing so [318]. Although recent research captures medical stakeholders' perspectives on GenAI's potential applications in practice [385], few works capture their perspectives on the technical challenges and domain-specific risks of generating synthetic medical imagery.

Several studies have identified and evaluated the challenges and risks of domain-agnostic T2I models [30, 32, 245]. However, it is unclear how these challenges and risks translate to domain-specific T2I models. To address these gaps in literature, we follow up a quantitative evaluation of a novel medical T2I model [380] with a human-centered qualitative evaluation. As an interdisciplinary team with expertise in human-computer interaction, machine learning, and radiology, we conducted a human-centered evaluation of the promises, technical challenges, and domain-specific risks of a text-guided model that generates Computed Tomography (CT) images of the lung (text-to-CT Scan GenAI) [380]. CT scans provide radiologists with cross-sectional views of the internal structures of the human body to aid in diagnosis [230].

Our human-centered approach in this work consists of two phases (formative discussions and model exploration) to facilitate discussions about: (**RQ1**) How do key stakeholders imagine using text-to-CT Scan GenAI for medical education, training, and practice? and (**RQ2**) What technical and domain challenges emerge from stakeholders' needs and workflows for medical text-to-CT Scan GenAI? To address these questions, we engaged with a total of eight radiologists, four trainees, and two medical students.

We will refer to this group as key stakeholders throughout the paper.

Our formative discussions included six radiologists and one senior trainee and took place at the end of the quantitative evaluation of the model in the study by [380]. These discussions focused on capturing stakeholders' ideas on how the model might be applied in practice (**RQ1**). This informed the type of tasks we gave participants during the second phase (model exploration), broadened our network for recruiting additional participants, and guided the selection of participant demographics. For the second phase, we recruited two medical students, three radiology trainees, and four radiologists (including two from the formative discussions) to prompt and evaluate the model in a semi-structured interview format. To facilitate easy prompting and output review, participants interacted with the model through a plugin that we developed for an existing open-source medical imaging interface (created by ohi [6]). By interacting with the model, participants imagined scenarios of how it can address real needs and challenges faced throughout radiology workflows (**RQ1**). The exploratory prompting activity also shed light on the technical challenges and potential domain-specific risks that researchers should consider when developing medical T2I models (**RQ2**).

Our work makes three contributions to the AIES and broader Responsible AI (RAI) community. First, we present a text-to-CT Scan GenAI plugin for a popular open-source medical imaging viewer (created by ohi [6]), which researchers can extend to explore additional approaches for human-centered evaluations of the ethical and safety challenges of domain-specific T2I models. Second, we are the first work to leverage a human-centered approach to explore a medical T2I model with medical stakeholders. As a result of this, our paper expands upon Yildirim et al. [385]'s human-centered GenAI work for applications of multimodal AI in radiology by mapping out applications of domain-specific T2I across medical education, radiology training, and practice. Third, we extend existing RAI work and build on del Cerro and Esteve [83]'s position on the implications of using medical T2I GenAI by presenting technical challenges and domain-specific risks that emerged from participants interacting with the model. These challenges and risks range from topics such as confirmation bias, misrepresentation of image findings, and output image resolution. We discuss the implications of developing medical T2I and suggest future research directions to consider exploring.

7.2 Formative Study

This section appears in the Appendix of the published manuscript. Designing human-centered generative AI systems in high-stakes domains can be difficult, especially when the applications of the technology are underexplored. We conducted short formative discussions to lay the groundwork for our design probe by capturing key stakeholders' perspectives on the potential value of text-to-CT Scan GenAI. These sessions were not intended to be comprehensive but to serve as an initial step to guide our design

probe development. We describe the methods used, followed by the resulting needs of stakeholders.

7.2.1 Semi-structured Interviews

We conducted IRB-approved semi-structured formative interviews with seven stakeholders to explore their needs. The sessions were designed to gather an initial understanding of radiologists' current perceptions and expectations of medical text-to-CT Scan GenAI. Although previous works suggest using medical text-to-image GenAI in medical education and clinical practice, we used these interviews to help capture any additional contexts throughout radiology where text-to-CT Scan GenAI could be useful to inform our design probe. Lastly, these sessions helped us finalize the different groups of participants to recruit for the design probe, such as medical students and junior trainees (who we initially did not consider stakeholders but found to play a critical role in the process).

Before the interview session, the radiologists were shown a series of pre-generated CT scans while completing the survey portion of Xu et al. [380]'s study. Similar to Waikel et al. [359], the radiologists were blinded to the exact prompts used to generate the CT scans. During the recorded interview session, they were told that the CT scans were generated and were asked questions about (**1**) their perceptions of text-to-image GenAI being used in radiology and (**2**) the potential areas in their workflows where they think text-to-CT Scan GenAI could be of value. These sessions were designed to be short (around 10 minutes) because radiologists' time is valuable, and we did not intend to use this to comprehensively capture rich data for our research questions. The 10 minutes were enough time to guide us through building a contextually accurate and needs-informed design probe. The interviews were recorded and transcribed to synthesize findings.

Participant Recruitment. Participants were recruited through our clinical collaborators' network. For each session, we provided the participants with an overview of the interview's purpose and obtained their consent for participation. All participants consented to audio recording for data analysis and were provided a form of monetary compensation for their time. We interviewed six attending radiologists and one senior radiology trainee (**F1**–**F7**) across two institutions; demographics are shown in Table **7.1**. Sessions were held on Zoom or Microsoft Teams, based on the participant's preference, and lasted for an average of 6.8 minutes.

Analysis. We used an affinity diagramming approach to analyze the qualitative data. Ideas from the transcripts were pulled out as quotes, and affinity mapping was conducted to group similar ideas across interviews based on common themes. Multiple authors sequentially reviewed the groupings individually, and all the quotes were grouped into primary themes. Sub-themes were collectively generated and agreed upon.

PID	Expertise Level	Experience (years)
F1	Expert	≥20
F2	Senior Trainee	4
F3	Expert	5–10
F4	Expert	5–10
F5	Expert	5–10
F6	Expert	5–10
F7	Expert	≥20

Table 7.1: Participant Demographics for the formative study.

7.2.2 Understanding Perceptions, Needs, & Workflows

As a result of our formative discussions, we identify training (after medical school but before becoming an attending radiologist) as an additional application context beyond medical education and clinical practice. Overall, we identified three needs across education, training, and practice to guide the development of our design probe.

7.2.2.1 ***Need 1: Visualize the Exact Pathology Searched for to Enhance Training.*** We learned from our formative discussions that radiologists rely heavily on a combination of community-driven and paid resources, such as StatDx [103], Google [121], and Radiopaedia [279], to explore different diseases and confirm pathology descriptions when encountering challenging or unfamiliar cases (**F1**, **F4**–**F6**). Radiopaedia is an online community-driven resource for trainees and radiologists to share and discover patient cases, imaging, and disease pathologies. StatDx is an encyclopedia for different pathologies. However, these tools can require time-intensive searches, and even then, they may not provide exactly what is needed for thorough differential diagnoses. For instance, **F2**, a senior trainee, described their current process when reviewing challenging cases:

> "*If this is something that is really abnormal and I don't know, we have some kind of library called StatDx. All residents have that. So* ***I usually look the disease up...but you have to search through it****. It's kind of useful sometimes. You google, for example, pulmonary edema, then it gives you a description and the findings in different modalities and images...But most often we use Google and Radiopaedia...*".

Even with these resources, trainees may still struggle to find exactly what they are looking for due to the lack of publicly accessible reference images. These community-driven or paid resources are ultimately limited by what others *can* publicly share based on patient privacy regulations, making it challenging to explore complex pathologies. As **F2** points out: "*...****there are some rare cases that you cannot see the actual images and you have to learn it from the textbook****...*". These limitations highlight the need for a tool that enables visualization and exploration of specific pathologies. Text-to-CT Scan GenAI offers a powerful solution to fill this gap by allowing trainees

to generate CT scans of the exact combination of pathologies they are searching for. Beyond training, **F2** and **F5** agree that this GenAI has potential in teaching and education. Acknowledging the limitations of existing resources, **F6** said

> "*It [text-to-CT Scan GenAI] helps us for learning about rare diseases that maybe they have not seen during residency or training curriculum...AI can give them [trainees] a chance to look at rare and interesting cases*".

These observations reveal that radiologists are already well accustomed to searching for pathologies on community-driven and paid resources to retrieve CT scans, albeit they can only, at best, find images *most similar* to what they searched for. Text-to-CT Scan GenAI can address this by providing the ability to create CT scans that exactly match the pathology searched for and handle combinations of pathologies in searches that are not publicly accessible. Additionally, the educational potential underscored by these findings suggests the importance of seeking input from junior trainees and medical students to further explore how this technology can enhance learning and training experiences.

7.2.2.2 ***Need 2: Integration of Visual Comparative Analyses in Clinical Practice***. Radiologists beyond training often rely on online resources, such as Google and Radiopaedia, to address challenging cases (**F1**, **F4**–**F6**). These tools are helpful in obtaining pathology descriptions and imaging in multiple modalities. However, they fall short when radiologists need to compare multiple CT scans for differential diagnoses or to validate their impressions of a case. As **F6** described: "*Radiopaedia gives you some structural explanation about a patient below the image and you can find related images*". **F6** imagined how text-to-CT Scan GenAI could enable a more interactive process while reviewing scans and writing up reports:

> "*Maybe we use AI to generate from the radiologist report a new CT scan and* ***compare this CT to the real one and understand our errors*** *and mistakes and which part may be ignored or may be misdiagnosed and help us to correct our reports.*"

This underscores the need for tools that not only provide reference images but also facilitate side-by-side comparisons of multiple CT scans, allowing radiologists to uncover overlooked details and validate their findings. Similarly, **F1** described how AI could help radiologists identify pathologies more effectively, while **F4** highlighted the potential of AI to assist in double-checking their impressions by helping them "*...find some clues that maybe our eyes lose when we want to put all the findings together to make a proper diagnosis*".

These perspectives suggest that text-to-CT Scan GenAI embedded in existing CT scan viewing applications could augment radiologists' clinical practice by enabling more efficient comparisons, uncovering overlooked details, and supporting decision-making in critical scenarios.

7.2.2.3 ***Need 3: Building Familiarity with text-to-CT Scan GenAI Functionalities to Bridge Novelty and Application***. Our formative discussions highlighted the im-

portance of understanding text-to-image GenAI's capabilities when coming up with potential applications for text-to-CT Scan GenAI. Radiologists were generally more familiar with traditional AI than text-to-image generative AI, which presented a barrier to imagining its applications. Initially, participants imagined use cases for text-to-text GenAI (**F2**) or traditional AI (*i.e.*, image classification) (**F3**, **F4**, **F7**) before identifying scenarios relevant to text-to-image GenAI. As **F5** acknowledged: "*I've never been exposed to images generated by AI before*," and struggled to envision practical applications, admitting "*honestly no ideas yet*."

Providing radiologists with the opportunity to experience and experiment with text-to-image GenAI directly could help address this novelty and unfamiliarity before being asked about a very specific, complex application of text-to-image GenAI. Doing so can help bridge the gap between novelty and practical application and alleviate fears about AI's role in radiology.

7.2.3 Formative Discussion Takeaways

We interviewed six radiologists and one senior radiology trainee (F1-F7) for an average of 6.8 minutes after they quantitatively evaluated outputs from the MedSyn model in Xu et al. [380]'s study. Ultimately, the formative discussions provided us with preliminary insights into radiologists' challenges and their perceptions of the value of text-to-CT Scan GenAI in education, training, and practice. However, participants struggled to detail specific scenarios since many had never used T2I GenAI before. We also uncovered participants' challenges with current online resources, such as visualizing the exact pathologies they search for, the importance of visual comparative analyses, and the lack of exposure to T2I GenAI. These challenges inspire us to integrate MedSyn into the medical imaging interface [6].

7.3 Study Design

We describe the protocol for our semi-structured interview and provide technical details on the text-to-CT Scan plugin that participants used during the interview session.

7.3.1 Text-to-CT Scan GenAI Plugin

Our formative discussions with radiologists revealed to us that they frequently search for image findings and pathologies on domain-specific search engines. However, they may not always find exactly what they are searching for. Therefore, we developed a GenAI plugin powered by the text-to-CT Scan GenAI model, MedSyn [380], and integrated it into a popular open-source medical image viewer [6][1]. The MedSyn

[1]Plugin Demo available at https://genai-radiology.web.app/.

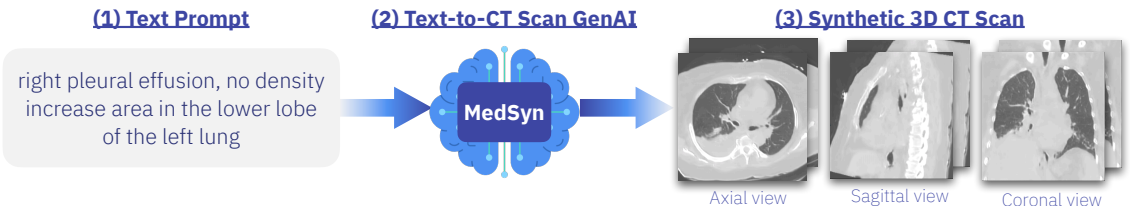


Fig. 7.1: Input/Output of the Text-to-CT Scan GenAI Model [380]. The text-to-CT Scan GenAI model that we use, called MedSyn [380], takes in a text prompt from the user to generate a 3D CT Scan. *Brain icon by Icons8.*

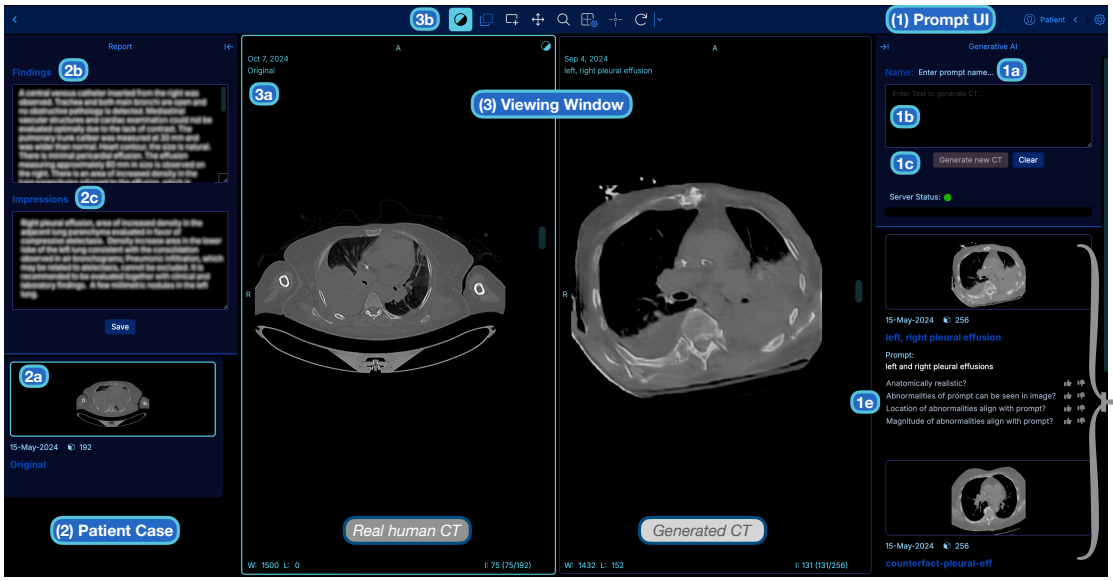


Fig. 7.2: Open-sourced medical imaging interface [6] with our GenAI plugin powered by MedSyn [380]. (2) Patient Case: The first column shows a real case consistent with severe right pleural effusion (sampled from CT-RATE dataset [129, 130, 131]). (3) CT Scan Viewing Window: The second and third columns show a CT scan comparison viewing window, where a human's CT scan (from [129, 130, 131]) is on the left, and a generated CT scan from MedSyn [380] is on the right. (1) Prompt UI: The fourth column is our text-to-CT Scan GenAI plugin, allowing users to generate a CT scan.

model enables users to generate CT scans based on exact and specific pathologies, extending capabilities beyond current search tools. As depicted in Figure 7.1 below, MedSyn takes a user's text prompt and generates a 3D CT Scan consisting of the axial, sagittal, and coronal views of the lungs. Generating the complete CT scan takes between three and four minutes when using an A100 GPU. Our plugin manifests in the interface (shown in Figure 7.2 below) as a text box where users can (1a) name their prompt, (1b) write their prompt, and (1c) generate a CT scan by calling the model API through the click of a button. Users can view a list of previous CT scans that they generated (1d) and have the opportunity to provide feedback on the generation (1e).

7.3.1.1 Front-end Interactions. When using this tool, a radiologist can view a patient's case history (2a), noting their findings (2b) and impressions (2c). This functionality is incorporated to align with how radiologists currently read CT scans in clinical practice. Viewing the patient's CT in the main viewer (3a) gives them access to numerous functionalities (3b) that they are used to having, such as adjusting the contrast to preset window levels (*e.g.*, soft tissue, bone, and lung)[2], adding annotations, comparing images side-by-side, and using a localizer tool. Many of these functionalities either expand upon what the existing radiology search engines offer or leverage findings from our formative interviews. For example, the tool allows users to view and compare a generated CT scan with a real CT scan, as well as annotate them.

7.3.2 Semi-structured Interviews

We conducted 45-minute semi-structured interviews with relevant medical stakeholders to elicit their perspectives on the potential applications and implications of text-to-CT Scan GenAI. These interviews were designed to be exploratory in nature; the goal was not to formally audit or analyze prompting behaviors. The interview sessions consist of four parts, inspired by Cai et al. [52]'s approach with a clinical decision-making design probe, which included a *pre-probe*, *probe*, and *post-probe* phase. We describe each part in the following subsections.

7.3.2.1 Introduction. The interviews started with an introduction and background discussion, where participants shared their experiences and challenges throughout their medical education, training, and practice. This portion was designed to understand the tools, methods, and resources they use to learn about different pathologies and image findings in CT scans.

7.3.2.2 Pre-probe. Participants were asked about their experiences with text-to-image GenAI. For participants who have never used text-to-image GenAI before, we showed them a short conversation made in ChatGPT where we already requested the generation of a medical image from DALL-E[3]. All participants were asked before interacting with the model if they could think of any use cases for text-to-image GenAI being valuable to them throughout their education, training, or practice to prime them for the rest of the interview session. Following this, we showed all participants an animated illustration of how the text-to-CT Scan GenAI model works, and participants were acclimated to the interface through a guided walkthrough video.

7.3.2.3 Probe. During the probe phase, participants generated two to four CT scans, depending on session duration and the depth of discussions. They were guided to use the model to generate CT scans based on different goals informed by the formative interviews, such as validating their impression of a real patient's CT scan loaded in the viewer from CT-RATE [129, 130, 131], learning about a pathology through

[2]Presets were determined by an author, who is a radiologist.
[3]Provided in the Appendix or the published manuscript.

differential diagnosis, and generating something of their choice. Throughout this process, participants provided feedback about the prompts they chose, the quality of the generation, and insights on the model's potential applications in radiology education, training, and practice.

7.3.2.4 Post-probe. Finally, after interacting with the model, participants were asked to reflect on the risks of generating synthetic CT scans in real-world applications and raise any concerns about the generated output.

7.3.3 Participant Recruitment

Participants were recruited through snowball sampling methods from our clinical collaborators' networks. For each session, we provided the participants with an overview of the study's purpose and methods and obtained their consent for participation. All participants consented to audio and screen recording for data analysis and were provided a form of monetary compensation for their time. We interviewed nine stakeholders (R1–R9) across three different institutions: two medical students, three radiology trainees[4], and four attending radiologists (two from the formative study and two new ones); experience distribution is provided in Table 7.2. Sessions were held on Zoom or Teams, based on preference, and lasted for an average of 46.34 minutes.

7.3.4 Analysis

We used an affinity diagramming approach to analyze the qualitative data. First, the transcripts from the interview sessions were verified for accuracy, and corrections were made by listening to the recordings as necessary. Quotes were pulled out from the

PID	Expertise Level	Radiology Experience (years)
R1	Expert	5–10
R2	Trainee	1
R3	Expert	> 20
R4	Expert	5–10
R5	Expert	> 20
R6	Trainee	1
R7	Student	0*
R8	Student	0**
R9	Trainee	1

Table 7.2: Participant Demographics for interview sessions. *Fourth-year med student. **Third-year med student.

[4]Residency programs in the United States come after someone has graduated from medical school, allowing them to gain further practice and experience in a specialization of their choice. In this work, we refer to residents as trainees.

transcripts, and affinity mapping was conducted by multiple authors to group similar quotes across interviews based on common themes. All quotes were grouped into primary themes and sub-themes that aligned with our research questions. Multiple authors grouped the quotes individually, and then the final groups were determined through discussion.

7.4 Findings

We first report on the identified promises of T2I in medical education, training, and practice (**RQ1**). We then follow up on these promises with the risks and challenges of T2I GenAI that stakeholders identified (**RQ2**). Table 7.3 and Table 7.4 show high-level representations of the findings for both research questions. A complete list of the prompts created by each participant, along with their comments about the prompt and the generated output, is provided in the supplemental materials.

7.4.1 Promises of T2I in Medical Education & Training

An application context that was brought up frequently in the formative interviews, as well as our semi-structured interviews, was the potential to use this technology in medical education and training. R7 presents a compelling case for using text-to-CT Scan GenAI in medical education because, "*...it gives you the opportunity to identify and recognize something that is potentially super dangerous, that you then don't have to guess about when you see it in real life*". We see the essence of this quote throughout our conversations with students and trainees. We have grouped participants' insights on specific use cases into three themes regarding the promises of T2I GenAI in medical education and training: (1) supplementing traditional lectures and practice materials, (2) expediting learning through variations, differentials, and rare cases, and (3) empowering trainees during on-call or emergency situations.

7.4.1.1 ***Supplement Lectures & Practice Materials.***. Despite numerous resources out there, "*...there isn't a great place to be like, Hey med students, let's go here and learn chest CT today*" (R8). R8 goes on further describing how they have never done

Phase	**Promises of Text-to-CT Scan GenAI**	Participants
Education & Training	Supplement Lectures & Practice Materials	F2, R1, R3, R7, R8
	Learning through Variations, Differentials, & Rare Cases	F6, R1, R2, R6, R9
	Empower Trainees when On-Call or in Emergency Situations	F3, R2, R7, R8
Practice	Improving Report Impression of Medical Images	F5, R1, R4–R6
	Planning & Communicating with Others	R2, R6, R7
	Visual Memory Support	F6, R4, R5

Table 7.3: (*RQ1*) Participants mapped to which applications of text-to-CT Scan GenAI they imagined across the different phases.

anything in their classes where they get to interact with and compare the differences between abnormal and normal anatomies on CT scans. This could be due to how medical school lectures are designed, showing static images on lecture slides. R3 corroborates this limitation in existing lecture style by describing how they get content for their lecture slides by searching online or through their archives: "*I have this list of typical cases for all of these different lesions...choosing representative images from them...or showing a student the entirety of the scan*". However, R3 went on to envision a scenario where they could use this T2I model to generate specific examples to use on their slides: "*I'm thinking about pre-populating...I'm gonna produce a lecture on this list of obscure topics...And I need specific examples of A, B, C, ... I feed those into your generator, and...I'll come back in an hour....and I'm ready to start producing my lecture.*" R1 provides another perspective, acknowledging how convenient text-to-CT Scan GenAI could be when "*...teaching the residents or medical students,...we just want to show them something, we can just type it, and it will pop up.*"

From the students' perspective, R7 and R8 saw the potential for text-to-CT Scan GenAI to support their studies from the earliest stages of their medical education. The process of medical students shadowing radiology residents is currently difficult as "*It's often been said that radiology even is very hard to shadow in because* ***it's like you're watching someone play a video game you don't know how to play***" (R8). R7 speaks to how this GenAI can "*lessen the learning curve*" as students transition to trainees, exposing them to various pathologies before reviewing real patients. R8 reflected on how the tool could help medical students grasp foundational concepts and prepare for exams: "*...being able to play with it...as you're doing questions, 'Hey, here's what a pneumothorax looks like under tension, under no tension...I think would be invaluable.*"

Despite being able to similarly perform these interactions through Radiopaedia or StatDx, the students instead mentioned relying on other resources, such as UWorld or Amboss, for licensing exam prep.

7.4.1.2 ***Expedite Learning through Variations, Differentials, and Rare Cases.***. As medical students transition into their training program, it becomes increasingly important to get as much exposure as possible to a wide variety of cases. As R1 says, "*the more you see, the better radiologist you will become*". However, despite relying on existing resources populated by the community and their training to see as many diverse images as possible, they are limited by what others *can* share based on patient privacy regulations or what patient cases come through at their hospital. Radiology trainee, R2, highlighted how this GenAI could accelerate their learning by allowing them to study variations of diseases: "*...this would expedite that if I could just be like, 'okay, I need to know the difference between this, this, and this'...I can just generate 5 different variances of those diseases on this tool with the different manifestations of those variants.*"

R2 exercises this during the interview session by prompting the model to generate '`lung cancer`', and then refines their prompt to generate a different variation of lung

cancer. In their review of their generations, they put the two generated CT scans side-by-side. R5 similarly experiments with the granularity of the prompts by adding '`3cm`' to their initial prompt '`irregular speculated lesion...`'.

As many diseases can mimic each other on a CT scan, it is important to learn not only through the variations of how a disease can manifest but also through differentials of how the pathology of one disease is different from another. For example, R6 emphasized how this GenAI could make positive findings more accessible in a low-risk environment: "*Being able to see those positive findings is more educational than commenting on negative findings...sometimes we need to see, like...maybe more than 200 cases to see one positive finding...But if I can be able to generate those positive findings by just giving my impressions...I think that's something that can be very helpful.*"

Beyond variations and differentials, R8 imagined how this GenAI could be applied in day-to-day training scenarios to prepare for region-specific diseases. For example, they described a scenario where generated CT scans could be "*...included in my docket of cases I have to look at for the day and one or two of them might be training, region-specific training*". They similarly mentioned the tool's value in training medical professionals in low-resource environments, such as being a "*...benefit to even Africa...And being able to say, 'Okay, here's what Cardiomegaly looks like with someone who might have tuberculosis.*"'

7.4.1.3 ***Empowering Trainees when On-Call or in Emergency Situations.***. Unlike clinical practice, where attending radiologists rely on their accumulated expertise, training is based on an apprenticeship model where trainees are closely supervised by their seniors. As trainees become more senior, they are increasingly put in situations where they may need to read cases in emergencies or on-call situations, where there is little to no time for doubt or waiting for an expert's approval. These situations demand quick and accurate decision-making, making tools like text-to-CT Scan GenAI an invaluable resource for trainees. Participants highlighted how this tool could serve as a critical support system during high-pressure scenarios. For instance, R2 described how it could aid in trauma cases by enabling them to create a checklist of generated CT scans based on diseases to look out for:

> "*...when you're on trauma, you have a trauma come in...if I had this tool up and I had like 5 different generated pictures of the 5 things I'm checking for when a trauma case comes in...and then when the surgery team is there with me, 'Hey, these are kind of like the things I'm looking out for', while we're searching through the image together*".

