

Supporting Transparency in AI Systems: Designing Training Dataset Explanations and Assessing AI Literacy

by

Ariful Islam Anik

A thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
of the University of Manitoba
in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

Department of Computer Science

University of Manitoba

Winnipeg, Manitoba, Canada

© Copyright 2026 by Ariful Islam Anik

Abstract

Training datasets play a central role in shaping the behavior, limitations, and impacts of AI systems, yet transparency about training datasets remains limited in practice. As a result, users can lack the context needed to meaningfully interpret system outputs, form informed judgments, and engage with AI systems effectively. One approach to addressing this lack of context is through training dataset explanations, which communicate key characteristics of the data used to train AI systems. This dissertation investigates how such explanations can support data-centric transparency in AI systems.

As part of this exploration, the dissertation presents three complementary studies. The first two studies examine how explanation design factors influence how users use training dataset explanations in an onboarding setting and how they assess the systems that provide them. The first study examines the influence of presentation style by comparing a data storytelling-based presentation with a Q&A-based presentation, while the second study investigates the effects of information depth by comparing summary and detailed training dataset explanations. These studies show that, in an onboarding context, differences in presentation style and information depth lead to different patterns of use, perceived understanding, and trust. These findings highlight how design choices in training dataset explanations can shape users' assessment of AI systems.

The third study introduces and validates a dual-format AI literacy scale grounded in an established AI literacy framework. The study focuses on scale development and psychometric validation of its measurement properties. The scale combines self-assessed and demonstrated knowledge to support systematic assessment of users' conceptual understanding of AI. Follow-up analyses using the scale reveal mismatches between perceived and demonstrated understanding, highlighting the limitations of relying solely on self-reported measures in explanation research.

Overall, this work contributes to the field of human-centered explainable AI by foregrounding training datasets as a critical aspect of system transparency and by contributing both design insights and measurement tools to support explanation research. In doing so, it supports explanation design and evaluation practices that can better account for user differences and interaction contexts, advancing the long-term objective of effective human–AI collaboration.

Acknowledgements

I would like to begin by expressing my deepest gratitude to my advisor, Dr. Andrea Bunt, for her guidance, support, and encouragement throughout both my MSc and PhD journey. I consider myself incredibly fortunate to have had the opportunity to learn from her over so many years, through the many ups and downs that are inevitable in graduate research. Her thoughtful feedback, patience, and mentorship have shaped not only this dissertation but also my growth as a researcher and as a person. I am especially grateful for the freedom she gave me to explore my own ideas while always helping me refine them with clarity and care. Beyond research, Andrea has taught me invaluable lessons about kindness, integrity, balance, and how to support others with generosity and thoughtfulness. She has been a role model for the kind of advisor, researcher, and person I hope to become. I am also grateful for her financial support throughout my degree. I would also like to thank the University of Manitoba for the UMGF award and the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding my research.

My gratitude extends to my advisory committee members, Dr. James Young and Dr. Max Turgeon, for their valuable feedback, questions, and suggestions throughout this dissertation. Their perspectives helped strengthen this work and pushed me to think more critically about my research and its broader contributions. I would also like to thank

Dr. Julita Vassileva for serving as the external examiner for my dissertation and for her thoughtful questions, constructive feedback, and engaging discussions. I am further grateful to Dr. Celine Latulipe, Dr. David Gerhard, Dr. Daniel Rea, and Dr. Houda Elmimouni for their constructive feedback and discussions during Human-Computer Interaction Lab meetings, which, at various points, helped shape the direction of my work in meaningful ways.