R2 expands this idea by imagining how they could use the tool to generate baseline CT scans to make sure they do not miss anything while on-call. R7 imagines how generating images for severe conditions ahead of shifts in the ER could accelerate trainees' decision-making, empowering them to handle emergency situations appropriately and promptly: "*...if you can generate your own images of what different aortic dissections look like,*

*then, when you have a patient that comes into the ER with ripping chest pain, and you take them to CT, and you see, like, 'Oh, sh*t! This is really bad, I need to call someone for help now', as opposed to 'Okay, I haven't seen this before, but this sounds like what I think it could be like. Let me wait and see if I can get CT to read this, and then I have to call my attending', and that all adds time in an emergency situation...time is life.*" R7 then prompted the model to generate an aortic dissection.

7.4.2 Promises of T2I in Clinical Practice

While GenAI can have a high potential for medical education and training, it remains to be seen whether and how it can augment clinical workflows. Clinical practice is slightly different from training because during training, the junior trainees, "*...are reading every case with the attending*" (R6). Our study further addresses *RQ1* by identifying how text-to-CT Scan GenAI can augment attending radiologists in clinical practice, including (**1**) improving report impressions of medical images, (**2**) planning and communicating with others, and (**3**) supplementing the existing resources they rely on when faced with challenging cases.

7.4.2.1 ***Improving Report Impression of Medical Images.***. Interpreting medical images to identify disease progression or abnormalities often depends on differential diagnosis and validating speculations. Participants highlighted how text-to-CT Scan GenAI could play a pivotal role in supporting these tasks by generating CT scans based on their interpretation to visualize hypotheses and refine interpretations. For example, R1 described how the GenAI could assist in determining conditions like cardiomegaly by providing reference images to compare against patient scans. Beyond aiding in diagnosis, the tool could support the refinement of more precise clinical impressions, which participants discovered as they continued generating more CT Scans. R5 described how the generation of reference images could refine the specificity of their observations, "*Instead of just right pleural effusion, I can compare the image and then specify which segment I expect the pleural effusion to be in*". Using this process, R5 reflected on how generating CT scans "*...can be useful to revise my diagnosis to add more details"* and "*...use the generation for differential diagnosis*".

The comparative functionality in the interface allowed R6 to discover that their impression of the original patient case presented to them was not completely correct when they compared the image findings in the generated image to the provided real CT: "*So, I said, smaller effusion, and sometimes people say mild effusion, some say severe effusion. But so this generated AI gives me a very small effusion. But* ***now I'm thinking this might be a mild effusion instead of a small effusion****, which can be very helpful actually...Something like when we use those subjective words and if we have something where somethings objective, could help us a lot actually.*"

Together, these scenarios underscore the potential of text-to-CT Scan GenAI to enhance the accuracy, efficiency, and precision of image interpretation in clinical practice, empowering clinicians while reviewing both routine and challenging cases. However,

it also raises implications for accountability, confirmation bias, and unintentional deskilling.

7.4.2.2 ***Planning & Communicating with Others.***. An interesting use case outside of direct diagnosis or education that was mentioned was in planning and communicating with other stakeholders. For example, R2 described how generating CT scans could be "*...**super helpful in surgical planning**...you can show the patient like, 'hey, this is going to be what you look like post-surgery.*" In a similar vein, R8 describes how they could imagine a scenario where they show a patient how their condition looks compared to a normal condition.

Radiologists are not the only ones who will be interacting with medical imaging throughout the clinical workflow. This could introduce multiple impressions that may disagree about a diagnosis. As such, R5 brings up how "***These generated images can act as additional evidence when I am communicating my diagnosis to other doctors***." R7 motivates R5's point, noting how some surgeons they have shadowed will "*...trust their own reading of the image over whatever the official [radiologist's] read is.*" This is an interesting dynamic in the clinical workflow where generated CT scans could aid in these collaborative workflows, helping radiologists provide validations of their reports by generating CTs that show differentials for other physicians to read.

7.4.2.3 ***Visual Memory Support.***. Radiologists are constantly learning beyond their formal training programs, often needing to recall visual features of pathologies they may not have encountered recently. As R4 describes it, "*I think the more it passes from graduation, yeah, the more you forget.*" As a result, participants highlighted how text-to-CT Scan GenAI could serve them "*by strengthening our visual memory*" (R4). This is particularly valuable for general radiologists working across different specialties where they encounter CT scans of the chest along with other regions of the body. As R5 explained: "*I'm working in a different field of radiology. If I am specialized in one branch, and I see a different CT scan, this can be useful to remember something you haven't seen in a long time.*" R5 goes on to say "*It would be easier to use this tool instead of Radiopaedia when I was a student, I would just type my prompt and see an image much faster than having to search for the right case.* R3 and R4 similarly conclude this could be used as "*...an alternative to looking something up on the Internet.*"

7.4.3 Risks & Challenges with Medical T2I GenAI

As a result of key stakeholders generating CT Scans for pathologies beyond what was shown to be validated by Xu et al. [380], we discovered technical challenges and domain-specific risks, addressing *RQ2*. This section is separated into two parts: (1) technical challenges and (2) domain-specific risks, each with sub-themes shown in Table 7.4. These groupings were identified from a thematic analysis of quotes related to model behavior and implications of model use in practice. For the interpretation

Category	Challenges	Participants
Technical Challenges	Generated CT scan resolution quality	R1, R3, R6, R7, R9
	Transparency and Hallucinations	R3, R4, R9
	Contrast-Enhanced CTs	R3, R4, R9
	Generation Expectations	R3–R5, R9
Domain-Specific Risks	Confirmation Bias & Distilling Incorrect Information	R2, R3, R6, R7, R9
	Over-exposure to synthetic CT scans during training	R3, R5, R7

Table 7.4: *RQ2*: Participants mapped to which technical challenges and domain-specific risks they brought up.

of our findings, technical challenges refer to model development, and domain-specific risks refer to downstream impacts from model use. Such risks and challenges may act as barriers to the adoption of medical T2I GenAI if left unaddressed.

7.4.3.1 ***Technical Challenges.***. Impressively, the model was able to generate pathologies that were not tested in the original study [380]. However, there are four primary technical and AI safety challenges that arose while participants were generating and exploring CT scans: (1) the resolution quality of the CT Scans, (2) the approach to validation and hallucinations, (3) generating contrast-enhanced CTs, and (4) technical expectations of the model generation.

Five participants commented on the resolution of the generated CT scans being of lower quality, even though the model we used was designed to develop high-resolution CTs. While looking through their generated CT scan, R7 found it difficult to see if the pathology they asked for was there "*because of the resolution*". They described the experience as if they were looking at a CT scan on a really bad computer instead of the expensive computers that the radiologists use, while R9 described it looking "*like they moved during*" the CT scan. R6 similarly asked during their session if they had a bad internet connection or if it was just the generation quality. However, R3 clarified that a radiologist may not necessarily require a high-resolution CT scan as long as it offers the required information to support the goal of the generation: "*It's not necessary to fool me into thinking it's a real CT. What you want it to do is* ***provide a sufficient example that I would recognize this on a real CT***."

Second, several participants encountered flaws in the generated output as they were prompting the model on image findings that were not tested for by the quantitative model evaluations. For example, R3 spent time closely auditing the outputs, commenting that the model has "*...thrown all sorts of pathology onto this patient...*" including "*...metastatic bone disease...and a plural recess down here that doesn't connect to anything.*" R4 and R9 similarly identified missing or injected pathologies in their generations. A vast majority of the ML literature refers to this type of model behavior as hallucinations, which come with domain-specific risks. As with any black box AI, it is important that users can see which parts of the prompts are associated with which regions of the generation. This becomes especially critical and challenging

with a 3D-generated output where multiple pathologies are combined within a single prompt. This task, however, is computationally expensive and underexplored for text-to-image GenAI [39].

Third, R3's prompt ("`Necrotizing Pneumonia on contrast-enhanced CT`") asked for a contrast-enhanced CT scan, which is a specific type of CT scan that shows reactions of a fluid given to a patient before the CT scan [230], to probe the model's limitations further. As discovered, the model currently cannot produce contrast-enhanced CT scans, as it was not trained on contrast-enhanced CT scans. R4 and R9 also acknowledged this limitation when they generated something that they expected to be shown on a contrast-enhanced CT. Without discussing with radiologists the importance of contrast-enhanced CTs in accurately determining certain image findings, models may be trained on data that does not represent what radiologists would benefit from. Similar to this, R1 and R6 mention how report impressions can sometimes be subjective (*i.e.*, how severe a disease actually is), which is especially consequential in training models depending on whether the reports are sourced from trainees or experts. R9 mentioned how they struggled to get a certain pathology to be of less severity, and they concluded that "*maybe all the patients you guys had were more like outpatient.*"

Lastly, despite the GenAI's potential value, R3 and R9 acknowledge that the generation time delay is substantial (3 minutes), especially when comparing it to searching for examples on Radiopaedia. R9 said that as residents, they have "*...limited time while you're reading images to do research.*" Beyond generation time, several participants commented on the GenAI's limitations compared to content provided by existing resources, such as Radiopaedia. For example, R5 contrasts the model & interface to how "*Radiopaedia also has some additional notes [about the image finding] which can be helpful for differential diagnosis*." These resources allow them to parse detailed descriptions, aiding in their conclusion about a diagnosis, whereas the GenAI is limited to taking a single prompt at a time to generate a CT scan and does not provide any detailed description in return.

7.4.3.2 ***Domain-Specific Risks***. The development of medical T2I that addresses stakeholders' needs and concerns goes beyond solving technical challenges. We must also consider the complex interplay between user expectations and domain-specific implications. Our findings revealed two domain-specific risks that need to be considered to ensure that T2I GenAI meets the needs of the key stakeholders and aligns with complex workflows: (1) mitigating confirmation bias and the distillation of incorrect information and (2) supporting continuous exposure to real medical images.

First, participants were unsure if and where the requested pathology from the prompt was generated (R2, R6, R7, R9). While reviewing the generations, they would say something like "*Maybe this is what they [AI] wanted*" (R2) or "*I potentially see that dissection*" (R7). This type of behavior with AI, masked as confirmation bias, can be detrimental to a trainee's learning progress. R9 highlighted potential consequences: "*If you're learning and then you get like false information, uh oh, you know, that's*

confusing. ***And you don't know if you're wrong or [the] software is wrong***." In the use case of practice exams, these unprompted pathologies could distract more than support the student (R3). Ultimately, it is possible for students generating CT scans who are unaware of the additional, missing, or misrepresented pathologies to unintentionally learn the wrong things. For example, after R7 generated a CT scan for aortic dissection, they describe a potential issue related to confirmation bias: "***Am I seeing what I'm actually seeing? Or am I seeing what I want to see?***" To this point, R7 emphasizes that research would need to "*...make sure that this is something that's not going to...indirectly, [give] poor quality education.*"

Second, despite the participants acknowledging that the generated CT scans were of lower resolution, they were surprised by the fact that they were looking at a generated CT scan. R7 remarked, "***If you hadn't told me I wouldn't have known [that the CT was AI generated]***." F2 reacted similarly, saying "*Oh wow that's really good because I couldn't tell which one is from which.*" R5 stated "*With the right prompt, the quality of the output is so good that I cannot distinguish that this is AI, if I did not create the prompt myself.*" This in itself raises potentially consequential implications as R5 states that "*Generated images can be useful for experienced doctors who have seen thousands of images, but* ***residents need to develop experience with real patient cases and images***." From a lecturer's perspective, the students want "*...to see real examples with validated pathology*" (R3). This underscores the ongoing theme that radiologists and those in the broader medical field need to continue seeing a plethora of real CT scans throughout their careers, given that the ones they need to interpret are real patients.

7.5 Discussion

Building advanced models, such as the text-to-CT Scan GenAI, often occurs without a deep understanding of key stakeholders' needs and potential safety risks. This lack of clarity raises important questions about misuse, misinformation, and the potential erosion of expert skills, particularly in high-stakes environments such as medicine [162, 262]. Below, we discuss the benefits of our study design and the implications associated with the use of text-to-image GenAI for radiology.

7.5.1 Evaluating Domain-Specific Text-to-Image GenAI

Human-Centered Evaluations of T2I GenAI with Stakeholders. As an outcome of conducting exploratory model prompting with key stakeholders, we observed how each participant in our study was interested in exploring and evaluating a different capability of the model. Unlike the formative interviews, where users could not prompt the model, the exploratory model prompting study design allowed us to uncover numerous technical challenges and domain-specific risks. For example, it helped expose that the thumbs-up/down feedback mechanisms offered for each generated image in the

interface did not address what participants really wanted to audit, which was the ability to generate contrast-enhanced CT scans, the resolution of the images (similar to Brisco et al. [42]'s observation), or the misrepresentation of image findings.

We also learned that CT scan impressions can be subjective, making it important to be concerned about the expertise of the radiologists who wrote the reports for the training data. Using reports by trainees or unverified by multiple radiologists could introduce inconsistencies that can impact the reliability of the model. All of these challenges are difficult to quantitatively capture; however, conducting exploratory human-centered AI evaluations to simultaneously elicit this will be useful to the advancement of RAI.

Lastly, unlike many existing risk taxonomies and evaluations for T2I models, we observe a specific set of risks that the radiologists were primarily concerned about: distilling incorrect information and losing exposure to real CT scans during training. These risks represent unintentional harms that could come about from this technology being widely available in practice. It is interesting that more granular risks or harms, such as various forms of bias, were not acknowledged by the participants. Despite this, it will be critical to identify domain-specific biases that arise with domain-specific T2I models and innovate solutions to mitigate them.

7.5.2 Implications of Text-to-CT Scan GenAI in Medical Education, Training, & Practice

We allude to *RQ1* as we discuss the implications of T2I for medical education, training, and practice. Our participants emphasized the role that online resources and pure experience play when faced with challenging cases and professional growth, as similarly observed by Yildirim et al. [385] and Rassie [281]. El-Ali et al. [102] and Quaia [277] have observed an increase in students' and trainees' performance when provided access to radiology-specific repositories, emphasizing the importance of supplemental resources to traditional lecture materials. Extending these insights, five participants referenced how getting exposure to variations can play an important role in competency within medical imaging-based professions. Generating variations as a design has similarly been explored by [118], highlighting the value that variation plays in sensemaking. However, the design of our GenAI plugin did not easily enable the generation and visualization of variations for a single prompt. This was also difficult due to the generation time for a single CT scan. As Weisz et al. [369] suggest, it is important to design GenAI applications to enable variability and expose differences from a prompt. While this generates variability aimed at meeting the user's needs, we also see how this can expedite education by showing several CT scans with the same pathology in several variations. This would help prepare students to practice identifying serious conditions before dealing with real patients, a benefit that del Cerro and Esteve [83] similarly mention.

Our participants also stressed the importance of being able to understand what real

human anatomy is like by looking at thousands of real CT scans. This opens the door for RAI work to explore various design opportunities for domain-specific T2I models, such as inpainting capabilities. Furthermore, RAI can explore the human-centered needs around explicitly identifying which case in the viewing window is generated and which is real to avoid misuse and abuse of the GenAI capabilities [83].

Considering Collaborative Workflows. Our participants highlighted the collaborative workflows that take place within medicine and how text-to-CT Scan GenAI could provide radiologists an opportunity to provide generations as rationalizations for their reading to enhance their collaboration with surgeons. Yildirim et al. [385] also saw the value of showing similar images as a use case for reassurance between radiologists and other medical professionals. Such an interaction throughout collaborative workflows might require designs that allow radiologists to leave additional notes or engage in an asynchronous discussion with the surgeon about their generations and patient case impressions.

7.5.3 Safety Implications of Medical T2I GenAI

With any powerful text-to-image GenAI model, it is necessary to consider the broader ethical and safety implications. We allude to *RQ2* while discussing design implications for safeguarding against hallucinations and confirmation bias, as well as addressing misuse of the generated CT scans.

Safeguard Against Hallucinations & Confirmation Bias. As we have seen through our own experiences and previous work, hallucinations are prevalent in GenAI systems. This is especially concerning because primarily the more experienced participants were commenting on misrepresentations of the prompt in their generated CT scan. This motivates future works to explore ways to express hallucinations to users, especially to novice users. Along with hallucinations, some participants raised concerns about confirmation bias, which can emerge when working with AI [291], especially among students and junior trainees. While it may be hard to avoid generating hallucinations, we can design mechanisms to safeguard end-users from confirmation bias during their learning experience. Designing mechanisms to minimize confirmation bias is especially important in an educational environment because recent work has shown how people can start reasoning incorrectly when given misleading explanations generated by AI [327]. These designs are especially important to avoid deskilling, slowing professional growth, and minimizing over-reliance. One potential solution that we see is by developing visual explanations that associate concepts from the prompt to regions of the generated image, such as the technical solutions proposed by Chefer et al. [65], Bradshaw et al. [39], and Evirgen et al. [107].

Preventing Misuse and Malpractice. Although this was not an emerging theme from the interviews, it is necessary to anticipate potential misuse and malpractice of synthetic medical images. Koohi-Moghadam and Bae [183] suggest the potential for AI-generated medical images to fabricate clinical results or patients' conditions,

which can lead to inaccurate conclusions or mistreating patients. The misuse and malpractice scenarios could go on, but it is more important to discuss practices to prevent such scenarios from becoming prevalent. This especially becomes important when these models increase competency by being able to do tasks such as inpainting original patient CT scans. Horvitz [145] proposes watermarking as a potential solution to distinguish synthetic from real content, promoting transparency.

7.6 Limitations & Future Work

This work comes with limitations that future work can build upon. First, our study involved a small sample size for both the formative study and semi-structured sessions. However, this sample size choice is restricted by domain stakeholders' availability and respect for their time. Additionally, our participants all had different levels of experience and perspectives about AI, which could have tainted their perspectives on the promises, risks, and challenges that they identified. While we had briefly asked participants about their prior experiences with AI in the pre-probe sessions, we have not done an extensive analysis of that data. Also, our study is limited to a very domain-specific T2I for chest CT scans, which could potentially limit how our conclusions translate to other medical text-to-image GenAI. Although not all participants were chest specialists, we felt that they were still able to successfully identify valid promises, risks, and challenges even beyond chest CTs. For example, two participants were able to see how this text-to-image generation process could generalize to other imaging modalities.

Additionally, we did not conduct a detailed analysis of all the prompts generated by participants, as this was not our primary goal with this study. Furthermore, we have a limited number of prompts/generations from participants, as each generation took between three and four minutes. We did not intend to further evaluate the model's performance or identify heuristics for prompts that work better with this model for this study. These areas remain for future exploration to better understand how key stakeholders would prompt generative models depending on their goals.

Similarly, we did not perform quantitative evaluations of the accuracy of the generated CT scans from the radiologists' perspective during these sessions. Future work should consider including quantitative model auditing activities with stakeholders using familiar medical interfaces and tools to maintain external validity and to ensure that evaluations appropriately target stakeholders' concerns.

7.7 Conclusion

Medical text-to-image GenAI capabilities are increasingly being investigated as models become increasingly competent. Given the lack of human-centered research involving key stakeholders on the promises, risks, and challenges of medical text-to-image GenAI,

our work seeks to bridge this gap by conducting exploratory model prompting sessions with a text-to-CT Scan generative model. Our study highlights the opportunities of text-to-image GenAI in medical education, training, and practice, such as empowering students to learn in low-stakes environments and augmenting radiologists when faced with challenging cases. These promising opportunities expose technical challenges and domain-specific risks of domain-specific T2I models that the RAI community can consider, such as confirmation bias. Overall, our work represents the importance of evaluating T2I models from a human-centered lens, taking domain-specific needs and challenges into consideration. We hope this work encourages RAI researchers to further explore the risks and challenges of domain-specific T2I models.

This chapter ultimately exposes clinical reasoning processes that are core to interpreting medical images. Simultaneously, it highlights how the text-to-image GenAI model can be repurposed as an explanation generator to support those clinical reasoning processes. This leads us to develop a process-oriented XAI system using this model in the next chapter.

Chapter 8

Augmenting Reasoning, Not Validating Outputs: Designing Process-Oriented XAI

This chapter expands on the following workshop paper:

Morrison, K., Mason, E., Chen, M., Lundi, S., Zandifar, A., Kim, D., Eslami, M., & Perer, A. (2026, April). Reframing Explainable Generative AI as a Tool for Thought. In ACM Conference on Human Factors in Computing Systems' Workshop on Tools for Thought (2026).

Morrison, K., Mason, E., Chen, M., Lundi, S., Zandifar, A., Kim, D., Dai, W., Eslami, M., & Perer, A. (2026, April). Augmenting Reasoning, Not Validating Outputs: Designing a Process-Oriented XAI System with Multimodal Generative AI. Manuscript under submission (2026).

Part II described how XAI had emerging uses across various processes in expert decision-making, such as supporting patient–clinician and clinician–clinician shared decision-making. However, in Chapter 7, we observed that the text-to-image model had significant explanatory potential for training and education efforts. Even though some participants in Chapter 7 did allude to supporting clinician–clinician or patient–clinician interactions, the more obvious explanatory potential of multimodal GenAI lies in scaffolding clinical reasoning during training. This marks a shift in our focus from XAI for decision-making in human–AI collaboration to designing process-oriented XAI as a form of human–AI collaboration to augment reasoning processes.

8.1 Introduction

Explainable AI (XAI) systems in human–AI collaboration have primarily been designed to help decision-makers use AI systems safely and effectively by serving as tools for verifying or interpreting AI outputs (*i.e.*, output-oriented XAI). As shown in Figure 8.1, this interaction paradigm consists of an AI system producing an output (*e.g.*, a

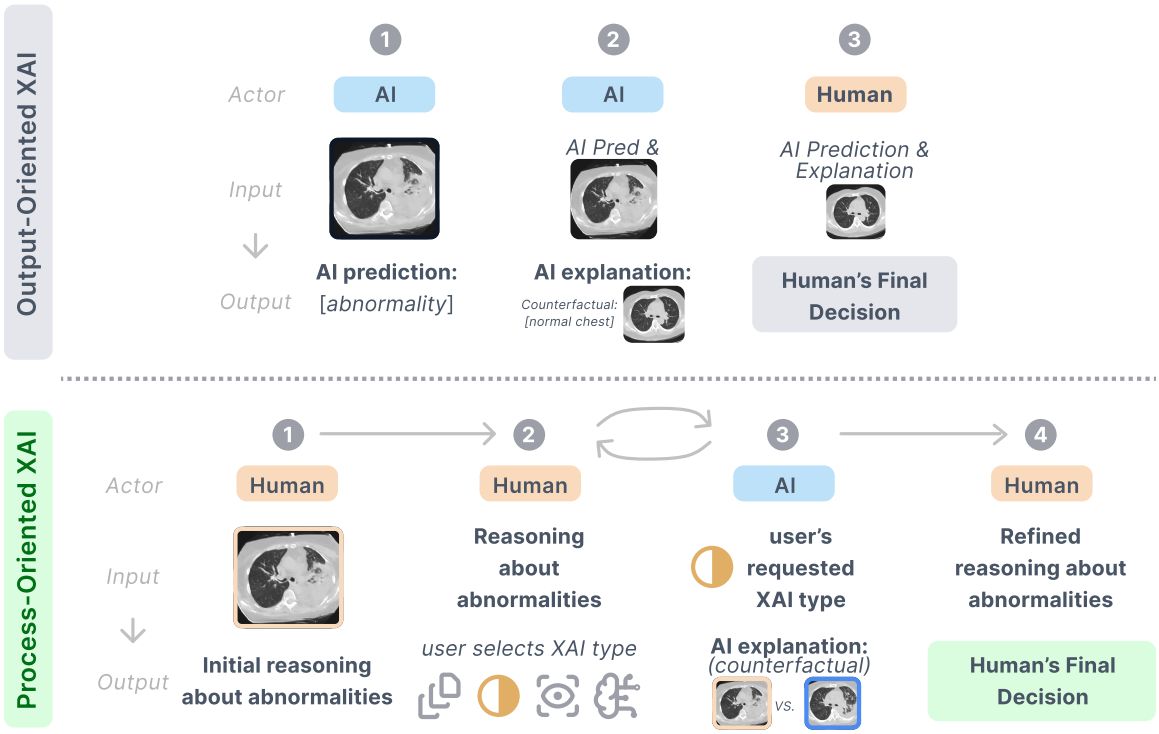


Fig. 8.1: Output-oriented XAI is framed as a way to help humans decide whether to accept, modify, or reject an AI's recommendation or generated artifact. Process-oriented XAI, in contrast, considers a workflow in which humans first reason without AI support, then generate AI explanations and use them to refine or shape their reasoning before finalizing their decision.

classification) and a corresponding AI explanation to justify its output for a human to use in making a decision. In this paradigm, the human is positioned to be a validator of AI outputs, which may unintentionally distract decision-makers from engaging in their own reasoning about the underlying data, context, and decision. Zhang and Reicherts [389] argue that this output-oriented framing forces users to reason backward, rationalizing if and why an AI recommendation is reliable based on its explanation. This interaction paradigm has been the primary focus when designing XAI systems, as the literature has primarily focused on output-oriented goals when evaluating XAI's impact in human–AI systems. For example, one canonical goal has been to calibrate reliance on AI outputs [49, 150, 200]. Other goals include supporting model understanding [49] and fostering trust in AI systems [200, 364].

Countless empirical studies have yielded mixed results regarding whether XAI fulfills those goals [348]. Even interventions designed to promote deeper cognitive engagement with AI explanations have shown limited success within this output-oriented framing [45, 115], motivating the need to explore new interaction paradigms with XAI. Researchers have recently leveraged GenAI to explain the outputs of another AI system [137, 249] or to make AI explanations themselves more human–centered [254]. These techniques are derived from an emerging field called explainable generative AI

(GenXAI) [302]. However, output-oriented interactions with GenXAI have also faced limitations in human–AI collaborations [239, 327].

As empirical evidence on addressing output-oriented goals (*e.g.*, calibrating reliance, supporting model understanding, and fostering trust in AI systems) with XAI remains inconsistent, the limitations of XAI as a tool to help users verify, interpret, or justify AI outputs become clear. These limitations reflect broader challenges with output-oriented interactions in AI systems. For example, recent organized efforts have called for departing from output-oriented interaction paradigms in AI systems by adopting a new design practice, Tools for Thought [339], with the goal of providing process-oriented support [389] that protects and augments human cognition. However, this reframing leaves the role of XAI in human–AI systems unclear, as XAI has traditionally been used to help humans interpret or verify AI outputs during AI-assisted decision-making.

Recent work has begun exploring XAI's role in human–AI collaboration beyond output-oriented interactions by introducing new frameworks, such as Evaluative AI [233], and by investigating how decision-makers appropriate XAI in practice [239]. However, system design for emerging interaction paradigms with XAI in human–AI collaboration remains underexplored.

Using cognitive psychology of explanations [171, 222], concepts from Tools for Thought [339], and Alan Dix's emphasis on appropriation-aware design [89] as a research lens, we introduce a new interaction paradigm for XAI in human–AI collaboration called process-oriented XAI. Process-oriented XAI, as depicted in Figure 8.1, reframes XAI as a tool that supports humans' workflow processes, rather than as a tool to help users verify AI outputs. This reframing is motivated by the fact that many XAI techniques draw on or emulate aspects of human explanation and reasoning patterns [232] (*i.e.*, counterfactual evidence, example-based reasoning, and localization). For example, in this interaction paradigm, the human considers the data on their own and what support they may need while reasoning about that data. This informs their choice to request an AI explanation from a specific XAI tool, such as a counterfactual explanation, to augment their reasoning and help them reach a final decision. We instantiate this interaction paradigm in a clinical reasoning tool for medical image interpretation using multimodal GenXAI to understand (**RQ1**) *How do people interact with process-oriented XAI, and how does that shape their reasoning?* (**RQ2**) *What are the challenges and benefits of using GenXAI to develop process-oriented interactions?*

To address these research questions, we situate our study in the context of medical image interpretation. We chose the task of interpreting medical images because (1) it remains a core focus across the machine learning (*e.g.*, [309]) and human-computer interaction communities (*e.g.*, [240, 385]), (2) it allows us to explore multiple modalities of explanations, ranging from visual to natural language, and (3) training in this domain ranges from shadowing experiences to "learning by doing" experiences [240]. We draw on findings from the literature (*e.g.*, [240, 385]) to inform a set of design goals for building a process-oriented XAI system that leverages a text-to-image medical imaging model developed by Xu et al. [380]. Then, we conduct a think-aloud study

paired with a semi-structured interview with ten clinicians-in-training to address our questions.

Overall, our paper contributes a conceptualization of a new interaction paradigm with XAI techniques, an instantiation of this paradigm in an interactive tool for medical image interpretation using GenXAI techniques, and an empirical study with ten clinicians-in-training. To address RQ1, we observe how five senior medical students and five junior radiology residents use our tool to interpret two different patient vignettes. We observe that participants used the same GenXAI Tools differently depending on their reasoning needs, such as needing a starting point, help refining their impression, or a second opinion. We also analyze how the tool appeared to shape participants' reasoning processes through a combination of behavioral observations and participants' reflections during the interview. Specifically, we capture how the GenXAI Tools can misguide, refine, or help expand reasoning. Lastly, we address RQ2 by synthesizing participants' reflections on the benefits and challenges of GenXAI for process-oriented support.

8.2 Process-Oriented XAI System

We instantiate the process-oriented XAI paradigm in a high-fidelity clinical reasoning support tool that leverages multimodal GenXAI.

8.2.1 Design Goals

Informed by literature on human–AI collaboration, we identify five design goals for process-oriented XAI. To support our medical imaging case study, we situate our design goals within radiologists' workflows and reasoning processes. However, these design goals could be adapted and applied more broadly.

DG1: Eliminate AI Predictions to Position Humans as Active Thinkers. Essential to the process-oriented XAI interaction paradigm and informed by recent work that explores designs that withhold the AI's recommendation (*e.g.*, [45, 339, 389]), AI should not make any predictions on the underlying data (*e.g.*, a patient's CT scan). Therefore, in our process-oriented XAI system, there is no AI-generated impression or classification for the data (*i.e.*, the patient's CT scan). This design choice inherently changes the role of the AI explanations, no longer making them output-oriented, since there is no AI output to interpret or verify. This design goal is demonstrated in step 1 for process-oriented XAI in Figure 8.1.

DG2: Align with Workflows to Augment Existing Processes. Another core aspect of process-oriented XAI is that users' interactions with XAI align with their existing workflow processes rather than introducing new processes to the workflow. This is inspired by empirical evidence from Buçinca et al. [44], showing the value of aligning AI explanations with gaps in human reasoning. As such, the XAI techniques in our

process-oriented XAI system should translate to reasoning processes, such as those involved in medical image interpretation [240, 385].

DG3: Enable User Agency by Prioritizing On-Demand Interactions. Output-oriented XAI studies have manipulated how users interact with AI explanations by either offering them on demand or immediately [59, 266]. In process-oriented XAI, offering explanations immediately would undermine the goal to support meaningful, user-driven interactions. Zhang and Reicherts [389] also discusses "timing of support" as an important factor when designing process-oriented AI. Therefore, the third design goal is that the user must decide when and how to engage with XAI throughout their reasoning process. This design is reflected in step 2 of process-oriented XAI in Figure 8.1.

DG4: Foster Self-reflection to Help Users Externalize Reasoning. Informed by design goals discussed by Zhang and Reicherts [389], process-oriented XAI interactions should foster self-reflection by enabling users to externalize their reasoning as they request explanations. By offering different AI explanation types, it forces the user to reflect on their own reasoning and which explanation will help augment it. This design goal is captured as the output of step 2 and input of step 3 in Figure 8.1.

DG5: Support Navigation of Imperfect AI Explanations. Explainable AI techniques have been shown to be imperfect [43, 186, 239, 292]. Another design goal suggested by Zhang and Reicherts [389] for process-oriented AI systems is the introduction of friction. In the context of imperfect AI explanations, rather than trying to compensate for these imperfections, we embrace them as a form of friction. Previous work has similarly leveraged AI's imperfections to support ideation [219] and richer conversations [127]. However, these imperfections can be misleading or unproductive. Thus, process-oriented XAI designs should support users in effectively and safely navigating these imperfections.

8.2.2 System Overview

We leverage an existing open-source medical imaging interface developed by the Open Health for Imaging Foundation (OHIF) [392] to create a high-fidelity prototype (depicted in Figure 8.2). We use this interface because it already provides several tools and capabilities that radiologists have when reviewing medical images. However, we adapted portions of the original OHIF interface design to better suit our think-aloud study protocol. For example, we provide a panel on the left-hand side of the interface (Figure 8.2.1) to showcase the current patient vignette that the participant is to review.

Figure 8.2.2 shows the medical imaging viewport, where the patient vignette CT scan (A) and any AI-generated CT scans from one of the GenXAI techniques (B) can be reviewed. Inherited from OHIF, this viewport component is flexible and easily allows users to review more than nine CT scans at a time if desired.

Lastly, we developed the right-hand side panel, shown in Figure 8.2.3, to feature the four process-oriented XAI techniques. To avoid participants getting confused

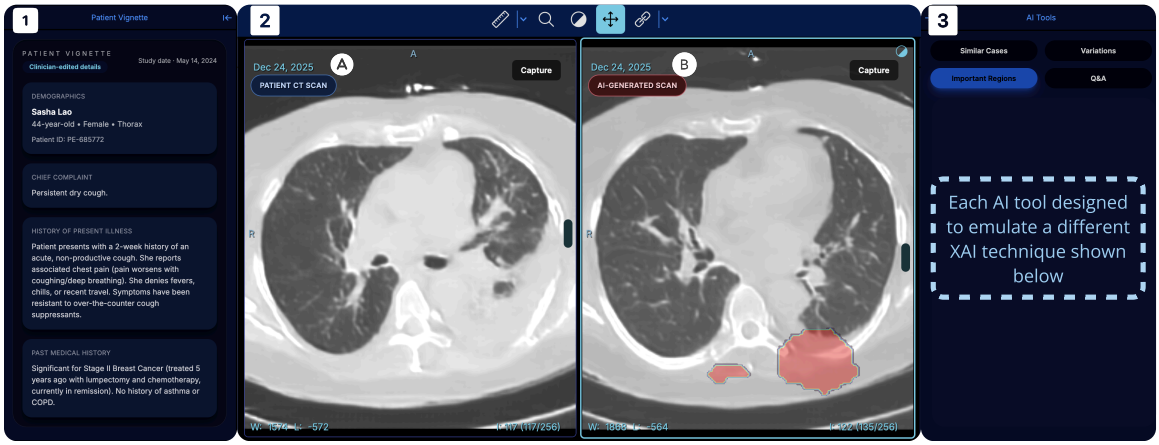


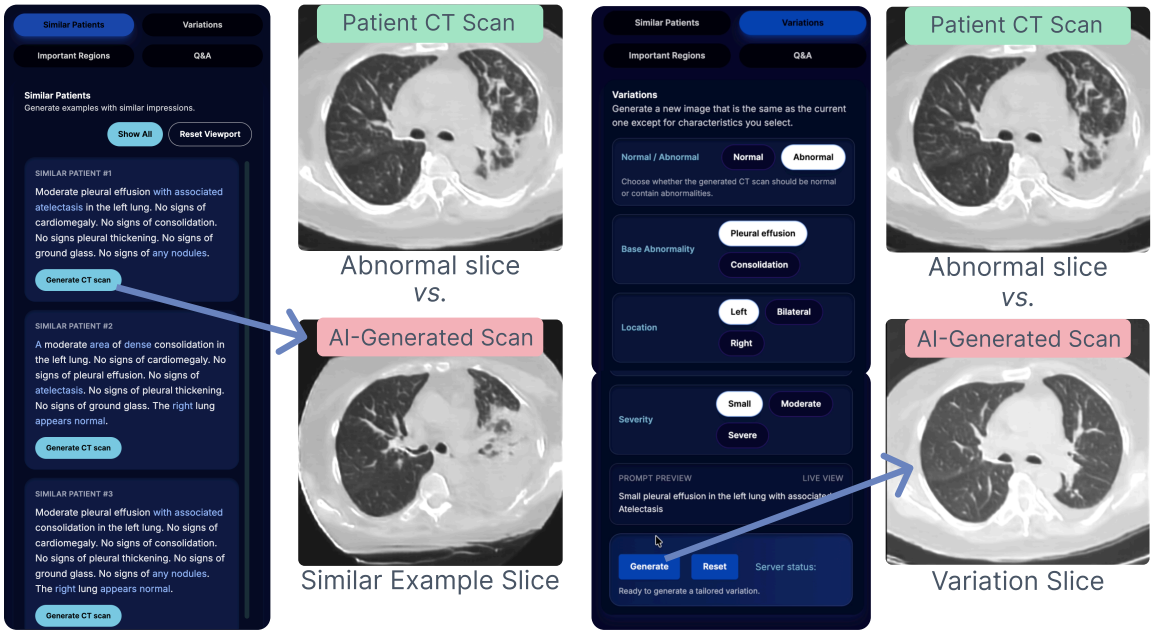
Fig. 8.2: GenXAI Clinical Reasoning Support Tool. (1) shows the patient vignette. (2) shows the viewing window: (a) shows the vignette CT scan and (b) shows a generated CT scan with important regions. (3) features the GenXAI Tools, shown below.

with technical jargon, we chose to call this panel 'AI Tools'. We will refer to these as GenXAI Tools.

8.2.3 Translating Reasoning Processes to GenXAI

Informed by findings from Morrison et al. [240] and Yildirim et al. [385] on applications of multimodal GenAI in radiology and radiologists' workflow processes when interpreting medical images, we follow DG2 to identify core reasoning processes for interpreting medical images and then translate those to GenXAI techniques derived from recent works [107, 214, 302]. To support DG1, we implement four GenXAI techniques that do not provide a direct AI recommendation for the underlying data to interpret. We describe each GenXAI Tool, its artifact, and the clinical reasoning process they aim to support below.

Similar Patients (Example-based Reasoning). Yildirim et al. [385] discuss how radiologists may search for similar example cases to help them reason. Grounded in our design goals, we built a GenXAI Tool that allows users to generate CT scans of similar abnormalities across different underlying anatomies. This explanation technique (Figure 8.3a) is inspired by comparative explanations from Cai et al. [51], the 'other possibilities' explanation from Lin et al. [214], and the example-based reasoning GenXAI technique [302]. Supporting DG3 and DG4, we show three similar patient impressions as suggestions for the user to choose from to generate, including one prompt that reflects the patient's CT scan and two others that are variations or potential mimics of the patient's condition. Because the underlying image-generation model can take more than 3 minutes to generate a new CT scan, we pre-generated CT scans for similar patients. We did not systematically assess whether the generated similar patients were clinically accurate.



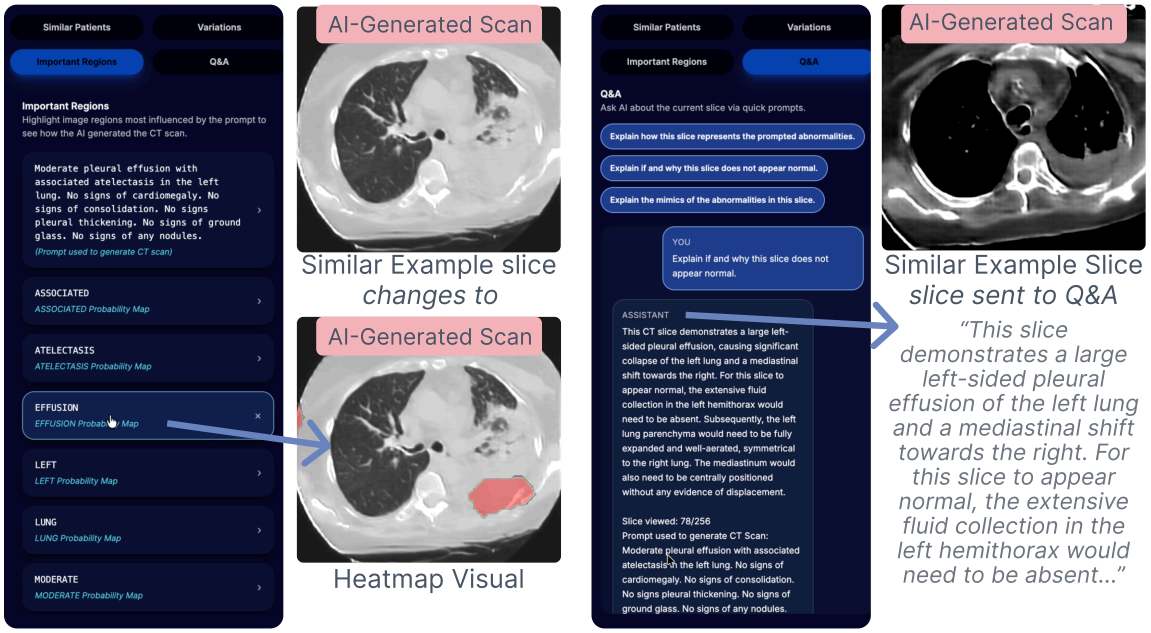
(a) Interface design for Similar Patients and an example slice from an AI-generated CT scan.

(b) Interface design for Variations and an example slice from an AI-generated variation.

Fig. 8.3: Interface designs and example AI explanations for Similar Patients (8.3a) and Variations (8.3b).

Variations (Counterfactual Generations). Reasoning by comparing to variations of abnormalities can also be valuable [240]. Inspired by recent counterfactual image generation approaches (*e.g.*, [140, 197, 214]), this explanation technique (Figure 8.3b) supports DG3 and DG4 by providing users with a prompt-building interface and is based on the counterfactual generations technique [302]. Users can generate CT scans by building prompt variations with different base abnormalities, severity, locations, and associated findings. We pre-generated every prompt combination to save time during the user study. Therefore, the prompt from the prompt builder is compared with the prompts from the generated set of CT scans to find its corresponding scan, which is then visualized in the viewport instantaneously. We did not systematically assess whether the generated scans contained clinically accurate or complete abnormalities.

Important Regions (Attention Visualizations). Localizing the abnormality is another important clinical reasoning process. We implement this as a visual attention mechanism (Figure 8.4a) that overlays a heatmap on the generated CT scan, highlighting image regions that receive high cross-attention from specific prompt tokens [39]. This tool only works on the generated CT scans, forcing users to reason about when and how to use it. In addition to supporting DG3 and DG4, this tool also targets DG5, allowing users to interpret or verify AI outputs if they wish. For example, users can highlight important regions to assess the validity of a generation or to try to localize an abnormality. The heatmaps were generated during the CT scan



(a) Interface design for Important Regions and an example heatmap revealed for an AI-generated CT scan.

(b) Interface design for Natural Language Q&A and an example response for an AI-generated slice.

Fig. 8.4: Interface designs and example AI explanations from the Important Regions (8.4a) and Natural Language Q&A tools (8.4b).

generation process. It is important to note that this explanation technique is prone to errors, as it has not been validated for generating precise pathology segmentation.

Natural Language Q&A (Surrogate Model). Lastly, seeking information from external resources to help explain differences and similarities between abnormalities is part of the clinical reasoning process. Based on the GenXAI techniques described by Schneider [302] and the ongoing attempts to use GenAI to explain domain concepts (*e.g.*, [132]), the Q&A tool (Figure 8.4b) provides three preset explanations that users can request for a specific slice of an AI-generated scan. Specifically, we allow users to ask for the AI to "*Explain how this slice represents the prompted abnormalities*", "*Explain if and why this slice does not appear normal*", and "*Explain the mimics of the abnormalities in this slice.*" These questions were piloted with a junior radiology resident. This GenXAI Tool also supports DG3, DG4, and DG5, enabling users to verify or interpret AI-generated scans. The Q&A feature is powered by Gemini 2.5 Flash, which was given system instructions to act as an expert radiology assistant with expertise in identifying chest abnormalities. We chose to use Gemini over GPT-5 because of its faster response time.

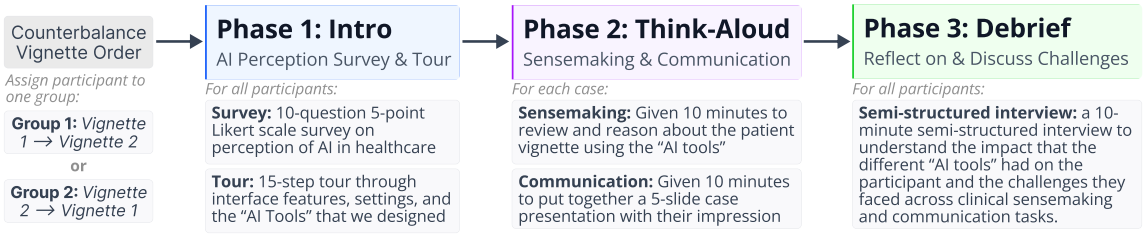


Fig. 8.5: Overview of user study protocol. Participants were assigned to either group one or group two to determine the order in which they would view the two patient vignettes. Phase 1, the Intro Phase, consisted of a 10-question survey and an interactive walkthrough tour of the interface. Phase 2, the Think-Aloud Phase, consisted of reviewing two patient vignettes and creating case presentations for each. Phase 3, the Debrief Phase, gave participants a space to reflect on the tools.

8.3 User Study

To better understand interaction patterns and how the GenXAI tools shape clinical reasoning (RQ1), as well as the benefits and challenges of process-oriented GenXAI designs (RQ2), we conducted a one-hour, three-phase user study with clinicians-in-training. To facilitate the different phases of the user study smoothly, we embedded our system into a secondary interface that we built[1]. We describe these phases below.

8.3.1 Protocol

Introduction Phase (10 minutes): Participants first took a 10-question Likert-scale survey on their perceptions of AI in healthcare [311]. Then, participants were provided an interactive walkthrough of the interface and the GenXAI Tools, so they had a better understanding of the artifacts that the GenXAI Tools produce. Participants were told they would only have access to the axial view (top-down view of the chest), given the limited time together, that the AI could not generate contrast-enhanced CT scans [240], and that all images may appear blurrier than expected [240].

Think-Aloud Phase (40 minutes): The Think-Aloud Phase lasted for the majority of the session, where participants reviewed two different patient vignettes in a pre-determined order based on their assigned group. While reviewing each vignette, participants were tasked with (1) making sense of the patient's condition (sensemaking task) and then (2) creating a four-slide case presentation to communicate their impression of the patient's condition (communication task). During both tasks, participants had access to all of the GenXAI Tools and could use them whenever they wanted. For both tasks, we use a concurrent think-aloud approach [104] to elicit participants' iterative reasoning processes. The interviewer encouraged think-aloud by asking clarifying questions when participants verbalized minimally.

Sensemaking Task (10 minutes per case): Participants were given ten minutes to

[1]Screenshots provided in the supplemental materials

think aloud as they made sense of the patient vignette and reviewed the patient's CT scan. They were told they can use the AI tools at any point during their sensemaking process and to think aloud as they use them.

Communication Task (10 minutes per case): Immediately following the sensemaking phase, participants were tasked with constructing a case presentation in Google Slides. This task was designed to mirror case presentations that are common in radiology residency programs. The case presentation consisted of six slides: an instruction slide, the patient vignette description, and four slides to be filled in by participants. Those four slides asked for the following information: key imaging findings, key imaging comparisons, differential diagnosis, and final impression. Again, participants could use the AI tools at any time and were encouraged to think aloud while doing so.

Debrief Phase (10 minutes): The study session concluded with a ten-minute semi-structured interview. The discussion focused on encouraging participants to reflect on the impact of GenXAI Tools on their reasoning and communication processes. It also aims to uncover the benefits and challenges of the different AI tools.

8.3.2 Patient Vignettes

Each vignette included fake patient demographics, a chief complaint, a history of present illness, a past medical history, and a corresponding CT scan. Our resident collaborator worked with us to generate this information for both vignettes to be reflective of the generated scans. Lastly, the patient CT scans for both patient vignettes are generated using MedSyn [380]. This allows users to generate a variation of the CT scan that preserves the same anatomical structures (*i.e.*, by reusing the same random noise distribution for a new generation). Both vignettes were reviewed by our radiology collaborator to have a ground-truth impression of the generated scans. The vignettes were informed by the abnormalities that the model has been shown to generate [380].

Patient Vignette A. The prompt for Patient A's CT scan was: "*Moderate pleural effusion with associated atelectasis in the left lung. No signs of cardiomegaly. No signs of consolidation. No signs pleural thickening. No signs of ground glass. No signs of any nodules.*" Our radiology collaborator reviewed the generated CT scan to confirm that the findings were consistent with moderate left pleural effusion and associated atelectasis. We mark the key findings in one slice shown in Figure 8.6. The prompts for the three similar patients are:

- "*Moderate pleural effusion with associated atelectasis in the left lung. No signs of cardiomegaly. No signs of consolidation. No signs pleural thickening. No signs of ground glass. No signs of any nodules.*"
- "*A moderate area of dense consolidation in the left lung. No signs of cardiomegaly. No signs of pleural effusion. No signs of atelectasis. No signs of pleural thickening. No signs of ground glass. The right lung appears normal.*"
- "*Moderate pleural effusion with associated consolidation in the left lung. No signs*

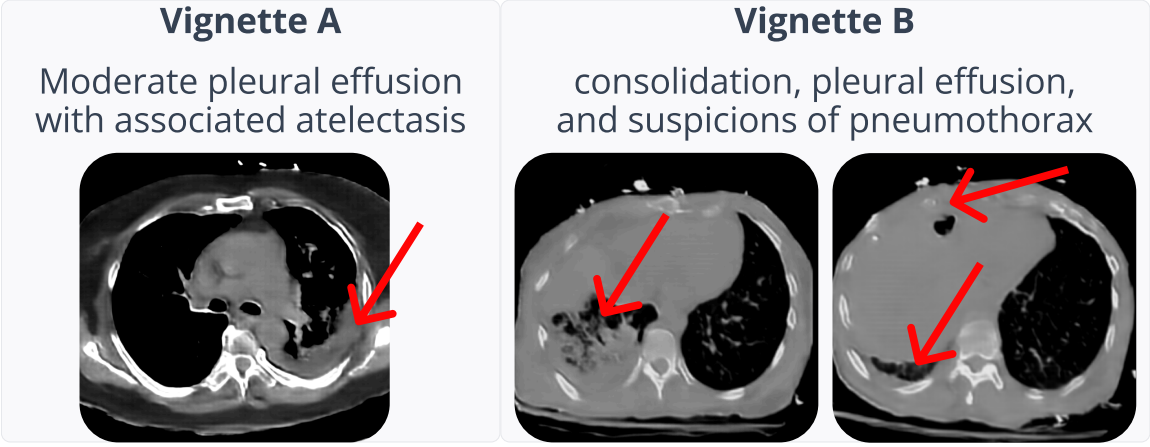


Fig. 8.6: Slices of key findings marked by red arrows in the CT scans for the two patient vignettes.

of cardiomegaly. No signs of consolidation. No signs of pleural thickening. No signs of ground glass. No signs of any nodules. The right lung appears normal."

Patient Vignette B. The prompt for Patient B's CT scan was: "*A large area of dense consolidation in the right lung. No signs of cardiomegaly. No signs of pleural effusion. No signs of atelectasis. No signs of pleural thickening. No signs of ground glass. The left lung appears normal.*" Our radiology collaborator reviewed the CT scan to confirm the presence of consolidation. However, the radiologist also noted a pleural effusion and suspected a pneumothorax. We mark these findings with red arrows in Figure 8.6. This case provides an opportunity to examine how participants interact with and reason through imperfect or internally inconsistent AI-generated explanations. The prompts for the three similar patients are:

- "*Severe pleural effusion in the right lung. No signs of cardiomegaly. No signs of consolidation. No signs of atelectasis. No signs of pleural thickening. No signs of ground glass. No signs of any nodules. The left lung appears normal.*"
- "*A large area of dense consolidation in the right lung. No signs of cardiomegaly. No signs of pleural effusion. No signs of atelectasis. No signs of pleural thickening. No signs of ground glass. No signs of any nodules. The left lung appears normal.*"
- "*Severe pleural effusion in the right lung with associated atelectasis. No signs of cardiomegaly. No signs of consolidation. No signs of pleural thickening. No signs of ground glass. No signs of any nodules. The left lung appears normal.*"

8.3.3 Study Interface

The top navigation bar shows the 10-question survey from the Introduction Phase (consent tab), the interactive walkthrough (walkthrough tab), the case viewer, case presentation, and case viewer 2. The case viewer tabs each display a different patient vignette, depending on the order the participant was assigned to see them.