I would like to thank my friends and fellow labmates in the Human-Computer Interaction Lab at the University of Manitoba. It has been a long journey, and the conversations, feedback, shared struggles, and encouragement I received from them made the PhD experience more meaningful and less isolating. I especially want to thank Afsane, Ananta, Danika, Emmanuel, Fairuz, James, Mahya, Mehran, Melika, Min, Patrick, Pouya, Raquel, Sam, Sarvnaz, Zach, Zachary, and many others for their friendship and support. The lab always felt like a calm and comforting place where I could focus and feel at ease. Coming to the lab was therefore something I truly enjoyed, and I am certain that the people made that experience even better. I am also grateful to my friends in Winnipeg, whose companionship and support made this city feel a bit more like home.

I am eternally grateful to my family for their unconditional love and support. Even from afar, their encouragement has been a constant source of strength, and their belief in me helped me continue through the many uncertain and difficult moments of this journey. I would not be where I am today if not for their sacrifices. I am especially grateful to my parents, who have always wanted the best for me and gave me the freedom to choose my own path, even when my choices may not have been easy for them to accept. My sister has been a cherished presence throughout my life, and I am deeply grateful for

her love, care, and support. I am also thankful to my brother-in-law for his kindness and support. My brothers and sisters-in-law practically raised me during the years I lived with them so that I could attend good schools and have better opportunities. I am deeply grateful for their love, generosity, and the sacrifices they made for me. I am also grateful to my in-laws for welcoming me into their family with warmth and kindness since the day they came to know me, and for their support throughout these years. I genuinely believe that very few people are lucky enough to receive such love and care. Finally, I want to send my love to my nephews and niece. Even across the distance, I can feel their love, and I am so grateful thankful to have them in my life. I love you all.

Above all, I want to thank my wife, Afsane Baghestani, whose love, patience, and support have carried me through this PhD. This journey is especially meaningful to me because I met her right at the beginning. I wanted a PhD. I did not expect to find my person along the way. Over the years, we got to know each other, grew together, and began building a life together alongside the many demands, uncertainties, and milestones of graduate school. She has been with me through the long days, stressful deadlines, moments of doubt, and small victories, always offering care, encouragement, and perspective when I needed them most. Her presence made this journey lighter, warmer, and more joyful, and I cannot imagine having gone through it without her. I am endlessly grateful for everything she has done for me — from being a sounding board for my ideas, to being the first pilot of my studies, the first proofreader of many submissions, my constant source of comfort, and my anchor through it all. Thank you for believing in me, for standing beside me, and for making this chapter of my life so much more beautiful. I love you.

Finally, I am grateful for the experiences, challenges, and people that shaped this journey, and to God for giving me the strength, patience, and guidance to carry on through the difficult moments. Completing this dissertation has been one of the most difficult and rewarding experiences of my life. As I look back, I realize that much of this journey was not made up of big moments, but of ordinary days: showing up, doing the work, facing uncertainty, and finding small reasons to continue. And I do not mean “ordinary” in a negative way. I have always believed that there is beauty in ordinary things. Perhaps, in its own quiet way, that was the point — something *The Office* taught me.

Dedication

To my parents, Ma and Abba. No degree or achievement can fully compensate for the distance from you during this journey. Still, this dissertation is a small tribute to your love, sacrifices, support, and the strength you continue to give me.

Publications

Certain portions of this thesis have been published in conference or workshop proceedings or are currently under peer review. Permissions to include these works in this dissertation have been granted by the respective publishers. The following is a list of publications that appear in this thesis, organized by chapter.

Chapter 3

Ariful Islam Anik and Andrea Bunt, "Supporting User Critiques of AI Systems via Training Dataset Explanations: Investigating Critique Properties and the Impact of Presentation Style," *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, Liverpool, United Kingdom, 2024, pp. 134-147, doi: 10.1109/VL/HCC60511.2024.00024.

Ariful Islam Anik and Andrea Bunt, "Diversity Challenges in Recruiting for Human-Centered Explainable AI Studies," *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, Liverpool, United Kingdom, 2024, pp. 429-431, doi: 10.1109/VL/HCC60511.2024.00067

Chapter 4

Ariful Islam Anik and Andrea Bunt. 2023. Understanding and Mitigating the Negative Consequences of Training Dataset Explanations. *In Proceedings of the ACM CHI 2023 Workshop on Human-centered Explainable AI (HCXAI)*.