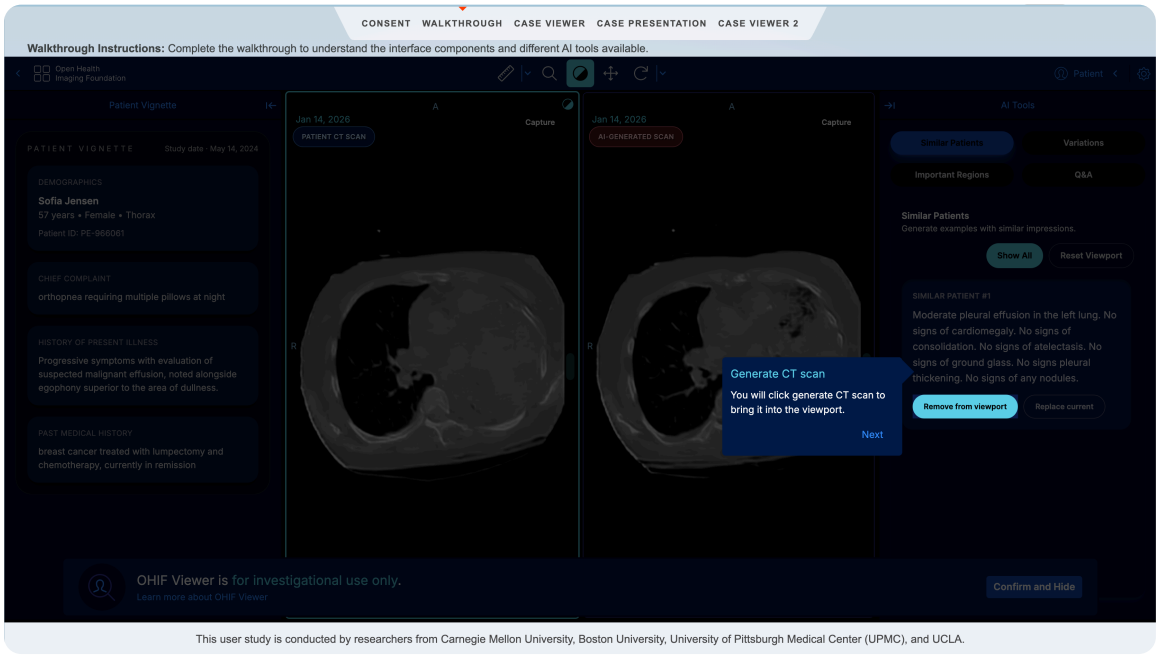


Fig. 8.7: Main User Interface where participants completed the user study. The navigation bar directs participants between the different phases of the study.

8.3.4 Recruitment & Data Analysis

We recruited senior medical students and junior radiology residents through the networks of our clinical collaborators at two institutions. Participants were compensated with 50USD for 60 minutes. Sessions were conducted remotely via Teams/Zoom for an average of 61.25 minutes ($SD = 4.3$ minutes).

Data collection included screen recordings, audio transcripts, and the case presentation decks. One author analyzed participants' interaction patterns with the GenXAI Tools during the sensemaking phase for each patient vignette and categorized them into similar interaction patterns. A second author reviewed the categorization of these interaction patterns and resolved any disagreements through discussions. Two authors conducted process coding [296] of the interview transcripts to identify how the GenXAI Tools shaped reasoning. Lastly, one author conducted structured coding [296] of quotes from the sessions to identify the benefits and challenges of designing with GenXAI. A second author joined for an affinity diagramming activity to identify themes within the benefits and challenges groups.

8.4 Findings

We conducted a pilot study and then recruited 10 clinicians-in-training: five 4th-year medical students (M1–M5) and five junior radiology residents (R1–R5) from two

PID	Expertise Level	Prof. Impact	Preparedness
R1	Resident (PGY2)	4.50	2.25
R2	Resident (PGY2)	3.67	2.25
R3	Resident (PGY3)	3.50	3.50
R4	Resident (PGY2)	2.67	2.25
R5	Resident (PGY2)	3.67	3.00
M1	4th-year Med Student	4.00	1.50
M2	4th-year Med Student	4.67	3.00
M3	4th-year Med Student	4.17	2.25
M4	4th-year Med Student	4.17	2.75
M5	4th-year Med Student	4.33	2.75

Table 8.1: Participant Demographics and scores for the Perception of AI in Healthcare [311] survey.

institutions. Table 8.1 provides a breakdown of each participant's scores from the survey based on factors (professional impact of AI and AI preparedness) identified by Shinners et al. [311]. In our study, the average AI preparedness score for the medical students was 2.45 ($SD = .53$) and 2.65 ($SD = .51$) for the residents. The average AI professional impact score was 4.27 ($SD = .23$) for medical students and 3.60 ($SD = .58$) for the residents. Across both groups, the AI preparedness score was lower than the professional impact score, consistent with the findings of Shinners et al. [311]. We also observed that the AI professional impact score was slightly higher among medical students than among residents. Despite these differences, a Welch's independent-samples t-test indicates that they are not significant ($t(5.18) = -2.13, p = .08$). These descriptive trends provide additional context for interpreting the observed interaction patterns and perceptions analyzed below.

The rest of this section presents findings for each research question. First, we identify three interaction patterns that participants tended to follow when using the GenXAI Tools to help them reason about the patient vignettes (RQ1). We complement this by examining how those different ways of engaging with the GenXAI Tools shaped clinical reasoning. Then, we present the benefits and challenges of using GenXAI for process-oriented support (RQ2).

8.4.1 RQ1: Interaction Patterns & Reasoning

We analyze behaviors across participants' sensemaking phases and denote the patient vignette by adding 'A' or 'B' after the participant identifier. Then, we group similar interactions to identify common patterns. As described in Table 8.2, we observed participants either using the GenXAI Tools as a starting point, to refine their reasoning, or to get a second opinion. Within these interaction patterns, we examine how these different ways of engaging with the AI tools can shape participants' reasoning processes.

All participants for the first vignette, except M1, chose to review the vignette and

	Starting-Point	**Refinement**	**Second-Opinion**
Impression Certainty	lack of impression	partial impression	coherent impression
Participants Observed	M1.A	M1.B, M2.A, M2.B, M3.A, M3.B, M4.B, M5.A, M5.B, R1.B, R3.A, R4.A, R5.A	M4.A, R1.A, R2.A, R2.B, R4.B

Legend: M = Medical Student — R = Radiology Resident
A = Vignette A — B = Vignette B

Table 8.2: Categorization of how participants interacted with the GenXAI Tools.

perform a top-to-bottom search of the CT scan before using the GenXAI Tools. Once the participants finished their initial reasoning, they used different GenXAI Tools as needed. M1 only reviewed the CT scan before using the GenXAI Tools and then later reviewed the patient's clinical information. R3 and R5 experienced technical difficulties that prevented them from completing the tasks for the second patient vignette, so their data for the second vignette were excluded from analysis.

8.4.1.1 Refining Impression of Medical Image. We observed that participants interacted most often (12/18) with the GenXAI Tools after forming an initial impression of the CT scan but remaining unsure about parts of it. To converge on a final impression, they used the GenXAI Tools to refine their reasoning by helping them discern the differences between pathologies (M1, M3, M4, M5, R4, R5), discern between different severity levels of a pathology (R1, M2, M3, M4), orient their attention to localizing abnormalities (M2, M5, R5), or expand their reasoning (R1, R2, R3, R5, M1, M4, M5).

Discern between different abnormalities/severities. As the example shows in Figure 8.8, M4 noted that, *"It does look like there is some sort of consolidation or fluid in the right lung"* while scrolling through the CT scan. They generated a similar patient with severe pleural effusion to compare against their patient's CT scan, and then generated another similar patient showing only consolidation. However, after generating and comparing two similar patients' CT scans with the patient's CT scan, they admitted that both generated CT scans looked similar to the patient's. This leads them to generate two variations, one for consolidation with pleural effusion and one for consolidation without pleural effusion, saying "*...the part that I was a little less sure about was the pleural effusion. So, then I made one without the pleural effusion.*" This interaction led them to correctly identify that vignette B had both pleural effusion and consolidation. For M2, they used the Variations Tool to help them reason about different severities. For example, they generated multiple variations, saying "*...I guess my thought process is more: 'How do I determine if it's small, moderate, or severe?'*"

Orient attention to abnormalities. Instead of discerning differences, we saw how M2,

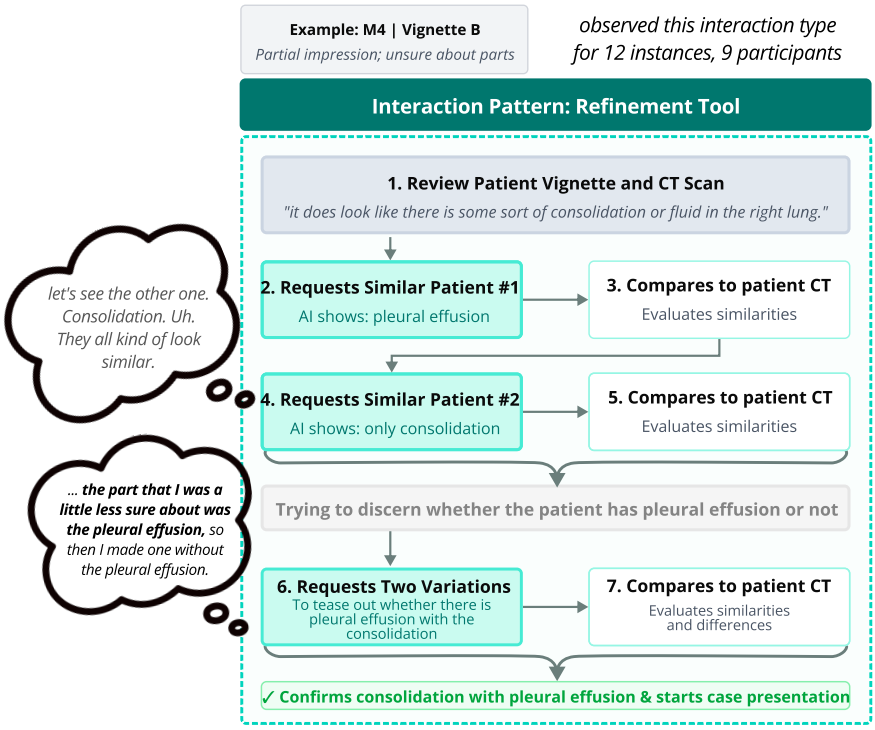


Fig. 8.8: Example flow of using the GenXAI Tools to refine an impression.

M5, and R5 leveraged the GenXAI Tools as a way to orient their attention towards where abnormalities were located to help them refine their impressions. For example, as M2 was going through the CT scan for vignette B, they used the Important Regions tool to see if it could help them tease out whether there was atelectasis: "*...I wanted to get an impression of where it [AI] thinks the atelectasis is, just cause then if I see it on the AI-generated, then I think it might help me know what to expect on the patient one.*" M5 similarly used the Important Regions tool to, "*...look for specific parts of the AI scan*" and orient their attention to where effusion was in the generated image. After reviewing the heatmap for "effusion", they felt "*...more confident that there's at least some level of an effusion going on in the patient CT scan.*" R5 used the Q&A tool on a slice of a generated scan to "*..see if it can give me any insight into how to identify it in the future on my own.*"

Expand reasoning. The GenXAI Tools helped several participants, including medical students and residents (R1, R2, R3, R5, M1, M4, M5), expand their reasoning during sensemaking and case presentations. For example, R1 experienced how the Q&A tools helped them ideate additional reasonable differentials to list in their case presentations: "*...So I mentioned hemothorax, but I didn't even mention all of these other possibilities...*". R2, R3, and R5 benefited similarly from the Q&A tool. For example, R2 said that, "*...the first one [vignette], I didn't think about a necrotizing pneumonia. I was like, 'Oh yeah, this is a good thing to think about'.*" M5

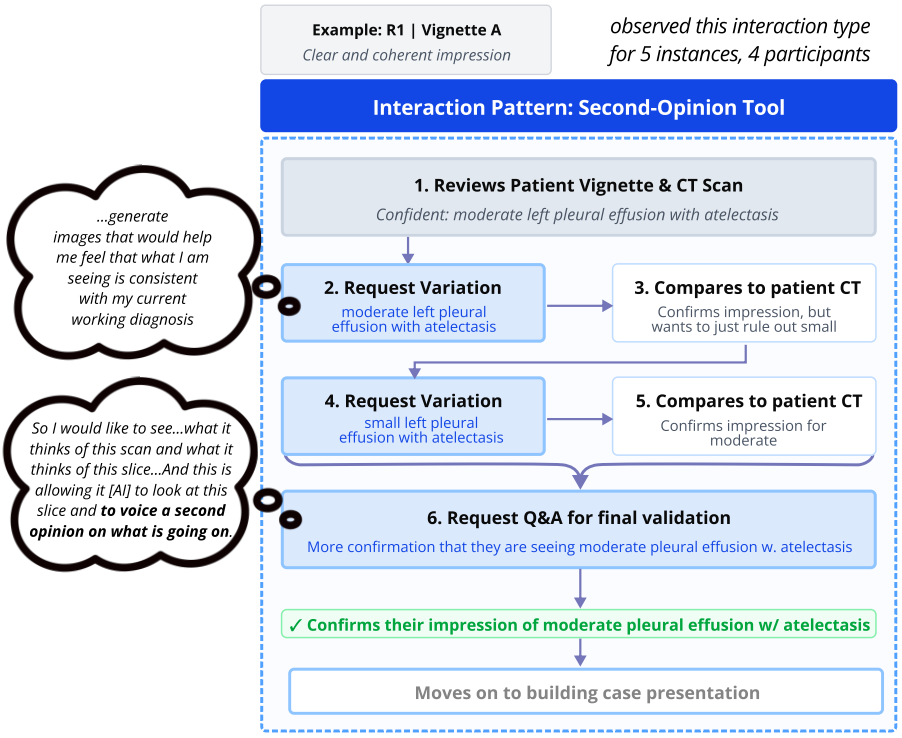


Fig. 8.9: Example flow of using the GenXAI Tools to get a second opinion.

clearly delineated the impact the Q&A tool had on their reasoning process and case presentation by showing their list of differentials and a list of additional differentials marked "*with AI consideration.*"

8.4.1.2 Second-Opinion Tool. We observed five instances across four participants who initially formed a coherent impression of the CT scan and interacted with the AI tools, hoping to reinforce their reasoning. This interaction pattern shaped reasoning by either providing reassurance to participants (R1, M4, R4) or serving as a space to challenge model behaviors (R2).

Provides Reassurance. As the sequence in Figure 8.9 shows, R1 concluded after reviewing vignette A and its corresponding CT scan that there were signs of moderate left pleural effusion with atelectasis. With that clear impression, they wanted to "*...generate images that would help me feel that what I am seeing is consistent with my current working diagnosis.*" So, they generated a variation for that exact impression. After comparing their generation to their patient's, they concluded "*It's pretty similar, which also makes me think that my diagnosis of this as moderate is not unreasonable.*" They go on to use the Q&A tool and feel reassured because "*...the AI is also looking at this image and...coming up with a similar diagnosis as what I am.*" M4 similarly started out by generating a variation of what their exact impression was after they finished reviewing the CT scan. R4 described how using AI-generated images as an additional point of comparison aligned with their existing radiology workflow when

confirming their line of reasoning: "*in radiology, we're always comparing stuff like how stuff looks with each other. So, as another point of comparison, it was nice.*"

Opportunity to Challenge Model Behavior. R2 shifted from using the GenXAI tools to initially get a second opinion to challenging the model. For example, R2 was the only participant who identified all three pathologies for vignette B and even spent time calling out hallucinations they saw in the images. This led them to generate a similar patient because they were "*...curious to see if the image it generates has an effusion anyway.*" Although this was just one instance, it is still a behavior to take note of when users have different experiences with the AI system.

8.4.1.3 Used as a starting point. When lacking an impression of the data, the GenXAI Tools could provide a starting point for reasoning. We observed one instance in which the GenXAI Tools could misguide the reasoning of less knowledgeable users (M1). In contrast, more knowledgeable users suggested that the GenXAI Tools could jumpstart

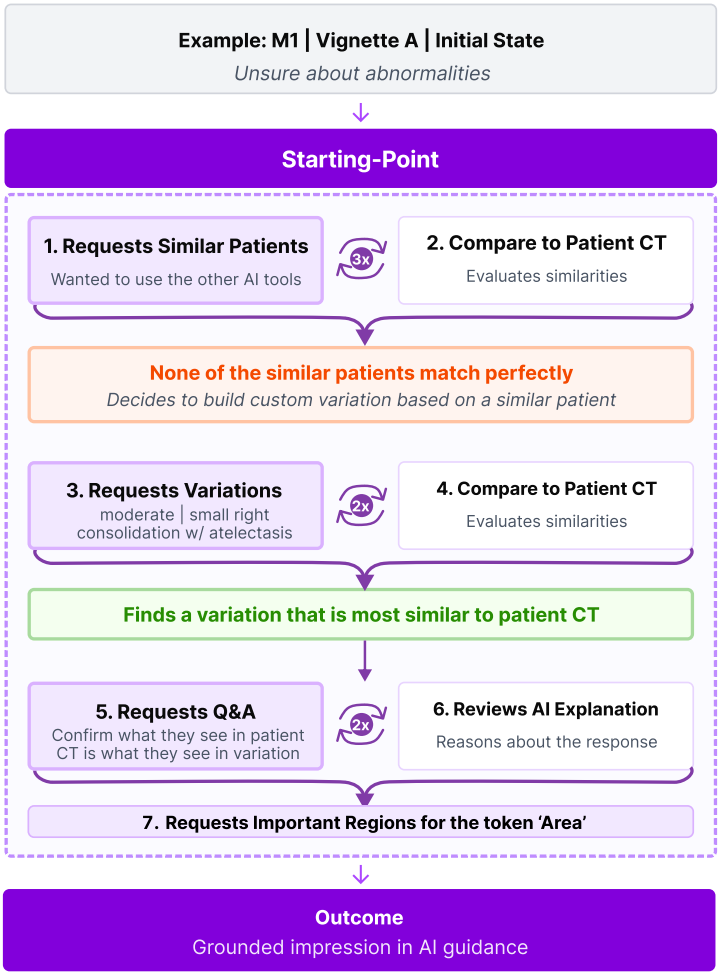


Fig. 8.10: Example flow of interactions that represent the process-oriented XAI system being used as a second-opinion tool.

their reasoning when they were stuck (R1, M2, M5).

Misguide Reasoning. We observed one instance in which one participant (M1.A) used the GenXAI Tools to establish a reference point for reasoning before deeply engaging in diagnostic reasoning themselves, which ultimately ended up misguiding their reasoning process. M1 generated a similar patient case within one minute of viewing the CT scan and without reading the clinical history from the vignette. They stated that their goal in doing this was "*Because the other tools don't work unless it's AI-generated*", and they wanted to start using the other tools. After trying to find a CT scan that most visually reflected their patient by comparing against all of the generated similar patient CT scans, they generated a more visually similar CT scan using the variation tool. However, they generated a variation using a prompt that did not accurately reflect the patient's CT. Despite this, they then used this generated scan as a proxy for their patient's abnormalities, reasoning with the Q&A explanations, and then copy/pasting the response from the Q&A into their case presentation. Despite acknowledging that they would crosscheck this response with external knowledge bases, they also said earlier in the session that the Q&A tool, "*...would save me a lot of time in terms of typing and thinking about what to say, which I don't think is a useful use of my skills, so I can copy and paste and edit quickly.*" M1 reflected on their use of the GenXAI Tools as a starting point for reasoning during the debrief, saying: "*...even I'll point out my own fault between the first and second case, like, after the first case, I went straight to the AI tool instead of working through [the case].*" Interestingly, M1 had the lowest AI preparedness score, shown in Table 8.1, which could explain their misuse of the AI tools.

Can Jumpstart Reasoning. Despite seeing how the GenXAI Tools could provide a starting point that is misleading, some participants suggested that the AI tools could be useful in providing a starting point when unsure how to begin reasoning about a case. For example, R1 said "*...if I see that something is wrong, but I can't even begin to think of how to describe it...I would use it [AI Tools]...as a starting point to get its [AI] view on what does it think is going on and potentially use that as a sounding room essentially to kind of get me on a process to thinking about what this could even be.*" M5 framed it to be "*...helpful to have a challenging viewpoint, one that's challenging what you're seeing and maybe throwing some other ideas out there.*" M2 similarly acknowledged that it could provide a good starting point when you have no idea what you are looking at. This is backed by what they said about their current workflow: "*I started off by doing the similar patients, which I think was nice because in my head that's more or less how I look at scans already*".

8.4.2 RQ2: Benefits & Challenges of GenXAI

Based on participants' comments while interacting with the GenXAI Tools and their reflections on how they engaged with the tools, we identify several benefits and challenges of using GenXAI for process-oriented support.

8.4.2.1 Benefits of Imperfect GenXAI. **Externalizes Reasoning.** AI explanations that participants perceived as questionable or inconsistent throughout reasoning prompted them to be skeptical of the content within the explanation. As a result, participants externalized their reasoning by agreeing or disagreeing with different parts of the explanation (R1, R2, R3, M1, M5). Imperfections in the Q&A response describing a similar patient prompted M5 to inspect the generated scan more closely, leading them to agree with some parts of the explanation while questioning others.

Encourages Critical Reflection. Three participants mentioned how AI explanations that they perceive as imperfect or questionable could encourage them to reflect more critically on their own interpretations and line of reasoning (R1, M1, M5). For example, R1 suggested how it could encourage a more critical review of the patient's scan: "*...if it said, oh, this looks more like consolidation or this looks more like hemothorax, for example, that would help me to rethink and revisit and look at the search pattern again, maybe look more closely for [additional findings]*". M5 acknowledged that,"*...it's helpful to have a challenging viewpoint, like one that's challenging what you're seeing and maybe throwing some other ideas out there.*"

Support in reducing anchoring bias. R2 acknowledged that imperfections can help mitigate anchoring bias by enabling someone to "*...avoid the satisfaction of the search and keep [their] mind open to other abnormalities.*" M1 similarly suggested that they could use our process-oriented XAI system to "*...not anchor on the first thing that comes to mind.*"

8.4.2.2 Challenges of Imperfect GenXAI. **Difficult to verify explanation quality**. Some participants spoke to the difficulty of verifying the quality of an AI explanation (M2, M3, M4). Lacking domain knowledge can make it more difficult for users to identify when a generation is inaccurate. For example, M3 said "*If you're someone with experience, you would look at it and you would know. But someone like me wouldn't know everything. Nothing really looks imperfect to me.*" M4 said "*I don't know if the AI generated scan is a good point of comparison...when I was scrolling through to try to find the pathology myself, I feel like there's some pathologies where I can't do that because I don't know what I'm looking for really.*" In contrast, participants with more expertise commented on how the variation they asked for had additional abnormalities (R1, R4). For example, R1 asked for a pleural effusion without atelectasis wondering how the model would behave because "*...in reality, if you had a pleural effusion this big, you would never have no atelectasis.*". They acknowledged that the generation did include atelectasis.

Over-reliance on AI Explanations. Because users' level of domain knowledge impacts whether they are able to notice imperfect explanations, it puts them at higher risk for over-relying on the information and reasoning conveyed by the generated explanation. M1 and M5 commented on the dangers of over-relying on these AI explanations. For example, M5 said that, "*...it could be hurtful if someone is just purely relying on what the AI is interpreting and just rolls with what it's seeing and describing without being skeptical.*"

Users turn to self-blame when they encounter imperfections. More than half of the medical students blamed themselves for not being able to generate the explanation they wanted (M1, M2, M3). For example, when M2 was using the Important Regions tool, it produced an overlay over a region that did not make sense. But, instead of blaming the model, they instead said, "*Like I feel like I have an idea of what it looks like...And also, I'm a med student. I don't know anything, which is the main answer.*" M3 acknowledged their lack of training, saying "*...similar patient three and similar patient two to me with my untrained eye look incredibly similar.*"

8.4.3 Other Challenges with GenXAI

Confirmation Bias. Several participants (R1, M3, M4, R4, M5) felt that the tool could have "*...a lot of confirmation bias in it...if I think it's something, I can get it to generate to confirm what I'm thinking*" (M3), which was a concern to them. R1 similarly brings up this concern, stating that "*...there still wasn't an exact ground truth...*" after generating a scan based on their impression to compare with their patient's scan. Similarly, R4 said that the Q&A explanation is, "*biased because [they're] the one that chose the prompts for it to generate [the referred CT scan].*"

8.5 Discussion

Designing XAI for human–AI collaboration, especially in clinical contexts, often focuses on providing AI recommendations and explaining them rather than supporting processes crucial to clinical reasoning [318]. Shifting away from this narrow focus on output-oriented XAI, this chapter introduces process-oriented XAI as an interaction paradigm for XAI design. Our findings suggest that process-oriented XAI can support clinicians-in-training's reasoning during medical image interpretation in multiple ways. We saw participants use the same XAI techniques differently (*e.g.*, as a starting point, a refinement aid, or a second opinion) depending on their uncertainty in reasoning and their goals while interpreting the image. In turn, these interaction patterns were associated with different reasoning effects, including misguiding, converging, and expanding reasoning. We discuss the implications of these findings below.

Process-oriented XAI designs change how people use AI explanations. We saw how participants primarily leveraged the different XAI components in our tool to refine incomplete impressions or confirm already coherent ones, rather than to verify a model-generated impression. This observation directly addresses RQ1 by showing that participants used process-oriented XAI as a starting point, a refinement aid, or a second opinion. These interaction patterns build upon the Evaluative AI framework proposed by Miller [233], in which XAI can be reframed as judgment support, trade-off support, or an additional source of options. They also align with ongoing Tools for Thought research that has called for shifting the design of AI systems toward supporting human cognition and reasoning processes [339, 389] rather than directly providing

recommendations. Our findings are also consistent with Gajos and Mamykina [115], who found that cognitive engagement with AI explanations is contingent on the interaction paradigm. By removing AI predictions (DG1), our process-oriented design may enable more effective analytical engagement. These findings, corroborated by findings from empirical studies, suggest that our process-oriented XAI interaction paradigm can foster new interactions with XAI during human–AI collaboration.

Participants appropriated the AI explanations based on their needs. We observed how each participant's workflows were slightly different in the order they used AI explanations to support them. This finding captures the outcome of designing AI explanations for appropriation. This diversity in interaction patterns that we observed extends findings from recent output-oriented AI studies that observed decision-makers negotiating with AI recommendations [317]. Dix [89] emphasizes that in some cases, it may be better to "*design a system so that a task can be done*" as opposed to "*designing a system to do the task*". Ultimately, by removing AI recommendations and offering AI explanations that translate to different reasoning processes, the process-oriented interaction paradigm may be able to expand the space of interaction possibilities.

Explanations perceived as imperfect may lead to productive friction, whereas imperfections that go unnoticed may lead to harmful friction. We observed several medical students struggling to assess the quality of their explanations or resorting to self-blame when they encountered imperfections. Eslami et al. [105] similarly observe how users resorted to self-blame when encountering imperfect AI systems. However, we also observed how participants externalized their reasoning when encountering questionable explanations. These findings suggest that perceived imperfections in explanations could create productive friction, whereas imperfections that go unnoticed may create harmful friction. As observed in related work, these imperfections could be seen less as flaws and more as design goals [127, 219]. Furthermore, Buçinca et al. [45] has observed how friction, in the form of different cognitive forcing functions, can reduce over-reliance on AI systems by enabling the user to think more. Imperfections can prompt overconfident thinkers to reconsider whether they are indeed thinking about the case correctly, which can be a form of productive friction. However, this could become a source of harmful friction for less knowledgeable users, as it could be misleading or lead them down unproductive rabbit holes. This dichotomy is captured by Morrison et al. [239] and Dakhel et al. [79], who conclude that domain expertise moderates the benefits gained from AI systems.

8.6 Limitations & Future Work

This study was limited in its controlled aspects, as we were primarily interested in exploring emerging interaction patterns and their impacts on reasoning without quantifying them.

Limited sample size with clinicians-in-training. Because we only had five partic-

ipants in each group, it is unclear if and how process-oriented XAI designs would generalize to clinicians-in-training from other institutions, more senior radiology residents, or expert radiologists. Future work could explore whether and how process-oriented XAI designs are adapted differently across levels of expertise, from medical students learning medical image interpretation to seasoned radiology experts.

Limited diversity in patient vignettes. We created two patient vignettes based on image findings that are fairly common and recognizable to experienced medical students and radiology residents. The lack of medical images with less familiar abnormalities could alter the interaction patterns we observed and their impacts on clinical reasoning. Future work could quantitatively explore how task difficulty and initial impression confidence moderate how people interact with process-oriented XAI, as prior work suggests that these factors can shape the impact of output-oriented XAI [364].

8.7 Conclusion

Explainable AI (XAI) systems have primarily been designed as tools to help humans verify, interpret, or justify AI outputs during human–AI collaboration. While countless empirical studies have highlighted the limitations of this output-oriented interaction paradigm for XAI, recent work shows promising potential for interacting with XAI during human–AI collaboration in more meaningful, user-driven ways. To this end, we propose a new interaction paradigm for XAI in human–AI collaboration, called process-oriented XAI, and instantiate it in an interactive tool that integrates four multimodal GenXAI techniques to assist in medical image interpretation. We capture how clinicians-in-training interact with our process-oriented XAI system to provide a starting point in reasoning, refine their reasoning, or give a second opinion. Our analyses also highlight the ongoing challenges with managing over-reliance on AI explanations in process-oriented XAI systems. We hope our contributions inspire human–AI collaboration researchers to instantiate our process-oriented XAI paradigm in other domains.

Part IV

Conclusions

Chapter 9

Discussion & Conclusion

9.1 Summary of Research Contributions

The purpose of the work in this thesis is to contribute to the ongoing research in human-centered XAI.

9.1.1 Part I: Imperfect Output-Oriented XAI in Human–AI Collaboration

Part I acknowledges the impacts of imperfect output-oriented XAI in human–AI collaboration by showing that explanations can be misleading, depending on users' level of domain expertise and the explanation type:

Chapter 3 identifies five human explanation strategies by eliciting explanations from crowdworkers when classifying building damage: causal argument, comparison, localization, extent of damage, and quantity of damaged structures. By presenting these explanation strategies to another set of 60 crowdworkers, we observed how causal explanations sometimes led humans astray when the AI's assessment was correct, but the identified location was incorrect for the damage. On the contrary, when the assessment was wrong and the localization was correct, causal explanations empowered humans to override the AI correctly.

Chapter 4 formalizes the concept of Imperfect XAI and presents an empirical study showing how the effects of imperfect, output-oriented XAI on human–AI collaboration are shaped by the decision-maker's level of domain expertise. This chapter demonstrates that the same explanation can be helpful or harmful, depending on who receives it and the type of explanation.

9.1.2 Part II: Emerging Uses of AI Explanations

Part II shifts from evaluating imperfect, output-oriented XAI in controlled decision-making studies with crowdworkers toward identifying emerging uses of AI explanations with domain stakeholders across real-world workflows.

Chapter 5 presents a holistic need-finding study with model builders, product designers, and five end-users to show how XAI is appropriated beyond output verification or interpretation in a widely deployed wildlife conservation tool.

Chapter 6 engages 18 clinicians in co-designing an interactive XAI component for a clinical decision-support tool. Through a design probe study with 7 physicians, we show how XAI can support multi-stakeholder interactions, sensemaking, and knowledge transfer.

9.1.3 Part III: Expanding the Design Space w/ Process-Oriented XAI

Part III formally expands the design space for AI explanations in human–AI collaborations by repurposing Explainable Generative AI (GenXAI) and introducing the process-oriented XAI paradigm. We instantiate this paradigm using GenXAI in a tool for medical image interpretation to augment reasoning processes throughout radiology workflows:

Chapter 7 develops a design probe to explore how text-to-image GenAI can support training and decision-making processes in radiology with medical students, radiology residents, and attendings. Our findings reveal opportunities to leverage GenAI to generate AI explanations.

Chapter 8 introduces the process-oriented XAI paradigm and instantiates it in a clinical tool for medical image interpretation using GenXAI techniques. In a think-aloud and semi-structured interview study with ten clinicians-in-training, we show how process-oriented XAI designs can serve as a starting point, a refinement aid, or a second opinion. We also observed how these interaction patterns can be associated with misguiding, converging, or expanding reasoning.

9.2 Discussion & Future Research Directions

This thesis expands the design space for XAI in human–AI collaboration by acknowledging the impact of imperfect AI explanations on human–AI decisions, uncovering the emerging uses of AI explanations in human–AI collaboration, and reframing XAI techniques as tools to support human reasoning processes. We discuss the implications and future work for each part of this thesis.

9.2.1 Navigating the Benefits and Harms of Imperfect AI Explanations.

Part I addresses RQ1 of this thesis by exposing the negative impact that imperfect explanations can have on human reasoning while working with an AI classifier, especially when the human's domain expertise is low. Recent follow-up work has further explored the impact that imperfect AI explanations can have on one's procedural knowledge later on when making decisions alone without the AI system [327]. In Part III, we saw a similar trend: domain expertise played a role in whether the imperfections

were harmful or helpful. Although this thesis emphasizes the importance of designing XAI with stakeholders and of expanding beyond narrow, output-oriented paradigms, it does not intend to argue that output-oriented mechanisms should be abandoned. In fact, the findings across Parts I and III suggest that there remains an important need for mechanisms that help users verify, interpret, and critically engage with AI explanations, even as the design space expands toward more process-oriented XAI designs.

Further research is needed to better understand when imperfect AI explanations are helpful versus harmful in human–AI collaborations. As a result, this future work could contribute to a taxonomy of factors beyond the domain-expertise level that influence when imperfections are beneficial versus harmful. Recent literature on imperfect AI systems has reframed these imperfections as design goals rather than flaws (*e.g.*, [127, 219]) in certain contexts, motivating the need to further explore if and how imperfect AI explanations during human–human–AI interactions could be supportive. However, it is also equally necessary to investigate when and how these imperfections become harmful.

9.2.2 Designing AI Explanations to Align with Stakeholders' Workflow Processes.

By working closely with stakeholders in Parts II and III, we transitioned from building XAI for stakeholders to building XAI with the stakeholders. This approach helped us address RQ2 of this thesis, uncovering emerging uses of AI explanations that aim to support unmet needs and augment, rather than replace, processes throughout their workflows. We also capture how the emerging uses of XAI may differ across domains and contexts. For example, Chapter 6 captures emerging uses of XAI for shared clinical decision-making among physicians and patients or between physicians and other specialists. In contrast, Chapter 7 observes how XAI in the domain of medical imaging interpretation may be more valuable as training-oriented scaffolds for reasoning, comparison, and reflection. One possible reason is that the multi-stakeholder shared decision-making context central to Chapter 6 is less directly present in radiology workflows, where image interpretation often centers more heavily on individual sensemaking. These differences in XAI use across contexts emphasize the importance of designing XAI with, not for, stakeholders to better align XAI with their existing workflows.

Drawing on Malle [228]'s work on how the mind explains behavior, this will be an interesting avenue for future researchers to explore how XAI can be designed to better support human–human interactions, shared decision-making, and collective reasoning processes. For example, AI explanations could be designed to support collective reasoning among care staff members as they try to agree on the best treatment plan for a patient. Recent work has explored the use of multi-agent systems by clinicians to help prepare for multidisciplinary tumor board meetings [206], a human–human

interaction. Future work could also explore lab studies examining how LLM-generated explanations influence a value-diverse group's ability to reach consensus on subjective problems, such as moral dilemmas and resource allocation.

9.2.3 Expanding the Design Space Responsibly.

Part III addresses RQ3 of this thesis by formally expanding the design space for AI explanations with the introduction of the process-oriented XAI paradigm using GenXAI techniques. We qualitatively observed how end-users' level of domain expertise can influence the kind of impact that imperfect AI explanations will have. However, it is necessary to further explore this design space of imperfect AI explanations in more depth. Future work can explore how to design safe process-oriented XAI tools that instill caution in users when allowing AI explanations to scaffold their reasoning. One contribution could be by developing safeguards for AI explanations. For example, researchers could investigate retrieval-augmented generation techniques that ground AI explanations in domain-relevant data [132] to mitigate their imperfections. The tool developed in Chapter 8 could have implemented this by grounding the Q&A tool in domain-specific knowledge [132], leveraging more advanced promptable segmentation techniques for the important regions tool (*e.g.*, [208, 229]), or grounding image generations in a similar real CT scan retrieved from a popular crowd-sourced database for radiology called Radiopaedia (*i.e.*, [237]). Future work can explore how these designs address challenges with imperfect AI explanations in workflows.

9.3 Conclusion

When I began my Ph.D., I was fascinated by the emerging field of human–AI collaboration and explainable AI. Early on in the Ph.D., I often described my research as trying to explain AI outputs to people so they could make better decisions. Over time, however, I came to see that this framing positioned humans as AI supervisors rather than as individual thinkers. As I became more interested in designing AI systems that augment rather than replace human judgment, my perspective shifted from explaining AI outputs to using AI explanations to help people explain, communicate, reason, and collaborate.

This thesis examines both the effects of explaining AI and the broader potential of using AI explanations to explain complex concepts. My hope is that this work encourages others to approach XAI not as a fixed technical problem aimed solely at making AI systems more acceptable or more efficient, but as a design space that can be reframed in response to evolving human goals, workflows, and contexts. For the XAI community, I hope this dissertation offers a broader perspective on the development, evaluation, and motivation of explanations. For the HCI community, I hope it points toward new opportunities for designing and operationalizing process-oriented XAI responsibly.

References

[1] Wild me. URL https://www.wildme.org/. Last accessed 2026-04-06. 4.1

[2] All about birds guide. URL https://www.allaboutbirds.org/guide/. Last accessed 2026-04-06. 4.3.3, 4.6

[3] inaturalist. URL https://www.inaturalist.org/. Last accessed 2026-04-06. 4.1

[4] What is explainable ai (xai)? URL https://www.ibm.com/think/topics/explainable-ai. Last accessed 2026-04-06. 1

[5] Merlin all about birds. URL https://merlin.allaboutbirds.org/. Last accessed 2026-04-06. 4.1, 4.3.1

[6] Open health imaging foundation (ohif) viewer. URL https://ohif.org. Last accessed 2026-04-06. (document), 7.1, 7.2.3, 7.3.1, 7.2

[7] Crowdai, 2022. URL no longer active. 3.2.1

[8] Mapswipe, 2022. URL https://mapswipe.org/en/. Last accessed 2026-04-06. 3.2.1

[9] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8466590. 1, 2.1, 2.2, 2.4

[10] Temi A Adekunle, Joy M Knowles, Sarah V Hantzmon, Maya N DasGupta, Kathryn I Pollak, and Sarah E Gaither. A qualitative analysis of trust and distrust within patient-clinician interactions. *PEC innovation*, 3:100187, 2023. URL https://www.sciencedirect.com/science/article/pii/S2772628223000675. 6.6.1.1

[11] Hüseyin Gökhan Akçay, Bekir Kabasakal, Duygugül Aksu, Nusret Demir, Melih Öz, and Ali Erdoğan. Automated bird counting with deep learning for regional bird distribution mapping. *Animals*, 10(7):1207, 2020. URL https://doi.org/10.3390/ani10071207. 4.3.1

[12] Stephan Alaniz. pytorch-gve-lrcn: Pytorch implementation of visual generation and execution for long-term predictions, 2018. URL https://github.com/salaniz/pytorch-gve-lrcn. 4.3.4

[13] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 275–285, 2020. URL https://dl.acm.org/doi/abs/10.1145/3377325.3377519. 2.2, 5.1

[14] 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): 5088, 2021. URL https://www.mdpi.com/2076-3417/11/11/5088. 1

[15] Joseph L Austerweil and Thomas L Griffiths. Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, 35(3):499–526, 2011. URL https://cocosci.princeton.edu/

tom/papers/seekconfisrational.pdf. 3.5.1

[16] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16, 2021. URL https://doi.org/10.1145/3411764.3445717. 2.3, 3.5.1, 4.2, 4.3.6, 4.5

[17] Catarina Barata and Carlos Santiago. Improving the explainability of skin cancer diagnosis using cbir. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III*, pages 550–559. Springer, 2021. URL https://doi.org/10.1007/978-3-030-87199-4_52. 4.1, 4.3.2, 4.3.4

[18] Woodrow Barfield. Expert-novice differences for software: Implications for problem-solving and knowledge acquisition. *Behaviour & Information Technology*, 5(1):15–29, 1986. URL https://doi.org/10.1080/01449298608914495. 4.2

[19] Shraddha Barke, Michael B James, and Nadia Polikarpova. Grounded copilot: How programmers interact with code-generating models. *Proceedings of the ACM on Programming Languages*, 7(OOPSLA1):85–111, 2023. URL https://doi.org/10.1145/3586030. 2.3

[20] Sarah Bayer, Henner Gimpel, and Moritz Markgraf. The role of domain expertise in trusting and following explainable ai decision support systems. *Journal of Decision Systems*, 32(1): 110–138, 2022. URL https://doi.org/10.1080/12460125.2021.1958505. 2.2, 4.2

[21] Michaela Benk and Tim Miller. Same performance, hidden bias: Evaluating hypothesis-and recommendation-driven ai. *arXiv preprint arXiv:2603.15824*, 2026. URL https://arxiv.org/abs/2603.15824. 2.6

[22] Raymond L Benza, Mardi Gomberg-Maitland, C Greg Elliott, Harrison W Farber, Aimee J Foreman, Adaani E Frost, Michael D McGoon, David J Pasta, Mona Selej, Charles D Burger, et al. Predicting survival in patients with pulmonary arterial hypertension: the reveal risk score calculator 2.0 and comparison with esc/ers-based risk assessment strategies. *Chest*, 156 (2):323–337, 2019. URL https://www.sciencedirect.com/science/article/pii/S0012369219301527. 6.2

[23] Isacco Beretta, Eleonora Cappuccio, and Marta Marchiori Manerba. User-driven counterfactual generator: A human centered exploration. In *xAI (Late-breaking Work, Demos, Doctoral Consortium)*, pages 83–88, 2023. URL https://openportal.isti.cnr.it/data/2023/489902/2023_489902.published.pdf. 1.1, 2.1, 6.3.2.2

[24] Tanya Y Berger-Wolf, Daniel I Rubenstein, Charles V Stewart, Jason A Holmberg, Jason Parham, Sreejith Menon, Jonathan Crall, Jon Van Oast, Emre Kiciman, and Lucas Joppa. Wildbook: Crowdsourcing, computer vision, and data science for conservation. *arXiv preprint arXiv:1710.08880*, 2017. URL https://arxiv.org/abs/1710.08880. 4.1, 5.1, 5.2.1

[25] Thales Bertaglia, Stefan Huber, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. Closing the loop: Testing chatgpt to generate model explanations to improve human labelling of sponsored content on social media. *arXiv preprint arXiv:2306.05115*, 2023. URL https://arxiv.org/abs/2306.05115. 4.1

[26] Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R Eagan, and Winston Maxwell. On selective, mutable and dialogic xai: A review of what users say about different types of interactive explanations. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–21, 2023. URL https://dl.acm.org/doi/10.1145/3544548.3581314. 1, 1.1, 6.1

[27] Hugh Beyer and Karen Holtzblatt. Contextual design. *Interactions*, 6(1):32–42, 1999. URL https://dl.acm.org/doi/fullHtml/10.1145/291224.291229. 5.2.2

[28] Gauri Bhardwaj, Yuvaraj Govindarajulu, Sundaraparipurnan Narayanan, Pavan Kulkarni, and Manojkumar Parmar. On the notion of hallucinations from the lens of bias and validity in synthetic cxr images. *arXiv preprint arXiv:2312.06979*, 2023. URL https://arxiv.org/abs/2312.06979. 7.1

[29] Aditya Bhattacharya, Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. Directive explanations for monitoring the risk of diabetes onset: introducing directive data-centric explanations and combinations to support what-if explorations. In *Proceedings of the 28th international conference on intelligent user interfaces*, pages 204–219, 2023. URL https://dl.acm.org/doi/pdf/10.1145/3581641.3584075. 1.1, 2.1, 2.2, 6.1, 6.3.2.2, 6.3.2.2, 6.3.2.3, 6.6.1.1

[30] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504, 2023. URL https://dl.acm.org/doi/abs/10.1145/3593013.3594095. 7.1

[31] Ahsan Bilal, David Ebert, and Beiyu Lin. Llms for explainable ai: A comprehensive survey. *arXiv preprint arXiv:2504.00125*, 2025. URL https://arxiv.org/abs/2504.00125. 2.5

[32] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 396–410, 2023. URL https://dl.acm.org/doi/abs/10.1145/3600211.3604722. 7.1

[33] Johanna Birkhäuer, Jens Gaab, Joe Kossowsky, Sebastian Hasler, Peter Krummenacher, Christoph Werner, and Heike Gerger. Trust in the health care professional and health outcome: A meta-analysis. *PloS one*, 12(2):e0170988, 2017. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0170988. 6.6.1.1

[34] Jason Bloomberg. Don't trust artificial intelligence? time to open the ai black box. *Forbes. September*, 16:2018, 2018. URL https://www.forbes.com/sites/jasonbloomberg/2018/09/16/dont-trust-artificial-intelligence-time-to-open-the-ai-black-box/. 1

[35] Drew Blount, Shane Gero, Jon Van Oast, Jason Parham, Colin Kingen, Ben Scheiner, Tanya Stere, Mark Fisher, Gianna Minton, Christin Khan, et al. Flukebook: an open-source ai platform for cetacean photo identification. *Mammalian Biology*, 102(3):1005–1023, 2022. URL https://doi.org/10.1007/s42991-021-00221-3. 5.1

[36] Kirsten Boehner, Janet Vertesi, Phoebe Sengers, and Paul Dourish. How hci interprets the probes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1077–1086, 2007. URL https://dl.acm.org/doi/abs/10.1145/1240624.1240789. 6.4.3

[37] Eve Bohnett, Jason Holmberg, Sorosh Poya Faryabi, Li An, Bilal Ahmad, Wajid Rashid, and Stephane Ostrowski. Comparison of two individual identification algorithms for snow leopards (panthera uncia) after automated detection. *Ecological Informatics*, 77:102214, 2023. URL https://doi.org/10.1016/j.ecoinf.2023.102214. 5.3.3.2

[38] Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. Role of human-ai interaction in selective prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5286–5294, 2022. URL https://doi.org/10.1609/aaai.v36i5.20465. 4.3.1

[39] Aidan Bradshaw, Katelyn Morrison, Arpit Mathur, Weicheng Dai, Motahhare Eslami, Adam Perer, et al. Toward interpretable 3d diffusion in radiology: Token-wise attribution for text-to-ct synthesis. In *Medical Imaging with Deep Learning-Short Papers*, 2025. URL https://openreview.net/pdf?id=DTYFRzRPQn. 7.4.3.1, 7.5.3, 8.2.3

[40] Virginia Braun and Victoria Clarke. Reflecting on reflexive thematic analysis. *Qualitative*

research in sport, exercise and health, 11(4):589–597, 2019. URL https://doi.org/10.1080/2159676x.2019.1628806. 3.3.1.5

[41] Andrea Brennen. What do people really want when they say they want "explainable ai?" we asked 60 stakeholders. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–7, 2020. URL https://dl.acm.org/doi/abs/10.1145/3334480.3383047. 3.1

[42] Ross Brisco, Laura Hay, and Sam Dhami. Exploring the role of text-to-image ai in concept generation. *Proceedings of the Design Society*, 3:1835–1844, 2023. URL https://www.cambridge.org/core/journals/proceedings-of-the-design-society/article/exploring-the-role-of-texttoimage-ai-in-concept-generation/BEAFFFED638C0F4C78E8AFA978FECD5A. 7.5.1

[43] Dieter Brughmans, Lissa Melis, and David Martens. Disagreement amongst counterfactual explanations: how transparency can be misleading. *Top*, 32(3):429–462, 2024. URL https://link.springer.com/content/pdf/10.1007/s11750-024-00670-2.pdf. 1.1, 8.2.1

[44] Zana Buçinca, Siddharth Swaroop, Amanda E Paluch, Finale Doshi-Velez, and Krzysztof Z Gajos. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–25, New York, NY, USA, 2025. ACM. doi: 10.1145/3706598.3713229. URL https://dl.acm.org/doi/full/10.1145/3706598.3713229. 2.6, 8.2.1

[45] 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. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021. URL https://dl.acm.org/doi/abs/10.1145/3449287. 1.1, 2.6, 8.1, 8.2.1, 8.5

[46] Eleanor R Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J Marc Overhage, Erika S Poole, and Jofish Kaye. Healthcare ai treatment decision support: Design principles to enhance clinician adoption and trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023. URL https://dl.acm.org/doi/full/10.1145/3544548.3581251. 2.2, 6.1, 6.5.4

[47] Garrick Cabour, Andrés Morales, Élise Ledoux, and Samuel Bassetto. Towards an explanation space to align humans and explainable-ai teamwork, 2021. URL http://arxiv.org/abs/2106.01503. 2.2, 5.1, 5.4

[48] Ángel Alexander Cabrera, Abraham J Druck, Jason I Hong, and Adam Perer. Discovering and validating ai errors with crowdsourced failure reports. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–22, 2021. URL https://doi.org/10.1145/3479569. 3.5.1, 5.4

[49] Ángel Alexander Cabrera, Adam Perer, and Jason I Hong. Improving human-ai collaboration with descriptions of ai behavior. *Proceedings of the ACM on Human-Computer Interaction*, 7 (CSCW1):1–21, 2023. URL https://doi.org/10.1145/3579612. 1, 4.3.1, 5.4, 6.1, 6.3.2.1, 8.1

[50] Ángel Alexander Cabrera, Marco Tulio Ribeiro, Bongshin Lee, Robert Deline, Adam Perer, and Steven M Drucker. What did my ai learn? how data scientists make sense of model behavior. *ACM Transactions on Computer-Human Interaction*, 30(1):1–27, 2023. URL https://dl.acm.org/doi/full/10.1145/3542921. 6.6.1.2

[51] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 258–262, 2019. URL https://doi.org/10.1145/3301275.3302289. 2.1, 4.3.2, 4.3.4, 8.2.3

[52] 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. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–24, 2019. URL https://dl.acm.org/doi/abs/10.1145/3359206. 3.1, 4.1, 5.2.2, 5.4, 6.1, 7.3.2

[53] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. Onboarding materials as cross-functional boundary objects for developing ai assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021. URL https://doi.org/10.1145/3411763.3443435. 5.4

[54] Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C Nascimento. Assertiveness-based agent communication for a personalized medicine on medical imaging diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023. URL https://doi.org/10.1145/3544548.3580682. 2.2, 4.1, 4.2, 4.2, 4.3.4, 4.5

[55] Oana-Maria Camburu and Zeynep Akata. Natural-xai: Explainable ai with natural language explanations. In *International Conference on Machine Learning*, 2021. URL https://icml.cc/media/icml-2021/Slides/10835.pdf. 1.1, 2.5

[56] Shiye Cao and Chien-Ming Huang. Understanding user reliance on ai in assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–23, 2022. URL https://dl.acm.org/doi/abs/10.1145/3555572. 6.1

[57] Josie Carmichael. Translating human-centred artificial intelligence for clinical decision support systems into practice: A medical retina case study. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–5, 2024. URL https://dl.acm.org/doi/full/10.1145/3613905.3638182. 6.6.3.1

[58] Nancy Carter, Denise Bryant-Lukosius, Alba DiCenso, Jennifer Blythe, and Alan J Neville. The use of triangulation in qualitative research. In *Oncology nursing forum*, volume 41, pages 545–547, 2014. URL https://educationaldevelopment.uams.edu/wp-content/uploads/sites/57/2025/01/5-Triangulation.pdf. 5.1, 5.4

[59] Federico Maria Cau and Lucio Davide Spano. The influence of curiosity traits and on-demand explanations in ai-assisted decision-making. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1440–1457, 2025. URL https://dl.acm.org/doi/full/10.1145/3708359.3712165. 1.1, 2.2, 8.2.1

[60] Federico Maria Cau and Lucio Davide Spano. Exploring the impact of explainable ai and cognitive capabilities on users' decisions. *User Modeling and User-Adapted Interaction*, 36(1):3, 2026. URL https://link.springer.com/article/10.1007/s11257-025-09438-0. 1, 6.1

[61] Cecilia Ka Yuk Chan. A comprehensive ai policy education framework for university teaching and learning. *International journal of educational technology in higher education*, 20(1):38, 2023. URL https://link.springer.com/article/10.1186/s41239-023-00408-3. 7.1

[62] Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. It takes two to tango: Towards theory of ai's mind. *arXiv preprint arXiv:1704.00717*, 2017. URL https://arxiv.org/abs/1704.00717. 4.3.2

[63] David S Channin, Pattanasak Mongkolwat, Vladimir Kleper, and Daniel L Rubin. The annotation and image mark-up project, 2009. URL https://web.stanford.edu/group/rubinlab/pubs/590.full.pdf. 3.1

[64] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018. doi: 10.1109/wacv.2018.00097. URL http://dx.doi.org/10.1109/WACV.2018.00097. 5.3.2.1

[65] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. URL https://dl.acm.org/doi/abs/10.1145/3592116. 7.5.3

[66] Haomin Chen, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. Explainable medical imaging ai needs human-centered design: Guidelines and evidence from a systematic review, September 2022. URL http://arxiv.org/abs/2112.12596. 2.2, 5.1

[67] 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. *Proceedings of the ACM on Human-computer Interaction*, 7(CSCW2):1–32, 2023. URL https://dl.acm.org/doi/abs/10.1145/3610219. 2.2, 4.1, 4.3.2, 4.6

[68] Wenting Chen, Pengyu Wang, Hui Ren, Lichao Sun, Quanzheng Li, Yixuan Yuan, and Xiang Li. Medical image synthesis via fine-grained image-text alignment and anatomy-pathology prompting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 240–250. Springer, 2024. URL https://link.springer.com/chapter/10.1007/978-3-031-72390-2_23. 7.1

[69] Michael Chromik and Andreas Butz. Human-xai interaction: a review and design principles for explanation user interfaces. In *IFIP Conference on Human-Computer Interaction*, pages 619–640. Springer, 2021. URL https://link.springer.com/chapter/10.1007/978-3-030-85616-8_36. 2.2

[70] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. I think i get your point, ai! the illusion of explanatory depth in explainable ai. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21, page 307–317, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380171. doi: 10.1145/3397481.3450644. URL https://doi.org/10.1145/3397481.3450644. 2.2

[71] Douglas Cirqueira, Dietmar Nedbal, Markus Helfert, and Marija Bezbradica. Scenario-based requirements elicitation for user-centric explainable ai: A case in fraud detection. In *International cross-domain conference for machine learning and knowledge extraction*, pages 321–341. Springer, 2020. URL https://doi.org/10.1007/978-3-030-57321-8_18. 2.2, 5.1