Ariful Islam Anik and Andrea Bunt. 2026. Designing Effective Training Dataset Explanations: The Impact of Information Depth and Progressive Disclosure. *In 31st International Conference on Intelligent User Interfaces (IUI'26)*, March 23–26, 2026, Paphos, Cyprus. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3742413.3789087>

Table of Contents

Abstract.....	i
Acknowledgements	iii
Dedication	vii
Publications.....	ix
Table of Contents.....	xi
List of Tables.....	xv
List of Figures	xvii
Contribution of Authors	xix
Chapter 1 – Introduction	1
1.1. Thesis Objective and Research Questions	4
1.2. Approach	4
1.3. Summary of Contributions.....	6
1.4. Thesis Outline	7
Chapter 2 – Background and Related Work.....	9
2.1. Explanations for Supporting AI Transparency.....	10

2.2. Impact of AI Explanations on Users	15
2.3. Prominent Design Factors for Explanation Design.....	16
2.4. Diversity and Assessment of Users' AI Literacy.....	19
2.5. Summary of Related Work.....	24
Chapter 3 – Investigating Users' Usage of Training Dataset Explanations and Impact of Presentation Style.....	25
3.1. Presentation Style Selection	27
3.2. Using Data Storytelling in Training Dataset Explanations	28
3.3. Designing Q&A-based Training Dataset Explanation.....	33
3.4. Study	35
3.5. Results.....	41
3.6. Discussion	56
3.7. Summary	62
Chapter 4 – Explainability Pitfalls: Investigating the Roles of Information Depth and Progressive Disclosure for Training Dataset Explanations.....	63
4.1. Explainability Pitfalls	65
4.2. Studying Information Depth and Progressive Disclosure	69
4.3. Results.....	79
4.4. Discussion	93
4.5. Summary	98

Chapter 5 – Developing a Psychometrically Validated Scale for Assessing AI Literacy	99
5.1. Defining AI Literacy	101
5.2. Developing a Scale to Assess AI Literacy	102
5.3. Results	111
5.4. Scale Validation.....	123
5.5. Relationship Between Self-Assessed and Factual AI Understanding	130
5.6. Discussion	139
5.7. Summary.....	147
Chapter 6 – Reflections on Designing Effective (Training Dataset) Explanations for Diverse Users	149
6.1. Insights.....	150
6.2. Considerations	154
6.3. Methodological Reflections and Challenges.....	157
6.4. Summary.....	162
Chapter 7 – Conclusion.....	163
7.1. Thesis Contributions	164
7.2. Limitations	167
7.3. Future Research Directions	169
Bibliography.....	173
Appendices.....	201

Appendix A – Presentation Style Study Documents	201
Appendix B – Information Depth Study Documents	231
Appendix C – AI Literacy Study Documents	259

List of Tables

Table 1: Example participant comments for each explanation topic.....	44
Table 2: Examples of biases identified by participants.	47
Table 3: Examples participant comments by critique accuracy type.	49
Table 4: Median (IQR) Likert-scale questionnaire values by presentation style.....	50
Table 5: Summary of the key results.....	51
Table 6: Potential negative consequences of explanations	67
Table 7: Median (IQR) Likert-scale questionnaire values by <i>Information Depth</i>	80
Table 8: Median (IQR) Likert-scale questionnaire values by <i>Progressive Disclosure</i>	82
Table 9: Illustrative examples of participants’ critique across <i>Information Depth</i>	86
Table 10: Illustrative critique examples by accuracy type.	88
Table 11: Abilities in Long and Magerko’s AI literacy framework.	104
Table 12: Retained Likert-scale items grouped by factor.	116
Table 13: Final set of 18 multiple-choice items included in the AI Literacy Scale.	121
Table 14: Summary of scale development and validation analyses.	128
Table 15: Median (IQR) values for Self-Assessed and Factual AI Understanding across AI training groups.....	135
Table 16: Participant age distribution in the development and validation samples.	280