[72] Herbert H Clark and Susan E Brennan. Grounding in communication. *Perspectives on socially shared cognition/American Psychological Association*, 1991. URL https://collablab.northwestern.edu/CollabolabDistro/nucmc/ClarkAndBrennan-GroundingInCommunication-1991.pdf. 6.6, 6.5, 6.6.1.1

[73] Eric Corbett and Emily Denton. Interrogating the t in facct. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1624–1634, 2023. URL https://doi.org/10.1145/3593013.3594104. 5.1

[74] United Therapeutics Corporation. Pah initiative risk assessment calculators, 2024. URL https://www.pahinitiative.com/hcp/risk-assessment/calculators. (document), 6.1, 6.2

[75] Lorenzo Corti, Rembrandt Oltmans, Jiwon Jung, Agathe Balayn, Marlies Wijsenbeek, and Jie Yang. "it is a moving process": Understanding the evolution of explainability needs of clinicians in pulmonary medicine. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024. URL https://doi.org/10.1145/3613904.3642551. 2.2, 2.4, 5.3.3, 6.1, 6.6.1.1

[76] Jonathan P Crall, Charles V Stewart, Tanya Y Berger-Wolf, Daniel I Rubenstein, and Siva R Sundaresan. Hotspotter—patterned species instance recognition. In *2013 IEEE workshop on applications of computer vision (WACV)*, pages 230–237. IEEE, 2013. URL https://doi.or

g/10.1109/wacv.2013.6475023. 5.2.1

[77] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 427–439, 2022. URL https://doi.org/10.1145/3531146.3533108. 5.4

[78] Karla Cruz Morel, Vinicio De Jesus Perez, and Arunabh Talwar. Gender and race disparities in pulmonary hypertension diagnosis and treatment. *Pulmonary Hypertension: Controversial and Emerging Topics*, pages 195–202, 2020. URL https://link.springer.com/chapter/10.1007/978-3-030-52787-7_12. 6.1

[79] Arghavan Moradi Dakhel, Vahid Majdinasab, Amin Nikanjam, Foutse Khomh, Michel C Desmarais, and Zhen Ming Jack Jiang. Github copilot ai pair programmer: Asset or liability? *Journal of Systems and Software*, 203:111734, 2023. URL https://doi.org/10.1016/j.jss.2023.111734. 2.3, 8.5

[80] Anindya Das Antar, Somayeh Molaei, Yan-Ying Chen, Matthew L Lee, and Nikola Banovic. Vime: Visual interactive model explorer for identifying capabilities and limitations of machine learning models for sequential decision-making. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–21, 2024. URL https://dl.acm.org/doi/abs/10.1145/3654777.3676323. 1

[81] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12, 2020. URL https://dl.acm.org/doi/abs/10.1145/3313831.3376638. 1

[82] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. The perils and pitfalls of explainable ai: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 39 (2):101666, 2022. URL https://doi.org/10.1016/j.giq.2021.101666. 4.2

[83] Josep Munuera del Cerro and Arnau Valls Esteve. How imaging generative ai will transform the medical radiological practice. *HealthManagement.org: The Journal*, 23(2), 2023. URL https://healthmanagement.org/c/healthmanagement/IssueArticle/how-imaging-generative-ai-will-transform-the-medical-radiological-practice. 7.1, 7.5.2

[84] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. Understanding practices, challenges, and opportunities for user-engaged algorithm auditing in industry practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023. URL https://doi.org/10.1145/3544548.3581026. 5.4

[85] Wesley Hanwen Deng, Michelle S Lam, Ángel Alexander Cabrera, Danaë Metaxa, Motahhare Eslami, and Kenneth Holstein. Supporting user engagement in testing, auditing, and contesting ai. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 556–559, 2023. URL https://doi.org/10.1145/3584931.3611279. 5.4

[86] Hannah Deters, Jakob Droste, Mathis Fechner, and Jil Klünder. Explanations on demand-a technique for eliciting the actual need for explanations. In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 345–351. IEEE, 2023. URL https://ieeexplore.ieee.org/abstract/document/10260835. 1.1

[87] Shipi Dhanorkar, Christine T Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. Who needs to know what, when?: Broadening the explainable ai (xai) design space by looking at explanations across the ai lifecycle. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, pages 1591–1602, 2021. URL https://dl.acm.org/doi/abs/10.1145/3

461778.3462131. 2.2

[88] Murat Dikmen and Catherine Burns. The effects of domain knowledge on trust in explainable ai and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, 162:102792, 2022. URL https://doi.org/10.1016/j.ijhcs.2022.102792. 2.2, 4.2

[89] Alan Dix. Designing for appropriation. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK*. BCS Learning & Development, 2007. URL https://pdfs.semanticscholar.org/7e5a/e497d0ca00a46d0f50e6cf57414a82ec3080.pdf. 1.2, 6.6, 6.5, 6.6.1.2, 8.1, 8.5

[90] Yuhan Du, Anna Markella Antoniadi, Catherine McNestry, Fionnuala M McAuliffe, and Catherine Mooney. The role of xai in advice-taking from a clinical decision support system: a comparative user study of feature contribution-based and example-based explanations. *Applied Sciences*, 12(20):10323, 2022. URL https://doi.org/10.3390/app122010323. 1, 1, 2.2, 4.6, 6.1

[91] Gunning D Aha DW. Darpa's explainable artificial intelligence program. *AI Magazine*, 40(2): 44, 2019. URL https://onlinelibrary.wiley.com/doi/epdf/10.1609/aimag.v40i2.2850?msockid=2d0e9bc100fb68401a248d3e0170697d. 1

[92] Sven Eckhardt, Niklas Kühl, Mateusz Dolata, and Gerhard Schwabe. A survey of ai reliance. *ACM Computing Surveys*, 58(6):1–37, 2025. URL https://dl.acm.org/doi/full/10.1145/3776528. 6.1

[93] Upol Ehsan and Mark O Riedl. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International conference on human-computer interaction*, pages 449–466. Springer, 2020. URL https://link.springer.com/chapter/10.1007/978-3-030-60117-1_33. 1.1, 2.2

[94] Upol Ehsan and Mark O Riedl. Explainability pitfalls: Beyond dark patterns in explainable ai. *arXiv preprint arXiv:2109.12480*, 2021. URL https://arxiv.org/abs/2109.12480. 5.1

[95] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2021. URL https://doi.org/10.1145/3411764.3445188. 1.1, 2.2

[96] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, and others. The who in xai: how ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509*, 2021. URL https://arxiv.org/abs/2107.13509v1. 1.1

[97] 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 *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–6, 2021. URL https://dl.acm.org/doi/abs/10.1145/3411763.3441342. 2.2

[98] 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 *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022. URL https://doi.org/10.1145/3491101.3503727. 2.2, 4.1

[99] Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O Riedl. Charting the sociotechnical gap in explainable ai: A framework to address the gap in xai. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–32, 2023. URL https://doi.org/10.1145/3579467. 5.1

[100] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daumé III. Seamful xai: Operationalizing seamful design in explainable ai. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–29, 2024. URL https://dl.acm.org/doi/abs/10.1145/3637396. 2.2, 5.1

[101] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. Bringing transparency design into practice. In *Proceedings of the 23rd international conference on intelligent user interfaces*, pages 211–223, 2018. URL https://dl.acm.org/doi/10.1145/3172944.3172961. 2.2, 5.1

[102] Alexander El-Ali, Fariha Kamal, C Lynn Cabral, and Judy H Squires. Comparison of traditional and web-based medical student teaching by radiology residents. *Journal of the American College of Radiology*, 16(4):492–495, 2019. ISSN 1546-1440. doi: 10.1016/j.jacr.2018.09.048. URL https://doi.org/10.1016/j.jacr.2018.09.048. 7.5.2

[103] Elsevier Inc. Statdx, 2025. URL https://statdx.com. 7.2.2.1

[104] K. Anders Ericsson and Herbert A. Simon. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA, 1984. URL https://direct.mit.edu/books/monograph/4763/Protocol-AnalysisVerbal-Reports-as-Data. 8.3.1

[105] Motahhare Eslami, Sneha R Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. Communicating algorithmic process in online behavioral advertising. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018. URL https://dl.acm.org/doi/abs/10.1145/3173574.3174006. 8.5

[106] Pouyan Esmaeilzadeh, Tala Mirzaei, and Spurthy Dharanikota. Patients' perceptions toward human–artificial intelligence interaction in health care: experimental study. *Journal of medical Internet research*, 23(11):e25856, 2021. URL https://www.jmir.org/2021/11/e25856. 6.6.1.1

[107] Noyan Evirgen, Ruolin Wang, and Xiang'Anthony Chen. From text to pixels: Enhancing user understanding through text-to-image model explanations. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 74–87, 2024. URL https://dl.acm.org/doi/fullHtml/10.1145/3640543.3645173. 7.5.3, 8.2.3

[108] Florian Eyert and Paola Lopez. Rethinking transparency as a communicative constellation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 444–454, 2023. URL https://doi.org/10.1145/3593013.3594010. 5.1

[109] Andrea Ferrario and Michele Loi. How explainability contributes to trust in ai. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1457–1466, 2022. URL https://dl.acm.org/doi/abs/10.1145/3531146.3533202. 2.2

[110] Raymond Fok and Daniel S Weld. In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. *arXiv preprint arXiv:2305.07722*, 2023. URL https://arxiv.org/abs/2305.07722v1. 4.4.3

[111] Courtney Ford and Mark T Keane. Explaining classifications to non experts: An xai user study of post hoc explanations for a classifier when people lack expertise. In *International Conference on Pattern Recognition*, pages 246–260. Springer, 2022. URL https://doi.org/10.1007/978-3-031-37731-0_15. 2.2, 4.3.4

[112] Timo Freiesleben and Gunnar König. Dear xai community, we need to talk! fundamental misconceptions in current xai research. In *World conference on explainable artificial intelligence*, pages 48–65. Springer, 2023. URL https://link.springer.com/chapter/10.1007/978-3-031-44064-9_3. 1, 6.1

[113] Nancy Friday, Tim D Smith, Peter T Stevick, and Judy Allen. Measurement of photographic quality and individual distinctiveness for the photographic identification of humpback whales,

megaptera novaeangliae. *Marine Mammal Science*, 16(2):355–374, 2000. URL https://doi.org/10.1111/j.1748-7692.2000.tb00930.x. 5.3.3.1

[114] Elisa Gagnon, Anouk de Regt, and Liette LaPointe. Clarity in complexity: Advancing ai explainability through sensemaking. In *Proceedings of the 58th Hawaii International Conference on System Sciences, HICSS 2025*, pages 1400–1409, 2025. URL https://research-portal.uu.nl/en/publications/clarity-in-complexity-advancing-ai-explainability-through-sensema/. 6.6.1.2

[115] Krzysztof Z. Gajos and Lena Mamykina. Do people engage cognitively with AI? impact of AI assistance on incidental learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, pages 794–806, New York, NY, USA, 2022. ACM. doi: 10.1145/3490099.3511138. URL https://dl.acm.org/doi/abs/10.1145/3490099.3511138. 8.1, 8.5

[116] Ziba Gandomkar and Claudia Mello-Thoms. Visual search in breast imaging. *The British journal of radiology*, 92(1102):20190057, 2019. URL https://doi.org/10.1259/bjr.20190057. 3.2.2

[117] Davide Gentile, Birsen Donmez, and Greg A Jamieson. Human performance effects of combining counterfactual explanations with normative and contrastive explanations in supervised machine learning for automated decision assistance. *International Journal of Human-Computer Studies*, 196:103434, 2025. URL https://www.sciencedirect.com/science/article/pii/S1071581924002179. 2.2

[118] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K Kummerfeld, and Elena L Glassman. Supporting sensemaking of large language model outputs at scale. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024. URL https://dl.acm.org/doi/full/10.1145/3613904.3642139. 7.5.2

[119] Dennis A Gioia, Kevin G Corley, and Aimee L Hamilton. Seeking qualitative rigor in inductive research: Notes on the gioia methodology. *Organizational research methods*, 16(1):15–31, 2013. URL https://doi.org/10.1177/1094428112452151. 4.1, 4.3.6

[120] Catalina Gomez, Sue Min Cho, Shichang Ke, Chien-Ming Huang, and Mathias Unberath. Human-ai collaboration is not very collaborative yet: A taxonomy of interaction patterns in ai-assisted decision making from a systematic review. *Frontiers in Computer Science*, 6:1521066, 2025. URL https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2024.1521066/full. 1

[121] Google. Google search, 2026. URL https://www.google.com/. 7.2.2.1

[122] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019. URL https://doi.org/10.1145/3359152. 4.3.2, 4.3.6

[123] Siân E Green, Jonathan P Rees, Philip A Stephens, Russell A Hill, and Anthony J Giordano. Innovations in camera trapping technology and approaches: The integration of citizen science and artificial intelligence. *Animals*, 10(1):132, 2020. URL https://doi.org/10.3390/ani10010132. 4.3.1

[124] David Gunning, Eric Vorm, Yunyan Wang, and Matt Turek. Darpa's explainable ai (xai) program: A retrospective. *Authorea Preprints*, 2021. URL https://onlinelibrary.wiley.com/doi/full/10.1002/ail2.61?msockid=2d0e9bc100fb68401a248d3e0170697d. 1, 1.1

[125] Grace Guo, Lifu Deng, Animesh Tandon, Alex Endert, and Bum Chul Kwon. Mimicri: Towards domain-centered counterfactual explanations of cardiovascular image classification models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1861–1874, 2024. URL https://dl.acm.org/doi/abs/10.1145/3630106.3659011. 2.2

[126] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 10–17, 2019. URL https://doi.org/10.1184/R1/8135576.v1. 3.1, 3.2, 3.2.1, 3.2.3, 3.3.1.3, 3.3.1.4, 3.5.2.4

[127] Jose A Guridi, Angel Hsing-Chi Hwang, Duarte Santo, Maria Goula, Cristobal Cheyre, Lee Humphreys, and Marco Rangel. From fake perfects to conversational imperfects: Exploring image-generative ai as a boundary object for participatory design of public spaces. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–33, 2025. URL https://dl.acm.org/doi/10.1145/3710912. 8.2.1, 8.5, 9.2.1

[128] Umm-E Habiba, Justus Bogner, and Stefan Wagner. Can requirements engineering support explainable artificial intelligence? towards a user-centric approach for explainability requirements. In *2022 IEEE 30th international requirements engineering conference workshops (REW)*, pages 162–165. IEEE, 2022. URL https://ieeexplore.ieee.org/abstract/document/9920137. 2.2, 5.1

[129] Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Dogan, Muhammed Furkan Dasdelen, et al. Generatect: Text-conditional generation of 3d chest ct volumes. *arXiv preprint arXiv:2305.16037*, 2023. URL https://arxiv.org/abs/2305.16037v1. (document), 7.1, 7.2, 7.3.2.3

[130] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arXiv:2403.17834*, 5, 2024. URL https://arxiv.org/abs/2403.17834v1. (document), 7.2, 7.3.2.3

[131] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. *arXiv preprint arXiv:2403.06801*, 2024. URL https://arxiv.org/abs/2403.06801. (document), 7.2, 7.3.2.3

[132] Ameer Hamza, Yong Hyun Ahn, Sungyoung Lee, Seong Tae Kim, et al. Llava needs more knowledge: Retrieval augmented natural language generation with knowledge graph for explaining thoracic pathologies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3311–3319, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/32342. 8.2.3, 9.2.3

[133] Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, and Kentaro Inui. Evaluation of similarity-based explanations. *arXiv preprint arXiv:2006.04528*, 2020. URL https://arxiv.org/abs/2006.04528. 6.3.2.1

[134] Andrew F Hayes. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications, 2017. URL https://doi.org/10.1111/jedm.12050. 4.1, 4.4.2, 4.4.2.1, 4.4.2.2

[135] Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. Is conversational xai all you need? human-ai decision making with a conversational xai assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 907–924, 2025. URL https://dl.acm.org/doi/full/10.1145/3708359.3712133. 1.1, 2.2

[136] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. Human-ai complementarity in hybrid intelligence systems: A structured literature review. *PACIS 2021 Proceedings*, 2021. URL https://aisel.aisnet.org/pacis2021/78/. 1.1, 2.3, 4.1, 4.3.6, 4.4, 4.4.3, 4.5

[137] Lisa Anne Hendricks, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Zeynep Akata.

Generating visual explanations with natural language. *Applied AI Letters*, 2(4):e55, 2021. URL https://doi.org/10.1002/ail2.55. 2.5, 4.1, 4.1, 4.3.2, 4.3.2, 4.3.3, 4.3.4, 4.6, 8.1

[138] Ralph Hertwig and Gerd Gigerenzer. The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of behavioral decision making*, 12(4):275–305, 1999. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/(sici)1099-0771(199912)12:4%3C275::aid-bdm323%3E3.0.co;2-m. 3.5.1

[139] Marius M Hoeper, Christine Pausch, Karen M Olsson, Doerte Huscher, David Pittrow, Ekkehard Grünig, Gerd Staehler, Carmine Dario Vizza, Henning Gall, Oliver Distler, et al. Compera 2.0: a refined four-stratum risk assessment model for pulmonary arterial hypertension. *European Respiratory Journal*, 60(1), 2022. URL https://publications.ersnet.org/content/erj/60/1/2102311.abstract. 6.2

[140] Elfi Inez Saïda Hofmeijer, Sheng-Chih Wu, Rozemarijn Vliegenthart, Cornelis Herman Slump, Ferdi van der Heijden, and Can Ozan Tan. Artificial ct images can enhance variation of case images in diagnostic radiology skills training. *Insights into Imaging*, 14(1):186, 2023. URL https://link.springer.com/content/pdf/10.1186/s13244-023-01508-4.pdf. 8.2.3

[141] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019. URL https://doi.org/10.1145/3290605.3300809. 1

[142] Kenneth Holstein and Vincent Aleven. Designing for human–ai complementarity in k-12 education. *AI Magazine*, 43(2):239–248, 2022. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/7399. 3.5.1

[143] Judith A Holton. The coding process and its challenges. *The Sage handbook of grounded theory*, 3:265–289, 2007. URL https://doi.org/10.4135/9781848607941.n13. 4.3.6

[144] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26, 2020. URL https://doi.org/10.1145/3392878. 5.1

[145] Eric Horvitz. On the horizon: Interactive and compositional deepfakes. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 653–661, 2022. URL https://dl.acm.org/doi/abs/10.1145/3536221.3558175. 7.5.3

[146] Benjamin Hou, Georgios Kaissis, Ronald M Summers, and Bernhard Kainz. Ratchet: Medical transformer for chest x-ray diagnosis and reporting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 293–303. Springer, 2021. URL https://doi.org/10.1007/978-3-030-87234-2_28. 1, 4.3.1

[147] Huiling Huang, Stephanie Q Liu, and Zhi Lu. When and why language assertiveness affects online review persuasion. *Journal of Hospitality & Tourism Research*, 47(6):988–1016, 2023. URL https://doi.org/10.1177/10963480221074280. 4.2

[148] Marc Humbert, Gabor Kovacs, Marius M Hoeper, Roberto Badagliacca, Rolf MF Berger, Margarita Brida, Jørn Carlsen, Andrew JS Coats, Pilar Escribano-Subias, Pisana Ferrari, et al. 2022 esc/ers guidelines for the diagnosis and treatment of pulmonary hypertension: Developed by the task force for the diagnosis and treatment of pulmonary hypertension of the european society of cardiology (esc) and the european respiratory society (ers). endorsed by the international society for heart and lung transplantation (ishlt) and the european reference network on rare respiratory diseases (ern-lung). *European heart journal*, 43(38):3618–3731, 2022. URL https://publications.ersnet.org/content/erj/61/1/2200879. 6.2, 6.6.1.2

[149] Christina Humer, Andreas Hinterreiter, Benedikt Leichtmann, Martina Mara, and Marc Streit. Comparing effects of attribution-based, example-based, and feature-based explanation methods

on ai-assisted decision-making. *OSF Preprints*, 2, 2022. URL https://www.researchgate.net/publication/364581372_Comparing_Effects_of_Attribution-based_Example-based_and_Feature-based_Explanation_Methods_on_AI-Assisted_Decision-Making. 1, 2.2, 4.3.2, 6.1

[150] Christina Humer, Andreas Hinterreiter, Benedikt Leichtmann, Martina Mara, and Marc Streit. Reassuring, misleading, debunking: Comparing effects of xai methods on human decisions. *ACM Transactions on Interactive Intelligent Systems*, 14(3):1–36, 2024. URL https://dl.acm.org/doi/full/10.1145/3665647. 1, 6.1, 6.6.1, 8.1

[151] David Avedis Injarabian, Joonas Ariva, Hendrik Šuvalov, and Dmytro Fishman. Interpreting ct-scans with clip: An explorative study of attribution methods for 3d vision-language models. In *Workshop on Clinical Image-Based Procedures*, pages 87–96. Springer, 2025. URL https://link.springer.com/chapter/10.1007/978-3-032-05479-1_10. 2.2

[152] Mir Riyanul Islam, Mobyen Uddin Ahmed, Shaibal Barua, and Shahina Begum. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, 12(3):1353, 2022. URL https://www.mdpi.com/2076-3417/12/3/1353. 1

[153] Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, 2021. URL https://doi.org/10.1145/3442188.3445901. 1

[154] Myeongjun Jang, Bodhisattwa Prasad Majumder, Julian McAuley, Thomas Lukasiewicz, and Oana-Maria Camburu. Know how to make up your mind! adversarially detecting and alleviating inconsistencies in natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 540–553, 2023. URL https://doi.org/10.18653/v1/2023.acl-short.47. 2.3, 4.1

[155] Mohammad Hossein Jarrahi, Sarah Kenyon, Ashley Brown, Chelsea Donahue, and Chris Wicher. Artificial intelligence: A strategy to harness its power through organizational learning. *Journal of Business Strategy*, 44(3):126–135, 2023. URL https://doi.org/10.1108/jbs-11-2021-0182. 4.5

[156] Neal Jean, Marshall Burke, Michael Xie, W Matthew Alampay Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016. doi: 10.1126/science.aaf7894. URL https://www.science.org/doi/full/10.1126/science.aaf7894. 3.2

[157] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Text-to-image models for counterfactual explanations: a black-box approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4757–4767, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Jeanneret_Text-to-Image_Models_for_Counterfactual_Explanations_A_Black-Box_Approach_WACV_2024_paper.html. 2.5

[158] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. Euca: the end-user-centered explainable ai framework, 2021. URL http://arxiv.org/abs/2102.02437. 2.2, 2.4, 5.1, 6.1, 6.6.1.1

[159] Khushi Kalasampath, K. N. Spoorthi, Sreeparvathy Sajeev, Sahil Sarma Kuppa, Kavya Ajay, and Angulakshmi Maruthamuthu. A literature review on applications of explainable artificial intelligence (xai). *IEEE Access*, 13:41111–41140, 2025. doi: 10.1109/ACCESS.2025.3546681. URL https://ieeexplore.ieee.org/document/10908240. 1

[160] Hyeonsu B Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. Synergi: A mixed-initiative system for scholarly synthesis and sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–19, 2023. URL https://dl.acm.org/doi/abs/10.1145/3586183.3606759. 2.6

[161] Manreet K Kanwar, Mardi Gomberg-Maitland, Marius Hoeper, Christine Pausch, David Pittrow, Geoff Strange, James J Anderson, Carol Zhao, Jacqueline V Scott, Marek J Druzdzel, et al. Risk stratification in pulmonary arterial hypertension using bayesian analysis. *European Respiratory Journal*, 56(2), 2020. URL https://publications.ersnet.org/content/erj/56/2/2000008.abstract. 6.4.1, 6.4.2.1

[162] Mert Karabacak, Burak Berksu Ozkara, Konstantinos Margetis, Max Wintermark, and Sotirios Bisdas. The advent of generative language models in medical education. *JMIR Medical Education*, 9:e48163, 2023. URL https://mededu.jmir.org/2023/1/e48163. 7.5

[163] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022. URL https://dl.acm.org/doi/full/10.1145/3527848. 6.1

[164] Lena Kästner, Markus Langer, Veronika Lazar, Astrid Schomäcker, Timo Speith, and Sarah Sterz. On the relation of trust and explainability: Why to engineer for trustworthiness. In *2021 IEEE 29th international requirements engineering conference workshops (REW)*, pages 169–175. IEEE, 2021. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9582305. 2.2

[165] 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 *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020. URL https://doi.org/10.1145/3313831.3376219. 1, 2.2, 3.5.1, 3.5.2.1

[166] Maxime Kayser, Cornelius Emde, Oana-Maria Camburu, Guy Parsons, Bartlomiej Papiez, and Thomas Lukasiewicz. Explaining chest x-ray pathologies in natural language. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 701–713. Springer, 2022. URL https://doi.org/10.1007/978-3-031-16443-9_67. 4.1, 4.3.1

[167] Maxime Kayser, Bayar Menzat, Cornelius Emde, Bogdan Bercean, Alex Novak, Abdala Espinosa, Bartlomiej W Papiez, Susanne Gaube, Thomas Lukasiewicz, and Oana-Maria Camburu. Fool me once? contrasting textual and visual explanations in a clinical decision-support setting. *arXiv preprint arXiv:2410.12284*, 2024. URL https://arxiv.org/abs/2410.12284. 6.6.2

[168] Majeed Kazemitabaar, Xinying Hou, Austin Henley, Barbara Jane Ericson, David Weintrop, and Tovi Grossman. How novices use llm-based code generators to solve cs1 coding tasks in a self-paced learning environment. In *Proceedings of the 23rd Koli calling international conference on computing education research*, pages 1–12, 2023. URL https://arxiv.org/abs/2309.14049. 4.3.2

[169] Tahira Kazimi, Ritika Allada, and Pinar Yanardag. Explaining in diffusion: Explaining a classifier through hierarchical semantics with text-to-image diffusion models. *arXiv preprint arXiv:2412.18604*, 2024. URL https://arxiv.org/abs/2412.18604. 2.5

[170] Bernard Keenan and Kacper Sokol. Mind the gap! bridging explainable artificial intelligence and human understanding with luhmann's functional theory of communication. *arXiv preprint arXiv:2302.03460*, 2023. URL https://arxiv.org/abs/2302.03460. 6.6.1.1

[171] Frank C Keil. Explanation and understanding. *Annu. Rev. Psychol.*, 57(1):227–254, 2006. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3034737/#S4. 1.2, 8.1

[172] Asim B Khajwal and Arash Noshadravan. An uncertainty-aware framework for reliable disaster damage assessment via crowdsourcing. *International Journal of Disaster Risk Reduction*, 55:

102110, 2021. URL https://www.sciencedirect.com/science/article/pii/S2212420921000765. 3.2.1

[173] Dajung Kim, Niko Vegt, Valentijn Visch, and Marina Bos-De Vos. How much decision power should (a) i have?: Investigating patients' preferences towards ai autonomy in healthcare decision making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024. URL https://dl.acm.org/doi/full/10.1145/3613904.3642883. 6.6.3.1