Table 17: Participant gender distribution in the development and validation samples.	280
Table 18: Highest education level in the development and validation samples.	281
Table 19: Primary educational background of participants in the development and validation samples.....	281
Table 20: Participant prior AI training in the development and validation samples.	282
Table 21: Factor loadings of Likert-scale items in the development sample.	289
Table 22: Factor loadings of Likert-scale items after dropping weak items.	292
Table 23: Item Response Theory (IRT) parameter estimates for Factor 1.	295
Table 24: Item Response Theory (IRT) parameter estimates for Factor 2.	297
Table 25: IRT estimates for multiple-choice items in the development sample.	299
Table 26: Model fit statistics for multiple-choice items in the development sample.	302
Table 27: IRT estimates for multiple-choice items in the validation sample.....	305
Table 28: Model fit statistics for multiple-choice items in the validation sample.....	307
Table 29: Measurement invariance testing across gender.	308
Table 30: Measurement invariance testing across prior AI training groups	308
Table 31: Differential item functioning (DIF) analyses of multiple-choice items across gender and training groups.....	309
Table 32: Regression of calibration mismatch on self-assessed AI understanding.	310
Table 33: Regression of calibration mismatch on standardized self-assessed AI understanding.	311

List of Figures

Figure 1: Setting up the context and overview in the Data Story	31
Figure 2: Visualizations of different demographic information in the Data Story.....	32
Figure 3: An example of a slideshow in the Data Story.....	32
Figure 4: An example of local interactivity in the Data Story.....	33
Figure 5: Overview of the Q&A-based presentation for training dataset explanation.....	34
Figure 6: Coverage of explanation topics according to Presentation Style.....	43
Figure 7: Type of biases identified by participants for each Presentation Style	46
Figure 8: Snippets from the different versions of training dataset explanations.....	75
Figure 9: Overview of the study procedure.	78
Figure 10: Participants' perceived understanding of the data across <i>Information Depth</i>	81
Figure 11: Average number of comments per participant across <i>Information Depth</i>	83
Figure 12: Distributions of topics in participants' critique across <i>Information Depth</i>	85
Figure 13: Distribution of participants' critique accuracy across <i>Information Depth</i>	87
Figure 14: Distribution of participants' critique accuracy across <i>Progressive Disclosure</i>	89
Figure 15: Overview of the process used to develop the dual-format AI literacy scale.	103
Figure 16: Demographic distribution of survey participants.....	108
Figure 17: Parallel analysis for factor selection for the Likert-scale items.....	112

Figure 18: Item Characteristic Curves (ICCs) for Likert-scale items within Factor 1. ...	114
Figure 19: Item Characteristic Curves (ICCs) for Likert-scale items within Factor 2. ...	115
Figure 20: Item Characteristic Curves (ICCs) for all multiple-choice items.	119
Figure 21: Test Information Function (TIF) plot for retained multiple-choice items.	120
Figure 22: Demographic distribution of the validation sample.	124
Figure 23: Item Characteristic Curves (ICCs) for retained multiple-choice items.	127
Figure 24: Distribution of standardized calibration scores by AI literacy group.	133
Figure 25: Boxplot of standardized calibration scores by gender.....	134
Figure 26: Medians for Self-assessed and Factual understanding by AI training levels.	135
Figure 27: Boxplot showing calibration scores across AI training groups.	136
Figure 28: Calibration curve of Self-Assessed Understanding vs. calibration Score.	137
Figure 29: Item characteristic curves (ICCs) for Factor 1.....	296
Figure 30: Item characteristic curves (ICCs) for Factor 2.....	298
Figure 31: ICCs for all multiple-choice items in the development sample.	302
Figure 32: ICCs for multiple-choice items in the validation sample.	306

Contribution of Authors

This dissertation represents my work conducted under the supervision of Dr. Andrea Bunt. I took the lead in all stages of the research presented in this dissertation, including conceptualization, research design, development of study materials and procedures, data collection, data analysis, interpretation of findings, and preparation and revision of the manuscripts.

Dr. Andrea Bunt provided supervision, guidance, and feedback throughout the research process. She contributed feedback on the research design, analysis, interpretation of findings, and written manuscripts, and provided guidance on the overall direction of the research.

Chapter 1

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) systems increasingly shape decisions that affect people’s lives, from everyday recommendations [49,136,175,341], to screening candidates in hiring [182,212,261], to assisting doctors in healthcare [20,50,63,115], and even informing judgments in criminal justice [7,42,43]. As these technologies expand their reach, it is critical that *interested parties*¹ can use these systems in an informed and responsible manner. This requires transparency, which is defined as the degree to which people can understand how a system operates and what factors influence its outputs [93,120,246].

AI systems are rarely transparent by default, given that their behavior is driven by “black-box” algorithms. This lack of transparency undermines users’ trust, hinders accountability, and limits people’s ability to make informed judgments about these

¹ We use the term *interested parties* to refer to people impacted by technology due to the colonial connotations of the term *stakeholders* [251].

systems [83,200,267]. To address these challenges, the field of Explainable AI (XAI) continues to develop a variety of explanation techniques that reveal aspects of how AI systems function [1,2,38,132]. While studies have demonstrated that explanations can increase perceived transparency [83,208,258], trust [167,198,278], and acceptance [83,173,336], XAI research has also been criticized for its predominant focus on algorithms [106,223,224]. Prior studies show that algorithm-centered explanations often fail to align with users' needs and can, at times, even contribute to detrimental effects, such as inappropriate trust, overreliance, and increased cognitive burden [24,25,47,50,163,255].

Several researchers, therefore, argue for moving beyond algorithmic transparency and toward socio-technical and human-centered transparency. AI systems are socio-technical in that they embody social choices in data collection, design, and deployment, and their impacts extend beyond technical outputs [8,103,106,108]. Furthermore, explainability is a human factor rather than an algorithmic property [105]. Explanations are socially situated [202,206,223] and reflect a shared sense-making process that occurs between an explainer and explainee. To be meaningful, explainability must account for more than algorithms. It should incorporate the data used to train the system, the design decisions that shaped the system, and the humans behind it [47,103,127,210,324].

A critical element in a socio-technical, human-centered view of transparency of AI systems is training data. AI systems are data-driven, meaning they learn patterns from training data and leverage this learning to make predictions and recommendations [30,218]. As a result, the properties of training data directly influence the fairness, accuracy, and reliability of their predictions. Further, any potential problems or biases in

the data often manifest in the system’s behavior, raising concerns about equity and accountability [26,40,57]. This makes the entire pipeline for training dataset formation, starting from the data collection through data processing (e.g., cleaning, annotating), to data evaluation, important [194].

Despite this importance, information about training data is rarely communicated to users [33,81,153]. As a result, users often lack critical context about the systems they use, which can lead to miscalibrated trust and inappropriate reliance on the system. Limited transparency about training data can be particularly consequential in early-use settings, such as system onboarding, because early interactions with the system shape mental model formation, set expectations, and calibrate trust [55,190]. While prior work has explored the initial feasibility of training dataset explanations [8], we still do not have a systematic understanding of how to design such explanations for diverse users in onboarding settings.

Motivated by this, in this dissertation, we study how to design effective training dataset explanations for onboarding contexts. We examine how design choices shape users’ understanding of explanations and their assessments of AI systems. We also take a critical step towards design for diverse users by developing and validating an instrument that assesses users’ AI literacy (i.e., the ability to recognize AI systems, understand AI’s strengths and limitations, interpret AI decision-making processes, and reflect on societal and ethical implications of AI) and provides a basis for assessing and potentially tailoring explanations and AI interventions to different user groups.