[174] Doha Kim, Yeosol Song, Songyie Kim, Sewang Lee, Yanqin Wu, Jungwoo Shin, and Daeho Lee. How should the results of artificial intelligence be explained to users?-research on consumer preferences in user-centered explainable artificial intelligence. *Technological Forecasting and Social Change*, 188:122343, 2023. URL https://doi.org/10.1016/j.techfore.2023.122343. 4.6

[175] Ja Ae Kim, Kenneth Mascola, and Yoan K Kagoma. Re-envisioning on-call resident workflows: Impact on resident experience, 2021. URL https://www.rsna.org/-/media/Files/RSNA/Practice-Tools/Quality-improvement/Quality-improvement-reports/2021/Re-Envisioning_On-Call_Resident-QDP-QI-20_RSNA_.ashx. 5.4

[176] Minjung Kim, Saebyeol Kim, Jinwoo Kim, Tae-Jin Song, and Yuyoung Kim. Do stakeholder needs differ? designing stakeholder-tailored explainable artificial intelligence (xai) interfaces. *International Journal of Human-Computer Studies*, 181:103160, 2024. URL https://www.sciencedirect.com/science/article/pii/S1071581923001696. 2.4, 6.6.1.1

[177] Sunnie SY 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 *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023. URL https://doi.org/10.1145/3544548.3581001. 2.2, 4.1, 4.3.1, 4.3.2, 4.5, 4.6, 6.1, 6.6.1.3

[178] Taenyun Kim and Hayeon Song. The effect of message framing and timing on the acceptance of artificial intelligence's suggestion. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020. URL https://doi.org/10.1145/3334480.3383038. 4.5

[179] Taenyun Kim and Hayeon Song. Communicating the limitations of ai: the effect of message framing and ownership on trust in artificial intelligence. *International Journal of Human–Computer Interaction*, 39(4):790–800, 2023. URL https://doi.org/10.1080/10447318.2022.2049134. 4.1, 4.5

[180] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018. URL https://doi.org/10.3386/w23180. 3.5.1

[181] 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 *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019. URL https://doi.org/10.1145/3290605.3300641. 2.3, 4.1, 4.2, 4.5

[182] Virve Koljonen. What could we make of ai in plastic surgery education. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, 81:94–96, 2023. URL https://www.jprasurg.com/article/S1748-6815(23)00205-X/fulltext. 2.5

[183] Mohamad Koohi-Moghadam and Kyongtae Ty Bae. Generative ai in medical imaging: applications, challenges, and ethics. *Journal of Medical Systems*, 47(1):94, 2023. URL https://link.springer.com/article/10.1007/s10916-023-01987-4. 7.5.3

[184] Brett Koonce. Efficientnet. In *Convolutional neural networks with swift for Tensorflow:*

image recognition and dataset categorization, pages 109–123. Springer, 2021. URL https://link.springer.com/chapter/10.1007/978-1-4842-6168-2_10. 5.3.2.2

[185] Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5686–5697, 2016. URL https://dl.acm.org/doi/pdf/10.1145/2858036.2858529. 1.1, 6.3.2.2

[186] Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022. URL https://arxiv.org/abs/2202.01602. 1.1, 6.6.2, 8.2.1

[187] Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. Are large language models post hoc explainers? *arXiv preprint arXiv:2310.05797*, 2023. URL https://arxiv.org/abs/2310.05797v1. 2.3, 4.1

[188] Ann Kronrod, Amir Grinstein, and Kerem Shuval. Think positive! emotional response to assertiveness in positive and negative language promoting preventive health behaviors. *Psychology & health*, 37(11):1309–1326, 2022. URL https://doi.org/10.1080/08870446.2021.1942876. 4.2

[189] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*, 2023. URL https://arxiv.org/abs/2312.14867v1. 7.1

[190] Ajay Kumar, Pierce Burr, Tim Michael Young, et al. Using ai text-to-image generation to create novel illustrations for medical education: current limitations as illustrated by hypothyroidism and horner syndrome. *JMIR Medical Education*, 10(1):e52155, 2024. URL https://mededu.jmir.org/2024/1/e52155/. 2.5

[191] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International conference on machine learning*, pages 5491–5500. PMLR, 2020. URL https://proceedings.mlr.press/v119/kumar20e.html. 2.2

[192] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: A survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021. URL https://arxiv.org/abs/2112.11471. 3.5.1

[193] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017. URL https://doi.org/10.1017/S0140525X16001837. 3.5.1

[194] Himabindu Lakkaraju and Osbert Bastani. "how do i fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85, 2020. URL https://doi.org/10.1145/3375627.3375833. 3.5.1

[195] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. URL https://doi.org/10.1609/aaai.v31i1.10821. 3.5.1

[196] Michelle S Lam, Mitchell L Gordon, Danaë Metaxa, Jeffrey T Hancock, James A Landay, and Michael S Bernstein. End-user audits: A system empowering communities to lead large-scale investigations of harmful algorithmic behavior. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–34, 2022. URL https://doi.org/10.1145/3555625. 5.4

[197] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim,

William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Lang_Explaining_in_Style_Training_a_GAN_To_Explain_a_Classifier_ICCV_2021_paper.pdf. 2.5, 8.2.3

[198] Christine P Lee, Min Kyung Lee, and Bilge Mutlu. The ai-dec: A card-based design method for user-centered ai explanations. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pages 1010–1028, 2024. URL https://dl.acm.org/doi/abs/10.1145/3643834.3661576. 6.1, 6.6.1.2, 6.6.1.3

[199] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004. URL https://journals.sagepub.com/doi/abs/10.1518/hfes.46.1.50_30392. 1

[200] Min Hun Lee and Chong Jun Chew. Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–22, 2023. URL https://dl.acm.org/doi/abs/10.1145/3610218. 1, 6.1, 8.1

[201] Min Hun Lee, Silvana Xin Yi Choo, Shamala D Thilarajah, et al. Improving health professionals' onboarding with ai and xai for trustworthy human-ai collaborative decision making. *arXiv preprint arXiv:2405.16424*, 2024. URL https://arxiv.org/abs/2405.16424. 2.2

[202] Yeonjoo Lee, Miyeon Ha, Sujeong Kwon, Yealin Shim, and Jinwoo Kim. Egoistic and altruistic motivation: How to induce users' willingness to help for imperfect ai. *Computers in Human Behavior*, 101:180–196, 2019. URL https://doi.org/10.1016/j.chb.2019.06.009. 4.2

[203] Benedikt Leichtmann, Andreas Hinterreiter, Christina Humer, Marc Streit, and Martina Mara. Explainable artificial intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival. *International Journal of Human–Computer Interaction*, pages 1–18, 2023. URL https://doi.org/10.31219/osf.io/68emr. 4.2

[204] Brett Levac, Sidharth Kumar, Ajil Jalal, and Jonathan I Tamir. Accelerated motion correction with deep generative diffusion models. *Magnetic Resonance in Medicine*, 92(2):853–868, 2024. URL https://onlinelibrary.wiley.com/doi/full/10.1002/mrm.30082. 7.1

[205] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021. URL https://dl.acm.org/doi/abs/10.1145/3411764.3445522. 4.2

[206] Jiachen Li, Mandi Hall, Ruican Zhong, Selin S. Everett, Alyssa Unell, Hanwen Xu, Matthias Blondeel, Jonathan M. Carlson, Katie Claveau, Thulasee Jose, Tristan Naumann, David C. Rhew, Naiteek Sangani, Frank Tuan, James Weinstein, Varun Mishra, Elizabeth D. Mynatt, Scott Saponas, Leonardo Schettini, Joseph S. Preston, Aiden Gu, Naoto Usuyama, Zelalem Gero, Cliff Wong, Noel C. Codella, Hoifung Poon, Shrey Jain, Matthew P Lungren, and Eric Horvitz. Exploring the future of ai in clinical collaboration: A study on tumor board case preparation. In *ACM CHI Conference on Human Factors in Computing Systems (CHI 2026)*, April 2026. URL https://www.microsoft.com/en-us/research/publication/exploring-the-future-of-ai-in-clinical-collaboration-a-study-on-tumor-board-case-preparation/. 9.2.2

[207] Rena Li, Sara Kingsley, Chelsea Fan, Proteeti Sinha, Nora Wai, Jaimie Lee, Hong Shen, Motahhare Eslami, and Jason Hong. Participation and division of labor in user-driven

algorithm audits: How do everyday users work together to surface algorithmic harms? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023. URL https://dl.acm.org/doi/full/10.1145/3544548.3582074. 5.4

[208] Yuheng Li, Yuxiang Lai, Maria Thor, Deborah Marshall, Zachary Buchwald, David S Yu, and Xiaofeng Yang. Towards universal text-driven ct image segmentation. *arXiv preprint arXiv:2503.06030*, 2025. URL https://arxiv.org/pdf/2503.06030. 9.2.3

[209] Q Vera Liao and Kush R Varshney. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*, 2021. URL http://arxiv.org/abs/2110.10790. 2.2, 3.5.2.1, 4.1

[210] Q Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941*, 10, 2023. URL https://doi.org/10.1162/99608f92.8036d03b. 5.1

[211] Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: Informing design practices for explainable ai user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, April 2020. doi: 10.1145/3313831.3376590. URL http://arxiv.org/abs/2001.02478. 2.1, 3.1, 5.1

[212] Q Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. Question-driven design process for explainable ai user experiences. *arXiv preprint arXiv:2104.03483*, 2021. URL http://arxiv.org/abs/2104.03483. 2.2, 5.1, 5.2.2, 5.2.2, 5.4

[213] Q Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. Designerly understanding: Information needs for model transparency to support design ideation for ai-powered user experience. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–21, 2023. URL https://doi.org/10.1145/3544548.3580652. 5.1

[214] Bo Lin, Yingjing Xu, Xuanwen Bao, Zhou Zhao, Zhouyang Wang, and Jianwei Yin. Skingen: An explainable dermatology diagnosis-to-generation framework with interactive vision-language models. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1287–1296, 2025. URL https://dl.acm.org/doi/10.1145/3708359.3712098. 8.2.3

[215] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024. URL https://link.springer.com/chapter/10.1007/978-3-031-72673-6_20. 7.1

[216] Martin Lindvall, Claes Lundström, and Jonas Löwgren. Rapid assisted visual search: Supporting digital pathologists with imperfect ai. In *26th International Conference on Intelligent User Interfaces*, pages 504–513, 2021. URL https://doi.org/10.1145/3397481.3450681. 4.1

[217] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018. URL https://spawn-queue.acm.org/doi/pdf/10.1145/3236386.3241340. 6.1

[218] Zachary Chase Lipton. The mythos of model interpretability. *CoRR*, abs/1606.03490, 2016. URL http://arxiv.org/abs/1606.03490. 3.5.1

[219] Fang Liu, Junyan Lv, Shenglan Cui, Zhilong Luan, Kui Wu, and Tongqing Zhou. Smart "error"! exploring imperfect ai to support creative ideation. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–28, 2024. URL https://dl.acm.org/doi/10.1145/3637398. 8.2.1, 8.5, 9.2.1

[220] Han Liu, Vivian Lai, and Chenhao Tan. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the*

ACM on Human-Computer Interaction, 5(CSCW2):1–45, 2021. doi: 10.1145/3479552. URL https://doi.org/10.1145/3479552. 1, 6.1

[221] Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed H Awadallah. On improving summarization factual consistency from natural language feedback. *arXiv preprint arXiv:2212.09968*, 2022. URL https://arxiv.org/abs/2212.09968v1. 4.3.3

[222] Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10 (10):464–470, 2006. URL https://www.sciencedirect.com/science/article/pii/S1364661306002117. 1.2, 8.1

[223] Jens Ludwig and Sendhil Mullainathan. Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827, 2024. URL https://academic.oup.com/qje/article/139/2/751/7515309?guestAccessKey=. 6.3.2.3

[224] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html. 2.1

[225] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018. URL https://www.nature.com/articles/s41551-018-0304-0. 2.2, 2.4

[226] Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. Teaching categories to human learners with visual explanations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3820–3828, 2018. URL https://doi.org/10.1109/cvpr.2018.00402. 4.5

[227] Michelle Mai, Fatima N Mirza, and Christopher DiMarco. Application of text-to-image translation algorithms in medicine: a systematic review. *JAAD Reviews*, 2:88–96, 2024. URL https://www.sciencedirect.com/science/article/pii/S2950198924000345. 7.1

[228] Bertram F Malle. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press, 2006. 1.2, 9.2.2

[229] Awais Mansoor, Ulas Bagci, Brent Foster, Ziyue Xu, Georgios Z Papadakis, Les R Folio, Jayaram K Udupa, and Daniel J Mollura. Segmentation and image analysis of abnormal lungs at ct: current approaches, challenges, and future trends. *Radiographics*, 35(4):1056–1076, 2015. URL https://pubs.rsna.org/doi/abs/10.1148/rg.2015140232. 9.2.3

[230] Mayo Clinic. Ct scan, 2023. URL https://www.mayoclinic.org/tests-procedures/ct-scan/about/pac-20393675. 7.1, 7.4.3.1

[231] Nathan J McNeese, Beau G Schelble, Lorenzo Barberis Canonico, and Mustafa Demir. Who/what is my teammate? team composition considerations in human–ai teaming. *IEEE Transactions on Human-Machine Systems*, 51(4):288–299, 2021. URL https://doi.org/10.1109/thms.2021.3086018. 5.4

[232] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. ISSN 0004-3702. URL https://www.sciencedirect.com/science/article/pii/S0004370218305988. 2.4, 3.5.1, 6.6.1.1, 8.1

[233] Tim Miller. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 333–342, 2023. URL https://doi.org/10.1145/3593013.3594001. 2.6, 5.1, 8.1, 8.5

[234] Dimitry Mindlin, Fabian Beer, Leonie Nora Sieger, Stefan Heindorf, Elena Esposito, Axel-

Cyrille Ngonga Ngomo, and Philipp Cimiano. Beyond one-shot explanations: a systematic literature review of dialogue-based xai approaches. *Artificial Intelligence Review*, 58(3):81, 2025. URL https://link.springer.com/article/10.1007/s10462-024-11007-7. 1.2, 6.6.1.1

[235] 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 *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019. URL https://doi.org/10.1145/3287560.3287596. 5.4

[236] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021. URL https://dl.acm.org/doi/abs/10.1145/3387166. 2.1, 2.2, 5.1

[237] Daniele Molino, Camillo Maria Caruso, Paolo Soda, and Valerio Guarrasi. Retrieval-augmented anatomical guidance for text-to-ct generation. *arXiv preprint arXiv:2603.08305*, 2026. URL https://arxiv.org/pdf/2603.08305. 9.2.3

[238] Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. Evaluating the impact of human explanation strategies on human-ai visual decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–37, 2023. URL https://doi.org/10.1145/3579481. 4.1, 6.6.1.1

[239] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. The impact of imperfect xai on human-ai decision-making. *Proceedings of the ACM on human-computer interaction*, 8(CSCW1):1–39, 2024. URL https://dl.acm.org/doi/abs/10.1145/3641022. 1.1, 2.5, 6.6.2, 8.1, 8.2.1, 8.5

[240] Katelyn Morrison, Arpit Mathur, Aidan Bradshaw, Tom Wartmann, Steven Lundi, Afrooz Zandifar, Weichang Dai, Kayhan Batmanghelich, Motahhare Eslami, and Adam Perer. A human-centered approach to identifying promises, risks, & challenges of text-to-image generative ai in radiology. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 1758–1770, 2025. URL https://ojs.aaai.org/index.php/AIES/article/view/36672/38810. 8.1, 8.2.1, 8.2.3, 8.3.1

[241] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020. URL https://dl.acm.org/doi/abs/10.1145/3351095.3372850. 2.1, 6.3.1.1

[242] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5323–5331, 2022. URL https://doi.org/10.1609/aaai.v36i5.20469. 5.4, 6.3.2.1

[243] Hussein Mozannar, Jimin J Lee, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Effective human-ai teams via learned natural language rules and onboarding. *Advances in neural information processing systems*, 36:30466–30498, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/61355b9c218505505d1bededede9da56b2-Paper-Conference.pdf. 5.4

[244] Imani Munyaka, Zahra Ashktorab, Casey Dugan, J Johnson, and Qian Pan. Decision making strategies and team efficacy in human-ai teams. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–24, 2023. URL https://doi.org/10.1145/3579476. 5.4

[245] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 786–808,

2023. URL https://dl.acm.org/doi/abs/10.1145/3600211.3604711. 7.1

[246] Mohammad Naiseh. Social explainable ai (social xai): Towards expanding the social benefits of xai. In *The Impact of Artificial Intelligence on Societies: Understanding Attitude Formation Towards AI*, pages 169–178. Springer, 2024. URL https://link.springer.com/chapter/10.1007/978-3-031-70355-3_13. 6.6.1.2, 6.6.1.3

[247] Catherine Nasrallah, Cherish Wilson, Alicia Hamblin, Christine Hariz, Cammie Young, Jing Li, Jinoos Yazdany, and Gabriela Schmajuk. Patient perceptions of an electronic-health-record-based rheumatoid arthritis outcomes dashboard: a mixed-methods study. *BMC Medical Informatics and Decision Making*, 24(1):1–15, 2024. URL https://link.springer.com/article/10.1186/s12911-024-02696-9. 6.5.2.2

[248] Catherine Nasrallah, Cherish Wilson, Alicia Hamblin, Cammie Young, Lindsay Jacobsohn, Mary C Nakamura, Andrew Gross, Mehrdad Matloubian, Judith Ashouri, Jinoos Yazdany, et al. Using the technology acceptance model to assess clinician perceptions and experiences with a rheumatoid arthritis outcomes dashboard: qualitative study. *BMC Medical Informatics and Decision Making*, 24(1):140, 2024. URL https://link.springer.com/article/10.1186/s12911-024-02530-2. 6.5.2.2

[249] Purushothaman Natarajan and Athira Nambiar. Vale: A multimodal visual and language explanation framework for image classifiers using explainable ai and language models. *arXiv preprint arXiv:2408.12808*, 2024. URL https://arxiv.org/pdf/2408.12808. 1.1, 2.5, 8.1

[250] Mohd Amril Nazir, Edmund Evangelista, Syed M Salman Bukhari, and Ravishankar Sharma. A survey of feature attribution techniques in explainable ai: Taxonomy, analysis and comparison. *Annals of Mathematics and Computer Science*, 28:115–126, 2025. URL https://annalsmcs.org/index.php/amcs/article/view/547/220. 2.1

[251] Joshua Nelson. World's first outdoor glass elevator: The el cortez hotel, 2019. URL https://www.elevatorscenestudio.com/blog/el-cortez-glass-elevator. 7

[252] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems*, 34:26422–26436, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/de043a5e421240eb846da8effe472ff1-Abstract.html. 4.3.4

[253] Giang Nguyen, Mohammad Reza Taesiri, and Anh Nguyen. Visual correspondence-based explanations improve ai robustness and human-ai team accuracy. *Advances in Neural Information Processing Systems*, 35:34287–34301, 2022. URL https://arxiv.org/abs/2208.00780. 4.3.1

[254] Truong Thanh Hung Nguyen, Tobias Clement, Phuc Truong Loc Nguyen, Nils Kemmerzell, Van Binh Truong, Vo Thanh Khang Nguyen, Mohamed Abdelaal, and Hung Cao. Langxai: Integrating large vision models for generating textual explanations to enhance explainability in visual perception tasks. *arXiv preprint arXiv:2402.12525*, 2024. URL https://arxiv.org/abs/2402.12525. 2.5, 8.1

[255] Robert Nimmo, Marios Constantinides, Ke Zhou, Daniele Quercia, and Simone Stumpf. User characteristics in explainable ai: The rabbit hole of personalization? In *Proceedings of the 2024 CHI conference on human factors in computing systems*, pages 1–13, 2024. URL https://dl.acm.org/doi/pdf/10.1145/3613904.3642352. 1.1

[256] Geoffroy PJC Noel. Evaluating ai-powered text-to-image generators for anatomical illustration: a comparative study. *Anatomical Sciences Education*, 17(5):979–983, 2024. URL https://anatomypubs.onlinelibrary.wiley.com/doi/full/10.1002/ase.2336. 2.5, 7.1

[257] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019. URL https://arxiv.org/abs/1909.09223. 1

[258] Geoffrey Norman. Research in clinical reasoning: past history and current trends. *Medical education*, 39(4):418–427, 2005. URL https://asmepublications.onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2929.2005.02127.x. 6.6, 6.5, 6.6.1.3

[259] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 97–105, 2019. URL https://doi.org/10.1609/hcomp.v7i1.5284. 2.3

[260] Mahsan Nourani, Joanie King, and Eric Ragan. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 112–121, 2020. URL https://doi.org/10.1609/hcomp.v8i1.7469. 2.2, 4.5

[261] Jeroen Ooge and Katrien Verbert. Trust in prediction models: a mixed-methods pilot study on the impact of domain expertise. In *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pages 8–13. IEEE, 2021. URL https://doi.org/10.1109/trex53765.2021.00007. 2.2

[262] Keng-Boon Ooi, Garry Wei-Han Tan, Mostafa Al-Emran, Mohammed A Al-Sharafi, Alexandru Capatina, Amrita Chakraborty, Yogesh K Dwivedi, Tzu-Ling Huang, Arpan Kumar Kar, Voon-Hsien Lee, et al. The potential of generative artificial intelligence across disciplines: Perspectives and future directions. *Journal of Computer Information Systems*, 65(1):76–107, 2025. URL https://www.tandfonline.com/doi/full/10.1080/08874417.2023.2261010. 7.1, 7.5

[263] Barak Oshri, Annie Hu, Peter Adelson, Xiao Chen, Pascaline Dupas, Jeremy Weinstein, Marshall Burke, David Lobell, and Stefano Ermon. Infrastructure quality assessment in africa using satellite imagery and deep learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 616–625, 2018. URL https://dl.acm.org/doi/abs/10.1145/3219819.3219924. 3.2

[264] António C Pacheco and Carlos Martinho. Alignment of player and non-player character assertiveness levels. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 15, pages 181–187, 2019. URL https://doi.org/10.1609/aiide.v15i1.5242. 4.3.4

[265] Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018. URL https://doi.org/10.1016/j.jbef.2017.12.004. 4.3.5

[266] Jörg Papenkordt. Navigating transparency: The influence of on-demand explanations on non-expert user interaction with ai. In *International Conference on Human-Computer Interaction*, pages 238–263. Springer, 2024. URL https://link.springer.com/chapter/10.1007/978-3-031-60606-9_14. 2.2, 8.2.1

[267] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652*, 2019. URL https://arxiv.org/abs/1907.12652. 1.1, 2.3, 4.1, 4.2, 4.5, 4.6

[268] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine*, 2 (1):1–10, 2019. URL https://doi.org/10.1038/s41746-019-0189-7. 3.1

[269] Avery B Paxton, Erica Blair, Camryn Blawas, Michael H Fatzinger, Madeline Marens, Jason Holmberg, Colin Kingen, Tanya Houppermans, Mark Keusenkothen, John McCord, et al. Citizen science reveals female sand tiger sharks (carcharias taurus) exhibit signs of site fidelity

on shipwrecks. *Ecology*, 100(8):1–4, 2019. URL https://doi.org/10.1002/ecy.2687. 4.1

[270] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017. URL https://doi.org/10.1016/j.jesp.2017.01.006. 4.3.5

[271] Fernando Pérez-García, Sam Bond-Taylor, Pedro P Sanchez, Boris van Breugel, Daniel C Castro, Harshita Sharma, Valentina Salvatelli, Maria TA Wetscherek, Hannah Richardson, Matthew P Lungren, et al. Radedit: stress-testing biomedical vision models via diffusion image editing. In *European Conference on Computer Vision*, pages 358–376. Springer, 2024. URL https://link.springer.com/chapter/10.1007/978-3-031-73254-6_21. 7.1

[272] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, volume 5, pages 2–4. McLean, VA, USA, 2005. URL https://andymatuschak.org/files/papers/Pirolli,%20Card%20-%202005%20-%20The%20sensemaking%20process%20and%20leverage%20points%20for%20analyst%20technology%20as.pdf. 6.6, 6.5, 6.6.1.2

[273] Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. Finding and fixing spurious patterns with explanations. *arXiv preprint arXiv:2106.02112*, 2021. URL https://arxiv.org/abs/2106.02112. 4.3.3

[274] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021. URL https://doi.org/10.1145/3411764.3445315. 3.5.1

[275] Nicoletta Prentzas, Antonis Kakas, and Constantinos S Pattichis. Explainable ai applications in the medical domain: A systematic review. *arXiv preprint arXiv:2308.05411*, 2023. URL https://arxiv.org/abs/2308.05411. 2.1, 6.1

[276] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826, 2022. URL https://doi.org/10.1145/3531146.3533231. 5.4

[277] Emilio Quaia. Electronic platform-based education for radiology residents: results of a two-year survey. *BMC Medical Education*, 23(1):202, 2023. ISSN 1472-6920. doi: 10.1186/s12909-023-04190-8. URL https://doi.org/10.1186/s12909-023-04190-8. 7.5.2

[278] Marissa Radensky, Doug Downey, Kyle Lo, Zoran Popovic, and Daniel S Weld. Exploring the role of local and global explanations in recommender systems. In *CHI conference on human factors in computing systems extended abstracts*, pages 1–7. ACM New York, NY, USA, 2022. ISBN 9781450391566. doi: 10.1145/3491101.3519795. URL https://doi.org/10.1145/3491101.3519795. 1, 6.1

[279] Radiopaedia.org. Radiopaedia, 2025. URL https://radiopaedia.org. 7.2.2.1

[280] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html?ref=journey. 2.5

[281] Kate Rassie. The apprenticeship model of clinical medical education: time for structural change. *The New Zealand Medical Journal (Online)*, 130(1461):66, 2017. URL https://nzmj.org.nz/media/pages/journal/vol-130-no-1461/e9dcfee880-1696469524/vol-130-no-1461.pdf#page=66. 7.5.2

[282] Johan Redström. Towards user design? on the shift from object to user as the subject of design. *Design studies*, 27(2):123–139, 2006. URL https://www.sciencedirect.com/science/article/pii/S0142694X05000359. 1.1

[283] Leon Reicherts, Zelun Tony Zhang, Elisabeth Von Oswald, Yuanting Liu, Yvonne Rogers, and Mariam Hassib. Ai, help me think—but for myself: Assisting people in complex decision-making by providing different kinds of cognitive support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2025. URL https://dl.acm.org/doi/full/10.1145/3706598.3713295. 2.6

[284] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: Artificial Intelligence*, 2(3):e190043, 2020. URL https://doi.org/10.1148/ryai.2020190043. 3.6

[285] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, August 2016. URL https://doi.org/10.1145/2939672.2939778. 1, 2.1, 2.2

[286] Lara Riefle, Patrick Hemmer, Carina Benz, Michael Vössing, and Jannik Pries. On the influence of cognitive styles on users' understanding of explanations. In *Proceedings of the Forty-Third International Conference on Information Systems (ICIS)*, 2022. URL https://arxiv.org/abs/2210.02123. 4.3.2