1.1. Thesis Objective and Research Questions

The primary objective of this dissertation is to address challenges in designing effective training dataset explanations for diverse users. Our specific approach within this greater objective is that we place a strong emphasis on understanding the impact of key design factors in training dataset explanations and take a first step toward enabling design and evaluation with a diverse set of users by assessing their AI literacy through an instrument that future work can use to guide the design and assessment of AI interventions. To advance this objective, this dissertation investigates the following research questions.

- i) How does the presentation style of training dataset explanations impact users' comprehension of the explanation and their assessments of AI systems?
- ii) How does the amount of information in training dataset explanations affect users' comprehension, cognitive load, and their assessments of AI systems?
- iii) What constitutes a robust scale to assess users' AI literacy, and how can such a scale support future research and system design?

1.2. Approach

To answer our research questions, we conducted three complementary studies. In the following sections, we provide a brief overview of the studies and how they connect with each other.

1.2.1. Presentation Style Study

To investigate our first research question, we conducted a controlled user study to examine how people use information about training data to assess an AI system and how

presentation style impacts their comprehension and assessments. We asked participants to assess an AI system based on an explanation on the system’s training data presented either in a narrative-driven “Data Story” format or a structured Q&A format. We collected data from participants’ written assessments of the system, responses to a self-report questionnaire, and semi-structured interview sessions. Through the study, we provide insights into what information users focus on when they interact with training dataset explanations, whether they seem to understand the material, and how the presentation style shapes their information use and understanding. This study was published at the IEEE VL/HCC 2024 conference and the materials presented in this dissertation are a revised and adapted version of the content from the publication [9].

1.2.2. Information Depth Study

To answer the second research question, we began with a scoping review of the negative consequences of explanations (i.e., explainability pitfalls [107]) and identified information depth (i.e., the level of detail and complexity in an explanation) as a key factor that can contribute to information overload. We then conducted a controlled user study in which participants assessed two AI systems, each accompanied by training dataset explanations that varied in information depth (summary vs. detailed). We collected participants’ written assessments and self-reported questionnaire responses for each system, and semi-structured interview responses. Through the study, we provide insights into how information depth in training dataset explanations affects users’ understanding, their assessment of an AI system, and cognitive load. This study was published at the ACM IUI 2026 conference and the materials presented in this dissertation are a revised and adapted version of the content from the publication [11].

1.2.3. Assessing Users' AI Literacy

To answer our final research question, we developed and psychometrically validated a dual-format scale for assessing users' AI literacy, combining self-assessment items with factual knowledge items. Following established scale development frameworks [61,92,143], we applied factor analysis and item response theory to develop and evaluate the psychometric properties of the scale. Through the study, we provide a tool that enables researchers and practitioners to reliably assess users' AI literacy, identify differences across user groups, and potentially tailor explanations and intervention strategies so that AI systems better support diverse users. We also examined the relationship between self-assessed and factual understanding of AI and analyzed how experience-related factors, such as prior AI training, influence users' AI literacy.

1.3. Summary of Contributions

We contribute to the field of human-computer interaction and explainability research in the following ways:

- (1) We provide empirical evidence on how training dataset explanations shape users' understanding and assessments of AI systems.
- (2) We identify design trade-offs and offer guidance on using presentation style and information depth to balance clarity, comprehensiveness, and usability of training dataset explanations.
- (3) We develop and psychometrically validate a dual-format instrument that combines subjective self-assessment with demonstrated understanding of AI,

enabling future research to more comprehensively assess users' AI literacy, as well as design and evaluate AI interventions for diverse users.

This dissertation will help researchers and practitioners by offering empirically grounded guidance for designing training dataset explanations and a validated instrument for assessing users' AI literacy. Accordingly, these contributions broaden the scope of XAI beyond algorithmic transparency by advancing the design of data-centric explanations and equipping researchers with a robust tool for characterizing users in future explainability research. By enabling both better designed training dataset explanations and more accurate assessment of users' AI literacy, this dissertation promotes more informed and responsible use of AI systems.