[287] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. Understanding the role of explanation modality in ai-assisted decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 223–233, 2022. URL https://doi.org/10.1145/3503252.3531311. 4.2, 4.3.2

[288] Tabea E Röber, Rob Goedhart, and Sİ Birbil. Clinicians' voice: Fundamental considerations for xai in healthcare. *arXiv preprint arXiv:2411.04855*, 2024. URL https://arxiv.org/abs/2411.04855. 6.6.3.1

[289] Katharina J Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M Buhl, Hendrik Buschmeier, Elena Esposito, Angela Grimminger, Barbara Hammer, Reinhold Häb-Umbach, et al. Explanation as a social practice: Toward a conceptual framework for the social design of ai systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13(3): 717–728, 2020. URL https://ieeexplore.ieee.org/abstract/document/9292993. 1.2, 6.6.1.1

[290] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence*, 46(4):2104–2122, 2023. URL https://ieeexplore.ieee.org/abstract/document/10316181. 1.1, 2.1, 2.2

[291] Emely Rosbach, Jonas Ammeling, Sebastian Krügel, Angelika Kießig, Alexis Fritz, Jonathan Ganz, Chloé Puget, Taryn Donovan, Andrea Klang, Maximilian C Köller, et al. "when two wrongs don't make a right"-examining confirmation bias and the role of time pressure during human-ai collaboration in computational pathology. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2025. URL https://dl.acm.org/doi/full/10.1145/3706598.3713319. 7.5.3

[292] Saumendu Roy, Gabriel Laberge, Banani Roy, Foutse Khomh, Amin Nikanjam, and Saikat Mondal. Why don't xai techniques agree? characterizing the disagreements between post-hoc explanations of defect predictions. In *2022 IEEE international conference on software*

maintenance and evolution (ICSME), pages 444–448. IEEE, 2022. URL https://doi.org/10.1109/icsme55016.2022.00056. 1.1, 2.3, 4.6, 8.2.1

[293] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 269–276, 1993. URL https://dl.acm.org/doi/abs/10.1145/169059.169209. 6.6, 6.5, 6.6.1.2

[294] Mahya Sadeghi, Parmit K Chilana, and M Stella Atkins. How users perceive content-based image retrieval for identifying skin images. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 1*, pages 141–148. Springer, 2018. URL https://doi.org/10.1007/978-3-030-02628-8_16. 4.1, 4.3.2

[295] Conrad W Safranek, Anne Elizabeth Sidamon-Eristoff, Aidan Gilson, and David Chartash. The role of large language models in medical education: applications and implications, 2023. URL https://mededu.jmir.org/2023/1/e50945. 2.5

[296] Johnny Saldaña. *The coding manual for qualitative researchers*. SAGE publications Ltd, 2021. 8.3.4

[297] Tracy L. Sanders, Alexandra D. Kaplan, Keith MacArthur, William G. Volante, and P. A. Hancock. Trust and human factors: Foundations of trust in automation. In Frank Krueger, editor, *The Neurobiology of Trust*, pages 77–98. Cambridge University Press, 2021. 6.1

[298] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G. Blankenberg, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *medRxiv*, Mar 2021. doi: 10.1101/2021.02.28.21252634v1. URL https://www.medrxiv.org/content/10.1101/2021.02.28.21252634v1.full-text. 3.1

[299] 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 *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 617–626, 2022. URL https://doi.org/10.1145/3514094.3534128. 4.1, 4.2, 4.2

[300] Max Schemmer, Andrea Bartos, Philipp Spitzer, Patrick Hemmer, Niklas Kühl, Jonas Liebschner, and Gerhard Satzger. Towards effective human-ai decision-making: The role of human learning in appropriate reliance on ai advice. *arXiv preprint arXiv:2310.02108*, 2023. URL https://arxiv.org/abs/2310.02108. 4.5

[301] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 410–422, 2023. URL https://doi.org/10.1145/3581641.3584066. (document), 4.1, 4.2, 4.1, 4.3.6, 4.3.6, 4.4, 4.4.2.3, 4.5

[302] Johannes Schneider. Explainable generative ai (genxai): A survey, conceptualization, and research agenda. *Artificial Intelligence Review*, 57(11):289, 2024. URL https://link.springer.com/article/10.1007/s10462-024-10916-x. 1.1, 2.5, 8.1, 8.2.3, 8.2.3

[303] Jakob Schoeffer, Johannes Jakubik, Michael Voessing, Niklas Kuehl, and Gerhard Satzger. On the interdependence of reliance behavior and accuracy in ai-assisted decision-making. In *HHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*, pages 46–59. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2023. URL https://doi.org/10.3233/faia230074. 4.2

[304] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. On explanations, fairness, and appropriate reliance in human-ai decision-making. *ACM Conference on Human Factors in Computing Systems (CHI)*, 2024. URL https://arxiv.org/abs/2209.11812. 4.2

[305] Tjeerd AJ Schoonderwoerd, Wiard Jorritsma, Mark A Neerincx, and Karel Van Den Bosch. Human-centered xai: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, 154:102684, 2021. URL https://www.sciencedirect.com/science/article/pii/S1071581921001026. 2.2

[306] Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38(5):3043–3101, 2023. ISSN 1384-5810. doi: 10.1007/s10618-022-00867-8. URL https://doi.org/10.1007/s10618-022-00867-8. 1, 5.1

[307] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. URL https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf. 1, 5.1

[308] Julian Senoner, Simon Schallmoser, Bernhard Kratzwald, Stefan Feuerriegel, and Torbjørn Netland. Explainable ai improves task performance in human-ai collaboration. *arXiv preprint arXiv:2406.08271*, 2024. URL https://arxiv.org/abs/2406.08271. 6.3.2.2

[309] Adrian Serapio, Gunvant Chaudhari, Cody Savage, Yoo Jin Lee, Maya Vella, Shravan Sridhar, Jamie Lee Schroeder, Jonathan Liu, Adam Yala, and Jae Ho Sohn. An open-source fine-tuned large language model for radiological impression generation: a multi-reader performance study. *BMC Medical Imaging*, 24(1):254, 2024. URL https://link.springer.com/article/10.1186/s12880-024-01435-w. 8.1

[310] Samer H Sharkiya. Quality communication can improve patient-centred health outcomes among older patients: a rapid review. *BMC Health Services Research*, 23(1):886, 2023. URL https://link.springer.com/article/10.1186/s12913-023-09869-8. 6.6.1.1

[311] Lucy Shinners, Sandra Grace, Stuart Smith, Alexandre Stephens, and Christina Aggar. Exploring healthcare professionals' perceptions of artificial intelligence: Piloting the shinners artificial intelligence perception tool. *Digital Health*, 8:20552076221078110, 2022. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8832586/. (document), 8.3.1, 8.1, 8.4

[312] Hidehiko Shishido, Koyo Kobayashi, Yoshinari Kameda, and Itaru Kitahara. Method to generate building damage maps by combining aerial image processing and crowdsourcing. *Journal of Disaster Research*, 16(5):827–839, 2021. URL https://www.jstage.jst.go.jp/article/jdr/16/5/16_827/_article/-char/ja/. 3.2.1

[313] Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *International Journal of Human–Computer Interaction*, 39(7):1390–1404, 2023. URL https://doi.org/10.1080/10447318.2022.2101698. 2.2

[314] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013. URL https://arxiv.org/pdf/1312.6034.pdf. 3.1

[315] Ronal Singh, Tim Miller, Liz Sonenberg, Eduardo Velloso, Frank Vetere, Piers Howe, and Paul Dourish. An actionability assessment tool for explainable ai. *arXiv preprint arXiv:2407.09516*, 2024. URL https://arxiv.org/abs/2407.09516. 6.1

[316] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022. URL https:

//openreview.net/forum?id=XVPqLyNxSyh. 5.1

[317] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023. URL https://dl.acm.org/doi/full/10.1145/3544548.3581075. 6.1, 6.5.4, 8.5

[318] Venkatesh Sivaraman, Katelyn Morrison, Will Epperson, and Adam Perer. Over-relying on reliance: Towards realistic evaluations of ai-based clinical decision support. *arXiv preprint arXiv:2504.07423*, 2025. URL https://arxiv.org/abs/2504.07423. 1, 6.1, 7.1, 8.5

[319] Venkatesh Sivaraman, Eric P. Mason, Mengfan Ellen Li, Jessica Tong, Andrew J. King, Jeremy M. Kahn, and Adam Perer. Intelligent reasoning cues: A framework and case study of the roles of AI information in complex decisions. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. ACM. doi: 10.1145/3772318.3790953. URL https://arxiv.org/abs/2602.00259. 2.6

[320] Vera Sorin, Yiftach Barash, Eli Konen, and Eyal Klang. Creating artificial images for radiology applications using generative adversarial networks (gans)–a systematic review. *Academic radiology*, 27(8):1175–1185, 2020. URL https://www.sciencedirect.com/science/article/abs/pii/S1076633220300210. 7.1

[321] Francesco Sovrano, Kevin Ashley, and Alberto Bacchelli. Toward eliminating hallucinations: Gpt-based explanatory ai for intelligent textbooks and documentation. In *CEUR Workshop Proceedings*, volume 3444, pages 54–65. CEUR-WS Tokyo, Japan, 2023. URL https://ceur-ws.org/Vol-3444/itb23_s3p2.pdf. 4.1, 4.5, 4.6

[322] Timo Speith. A review of taxonomies of explainable artificial intelligence (xai) methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250, 2022. URL https://doi.org/10.1145/3531146.3534639. 5.4

[323] Laura Spillner, Nima Zargham, Mihai Pomarlan, Robert Porzel, and Rainer Malaka. Finding uncommon ground: A human-centered model for extrospective explanations. *arXiv preprint arXiv:2507.21571*, 2025. URL https://arxiv.org/abs/2507.21571. 6.6.1.1

[324] Philipp Spitzer, Niklas Kühl, and Marc Goutier. Training novices: The role of human-ai collaboration and knowledge transfer. *arXiv preprint arXiv:2207.00497*, 2022. URL https://arxiv.org/abs/2207.00497. 4.5

[325] Philipp Spitzer, Niklas Kühl, Daniel Heinz, and Gerhard Satzger. Ml-based teaching systems: A conceptual framework. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–25, 2023. URL https://doi.org/10.1145/3610197. 4.1, 4.5

[326] Philipp Spitzer, Niklas Kühl, Marc Goutier, Manuel Kaschura, and Gerhard Satzger. Transferring domain knowledge with (x) ai-based learning systems. *arXiv preprint arXiv:2406.01329*, 2024. URL https://arxiv.org/abs/2406.01329. 6.1, 6.6.1.3

[327] Philipp Spitzer, Joshua Holstein, Katelyn Morrison, Kenneth Holstein, Gerhard Satzger, and Niklas Kühl. Don't be fooled: the misinformation effect of explanations in human–ai collaboration. *International Journal of Human–Computer Interaction*, pages 1–29, 2025. URL https://www.tandfonline.com/doi/pdf/10.1080/10447318.2025.2574511. 1.1, 2.5, 6.6.2, 7.5.3, 8.1, 9.2.1

[328] Parvathaneni Naga Srinivasu, N Sandhya, Rutvij H Jhaveri, and Roshani Raut. From blackbox to explainable ai in healthcare: existing tools and case studies. *Mobile Information Systems*, 2022(1):8167821, 2022. URL https://onlinelibrary.wiley.com/doi/full/10.1155/2022/8167821. 1

[329] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379, 2020. URL https://arxiv.org/pdf/2002.08596. 2.1

[330] Simone Stumpf, Adrian Bussone, and Dympna O'sullivan. Explanations considered harmful? user interactions with machine learning systems. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2016. URL https://www.yumpu.com/cs/document/view/55465112/explanations-considered-harmful-user-interactions-with-machine-learning-systems. 5.1

[331] Qiyang Sun, Alican Akman, and Björn W Schuller. Explainable artificial intelligence for medical applications: A review. *ACM Transactions on Computing for Healthcare*, 6(2):1–31, 2025. URL https://dl.acm.org/doi/full/10.1145/3709367. 2.1, 2.2

[332] Susu Sun, Lisa M Koch, and Christian F Baumgartner. Right for the wrong reason: Can interpretable ml techniques detect spurious correlations? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 425–434. Springer, 2023. URL https://link.springer.com/chapter/10.1007/978-3-031-43895-0_40. 1.1

[333] William R Swartout and Johanna D Moore. Explanation in second generation expert systems. In *Second generation expert systems*, pages 543–585. Springer, 1993. URL https://link.springer.com/chapter/10.1007/978-3-642-77927-5_24. 1, 1.1

[334] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022. URL https://doi.org/10.1007/978-3-030-34372-9. 3.1

[335] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th international conference on intelligent user interfaces*, pages 109–119, 2021. URL https://doi.org/10.1145/3397481.3450662. 2.2, 4.3.2, 4.6

[336] Maxwell Szymanski, Vero Vanden Abeele, and Katrien Verbert. Designing and evaluating explanations for a predictive health dashboard: A user-centred case study. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2024. URL https://dl.acm.org/doi/full/10.1145/3613905.3637140. 6.6.1.1

[337] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023. URL https://doi.org/10.1109/cvpr52729.2023.00718. 4.1

[338] Lev Tankelevitch, Elena L Glassman, Jessica He, Majeed Kazemitabaar, Aniket Kittur, Mina Lee, Srishti Palani, Advait Sarkar, Gonzalo Ramos, Yvonne Rogers, et al. Tools for thought: Research and design for understanding, protecting, and augmenting human cognition with generative ai. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2025. URL https://dl.acm.org/doi/full/10.1145/3706599.3706745. 2.6

[339] Lev Tankelevitch, Elena L Glassman, Jessica He, Aniket Kittur, Mina Lee, Srishti Palani, Advait Sarkar, Gonzalo Ramos, Yvonne Rogers, and Hari Subramonyam. Understanding, protecting, and augmenting human cognition with generative ai: A synthesis of the chi 2025 tools for thought workshop. *arXiv preprint arXiv:2508.21036*, 2025. URL https://arxiv.org/abs/2508.21036. 1.2, 2.6, 8.1, 8.2.1, 8.5

[340] Heliodoro Tejeda Lemus, Aakriti Kumar, and Mark Steyvers. How displaying ai confidence affects reliance and hybrid human-ai performance. In *HHAI 2023: Augmenting Human Intellect*, pages 234–242. IOS Press, 2023. URL https://mstrep.s3.us-east-1.amazonaws.com/pub

lications/FAIA-368-FAIA230087.pdf. 4.6

[341] Jérôme Théau. Change detection. In *Springer Handbook of Geographic Information*, pages 75–94. Springer, 2012. URL https://doi.org/10.1007/978-0-387-35973-1_129. 3.2.1

[342] Thomas Thesen, Nsomma A Alilonu, and Simon Stone. Ai patient actor: An open-access generative-ai app for communication training in health professions. *Medical Science Educator*, 35(1):25–27, 2025. URL https://link.springer.com/article/10.1007/s40670-024-02250-2. 7.1

[343] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020. URL https://www.nature.com/articles/s41591-020-0942-0. 2.2, 4.1, 4.3.1, 4.3.4

[344] Stratis Tsirtsis and Manuel Gomez Rodriguez. Decisions, counterfactual explanations and strategic behavior. *Advances in Neural Information Processing Systems*, 33:16749–16760, 2020. URL https://dl.acm.org/doi/abs/10.5555/3495724.3497129. 2.1

[345] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022. URL https://www.nature.com/articles/s41467-022-27980-y. 4.1, 4.3.1

[346] Osman Tursun, Simon Denman, Sridha Sridharan, and Clinton Fookes. Towards self-explainability of deep neural networks with heatmap captioning and large-language models. *arXiv preprint arXiv:2304.02202*, 2023. URL https://arxiv.org/abs/2304.02202. 4.1

[347] Sohini Upadhyay, Himabindu Lakkaraju, and Krzysztof Z Gajos. Counterfactual explanations may not be the best algorithmic recourse approach. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 446–462, 2025. URL https://dl.acm.org/doi/full/10.1145/3708359.3712095. 6.1

[348] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12): 2293–2303, 2024. URL https://www.nature.com/articles/s41562-024-02024-1. 1, 1.1, 2.2, 6.1, 8.1

[349] Jasper Van Der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial intelligence*, 291:103404, 2021. URL https://www.sciencedirect.com/science/article/pii/S0004370220301533. 2.2

[350] Arnaud Van Looveren, Janis Klaise, Giovanni Vacanti, and Oliver Cobb. Conditional generative models for counterfactual explanations. *arXiv preprint arXiv:2101.10123*, 2021. URL https://arxiv.org/abs/2101.10123. 2.5

[351] Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q Vera Liao, and Jennifer Wortman Vaughan. Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in ai-powered code completions. *arXiv preprint arXiv:2302.07248*, 2023. URL https://arxiv.org/abs/2302.07248v1. 2.3, 4.1

[352] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–39, 2021. URL https://dl.acm.org/doi/abs/10.1145/3476068. 1, 3.4.2.3

[353] David Vergun. Dod announces ai competition to detect, defeat illegal fishing, Jul 2021. URL https://www.war.gov/News/News-Stories/Article/Article/2703739/dod-announces-a

i-competition-to-detect-defeat-illegal-fishing/. 3.2

[354] Roel Visser, Tobias M Peters, Ingrid Scharlau, and Barbara Hammer. Trust, distrust, and appropriate reliance in (x) ai: a survey of empirical evaluation of user trust. *arXiv preprint arXiv:2312.02034*, 2023. URL https://arxiv.org/pdf/2312.02034. 1.1, 1.2, 2.2

[355] Michael Vössing, Niklas Kühl, Matteo Lind, and Gerhard Satzger. Designing transparency for effective human-ai collaboration. *Information Systems Frontiers*, 24(3):877–895, 2022. URL https://doi.org/10.1007/s10796-022-10284-3. 2.2, 5.2.2

[356] Alexandra Vultureanu-Albişi and Costin Bădică. A survey on effects of adding explanations to recommender systems. *Concurrency and Computation: Practice and Experience*, 34(20):e6834, 2022. URL https://onlinelibrary.wiley.com/doi/full/10.1002/cpe.6834. 1

[357] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *arXiv:1711.00399*, Mar 2018. URL http://arxiv.org/abs/1711.00399. arXiv: 1711.00399. 2.1

[358] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011 (cub-200-2011). Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://www.vision.caltech.edu/datasets/cub_200_2011/. 4.3.3

[359] Rebekah L Waikel, Amna A Othman, Tanviben Patel, Suzanna Ledgister Hanchard, Ping Hu, Cedrik Tekendo-Ngongang, Dat Duong, and Benjamin D Solomon. Generative methods for pediatric genetics education. *medRxiv*, 2023. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10543060/. 2.5, 7.1, 7.2.1

[360] Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Nasif Zaman, Sharif Amit Kamran, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. Text-to-image artificial intelligence to aid clinicians in perceiving unique neuro-ophthalmic visual phenomena. *Irish Journal of Medical Science (1971-)*, 192(6):3139–3142, 2023. URL https://link.springer.com/article/10.1007/s11845-023-03315-8. 7.1

[361] 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. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–24, 2019. URL https://doi.org/10.1145/3359313. 4.1

[362] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–6, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368193. doi: 10.1145/3334480.3381069. URL https://doi.org/10.1145/3334480.3381069. 3.1

[363] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019. URL https://doi.org/10.1145/3290605.3300831. 3.5.1, 4.1

[364] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021. URL https://doi.org/10.1145/3397481.3450650. 1, 2.2, 6.1, 6.6.1, 8.1, 8.6

[365] Xinru Wang and Ming Yin. Effects of explanations in ai-assisted decision making: Principles and comparisons. *ACM Transactions on Interactive Intelligent Systems*, 12(4):1–36, 2022. URL https://dl.acm.org/doi/full/10.1145/3519266. 2.2

[366] Zijie J Wang, Jennifer Wortman Vaughan, Rich Caruana, and Duen Horng Chau. Gam coach: Towards interactive and user-centered algorithmic recourse. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023. URL https://doi.org/10.1145/3544548.3580816. 2.2, 6.1, 6.3.1.1, 6.3.2.3

[367] Frederik Warburg, Martin Jørgensen, Javier Civera, and Søren Hauberg. Bayesian triplet loss: Uncertainty quantification in image retrieval. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, pages 12158–12168, 2021. URL https://doi.org/10.1109/iccv48922.2021.01194. 4.1

[368] Rebecca M Warner. *Applied statistics: From bivariate through multivariate techniques*. Sage publications, 2012. 4.4.2.1, 4.4.2.2

[369] Justin D Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. Design principles for generative ai applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2024. URL https://dl.acm.org/doi/full/10.1145/3613904.3642466. 7.5.2

[370] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019. URL https://doi.org/10.1109/tvcg.2019.2934619. 1, 1

[371] WildMe. Image analysis pipeline - wild me documentation, 2023. 5.2.1

[372] Uta Wilkens. Artificial intelligence in the workplace–a double-edged sword. *The International Journal of Information and Learning Technology*, 37(5):253–265, 2020. URL https://doi.org/10.1108/ijilt-02-2020-0022. 4.5

[373] Michael Williams and Tami Moser. The art of coding and thematic exploration in qualitative research. *International management review*, 15(1):45–55, 2019. URL https://www.academia.edu/download/82465402/imr-v15n1art4.pdf. 5.2.2

[374] Stephan Winter, Nicole C Krämer, Leonie Rösner, and German Neubaum. Don't keep it (too) simple: How textual representations of scientific uncertainty affect laypersons' attitudes. *Journal of Language and Social Psychology*, 34(3):251–272, 2015. URL https://doi.org/10.1177/0261927x14555872. 4.3.4

[375] William Van Woensel, Floriano Scioscia, Giuseppe Loseto, Oshani Seneviratne, Evan Patton, Samina Abidi, and Lalana Kagal. Explainable clinical decision support: towards patient-facing explanations for education and long-term behavior change. In *International Conference on Artificial Intelligence in Medicine*, pages 57–62. Springer, 2022. URL https://link.springer.com/chapter/10.1007/978-3-031-09342-5_6. 2.4

[376] Joost F Wolfswinkel, Elfi Furtmueller, and Celeste PM Wilderom. Using grounded theory as a method for rigorously reviewing literature. *European journal of information systems*, 22(1):45–55, 2013. URL https://doi.org/10.1057/ejis.2011.51. 4.3.6

[377] Melisa L Wong, Francesca M Nicosia, Alexander K Smith, Louise C Walter, Vivian Lam, Harvey Jay Cohen, Kah Poh Loh, Supriya G Mohile, Carling J Ursem, and Margaret L Schwarze. "you have to be sure that the patient has the full picture": Adaptation of the best case/worst case communication tool for geriatric oncology. *Journal of geriatric oncology*, 13(5):606–613, 2022. URL https://www.sciencedirect.com/science/article/pii/S1879406822000157. 6.6.3.1

[378] Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng, and Fei Wang. Faithful ai in medicine: A systematic review with large language models and beyond. *medRxiv*, 2023. URL https://doi.org/10.21203/rs.3.rs-3661764/v1. 4.3.3

[379] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang 'Anthony' Chen. Chexplain: Enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13. ACM, April 2020. doi: 10.1145/3313831.3376807. URL https://doi.org/10.1145/3313831.3376807. 2.2

[380] Yanwu Xu, Li Sun, Wei Peng, Shuyue Jia, Katelyn Morrison, Adam Perer, Afrooz Zandifar, Shyam Visweswaran, Motahhare Eslami, and Kayhan Batmanghelich. Medsyn: Text-guided anatomy-aware synthesis of high-fidelity 3d ct images. *IEEE Transactions on Medical Imaging*, 2024. URL https://ieeexplore.ieee.org/abstract/document/10566053. (document), 1.1, 1.2, 7.1, 7.2.1, 7.2.3, 7.3.1, 7.1, 7.2, 7.4.3, 7.4.3.1, 8.1, 8.3.2

[381] Chhavi Yadav, Michal Moshkovitz, and Kamalika Chaudhuri. Xaudit: A theoretical look at auditing with explanations. *arXiv preprint arXiv:2206.04740*, 2022. URL https://arxiv.org/abs/2206.04740. 1

[382] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 189–201, 2020. URL https://doi.org/10.1145/3377325.3377480. 2.2

[383] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. Harnessing biomedical literature to calibrate clinicians' trust in ai decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2023. URL https://doi.org/10.1145/3544548.3581393. 4.1

[384] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications*, 11(1):2583, 2020. doi: 10.1038/s41467-020-16185-w. URL https://doi.org/10.1038/s41467-020-16185-w. 3.2

[385] Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho De Castro, Shruthi Bannur, Stephanie Hyland, et al. Multimodal healthcare ai: identifying and designing clinically relevant vision-language applications for radiology. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2024. URL https://arxiv.org/pdf/2402.14252. 7.1, 7.5.2, 8.1, 8.2.1, 8.2.3

[386] Daniel Zhang, Yang Zhang, Qi Li, Thomas Plummer, and Dong Wang. Crowdlearn: A crowd-ai hybrid system for deep learning-based damage assessment applications. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 1221–1232. IEEE, 2019. URL https://doi.org/10.1109/icdcs.2019.00123. 4.1

[387] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M Padilla, Jeffrey Caterino, Ping Zhang, et al. Rethinking human-ai collaboration in complex medical decision making: A case study in sepsis diagnosis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024. URL https://dl.acm.org/doi/full/10.1145/3613904.3642343. 2.2, 6.3.2.3, 6.6.1.2

[388] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305, 2020. URL https://doi.org/10.1145/3351095.3372852. 3.4.2.3

[389] Zelun Tony Zhang and Leon Reicherts. Augmenting human cognition with generative ai: Lessons from ai-assisted decision-making. *arXiv preprint arXiv:2504.03207*, 2025. URL https://arxiv.org/abs/2504.03207. 1, 1.2, 2.6, 6.6.1, 6.6.1.2, 8.1, 8.2.1, 8.5

[390] Zibin Zhao and Cagatay Turkay. Exploring how expertise impacts acceptability of ai explanations: a case study from manufacturing. In *Proceedings of the ACM CHI Workshop Human-Centered Perspectives in Explainable AI*. ACM, 2023. URL https://wrap.warwick.ac.uk/175429/. 2.2

[391] Lingxuan Zhu, Weiming Mou, Keren Wu, Jian Zhang, and Peng Luo. Can dall-e 3 reliably generate 12-lead ecgs and teaching illustrations? *Cureus*, 16(1), 2024. URL https://assets.cureus.com/uploads/technical_report/pdf/211073/20240221-2604-dsrwc6.pdf. 2.5

[392] Erik Ziegler, Trinity Urban, Danny Brown, James Petts, Steve D Pieper, Rob Lewis, Chris Hafey, and Gordon J Harris. Open health imaging foundation viewer: an extensible open-source framework for building web-based imaging applications to support cancer research. *JCO clinical cancer informatics*, 4:336–345, 2020. URL https://ascopubs.org/doi/pdfdirect/10.1200/CCI.19.00131. 8.2.2