1.4. Thesis Outline

Chapter 2 presents background and related work on AI transparency and explanations, which motivates the need for training dataset explanations, and situates the contributions of this dissertation within this literature. Chapter 3 describes the first study, where we investigate the impact of presentation style. Chapter 4 presents the second study, in which we examine the impact of information depth. Chapter 5 describes the development and validation of the scale to assess users' AI literacy and examines the relationship between users' self-assessed and demonstrated understanding of AI. Chapter 6 reflects on the findings and approaches used across these studies, synthesizing insights and offering considerations for future research. Finally, Chapter 7 concludes the dissertation, summarizing key takeaways, and outlining potential avenues for further investigation.

Chapter 2

Background and Related Work

To situate this dissertation in existing literature, we first review work that investigates explanations as a promising way to support effective human–AI collaboration. This includes examining the diverse goals that explanations are intended to support and characterizing the main types of explanations explored in prior research. We then examine studies that explore the impacts of explanations on users, paying attention to both benefits and potential negative consequences. Next, we survey research on how design choices such as presentation style and information depth shape the effectiveness of explanations. Finally, we review studies that address the diversity of users’ AI literacy and knowledge and highlight related work that emphasizes the importance of building robust and validated instruments for assessing AI literacy. We end with reviewing existing tools developed to capture users’ AI literacy.

2.1. Explanations for Supporting AI Transparency

In this section, we examine how explanations in AI systems have been conceptualized and what roles they are intended to play in human–AI collaboration. We summarize the main goals of explanations, review prominent explanation approaches, and highlight a shift from algorithm-centered to socio-technical, human-centered perspectives that motivates our focus on training dataset explanations.

2.1.1. Explanations for Effective Human–AI Collaboration

In recent research on AI systems, the concept of “human–AI collaboration” has gained prominence, with a growing body of work exploring ways to design and support effective partnerships between humans and AI systems [25,55,71,155,165]. The objective of such collaboration is to leverage the complementary strengths of both to enhance overall collaborative performance. While some studies have found that human–AI collaboration leads to improved performance compared to individuals or AI working alone [55,71,201,265], others have shown no improvement or even a detrimental impact on performance [25,151,311,340]. Insufficient transparency is cited as a major barrier to effective collaboration as users’ struggle to interpret how the AI produces its outputs and to make informed judgments about whether to rely on them [50,146,322]. For example, prior research shows that people may be skeptical of useful AI outputs or overly reliant on biased AI outputs, both of which can undermine human–AI collaboration [47,235].

A promising way to address the lack of transparency in AI systems is through explanations [1,2,198]. While many definitions exist for explanations, broadly, an explanation is a human-oriented account of how an AI system produced (or would

produce) a specific outcome for a given audience and context [97,202,223]. An explanation addresses users’ “why”, “what”, or “how” questions by identifying and communicating the most relevant causes or factors behind the system’s behavior, enabling users to understand, evaluate, and act upon the information [72,179,198,206]. Prior work characterizes many explanations as contrastive (why one outcome occurred instead of another) [202,223,267] and social (tailored to an audience’s needs, knowledge, and context) [103,202,223], and suggests that they tend to emphasize usable causal factors over exhaustive detail [184,206,223].

Explanations have the potential to enhance the transparency of systems, informing users about whether to trust the system and promoting appropriate reliance on AI systems [38,54,96,198]. However, many AI explanations are algorithm-centered and are often too technical for the intended receivers [106,223,224]. Consequently, instead of supporting users in achieving better human–AI collaboration, explanations may inadvertently burden users [103,167,181]. Moreover, since AI systems are deeply embedded in socio-technical contexts, transparency must extend beyond model internals and include factors such as training data, system purpose, and user goals [8,37,103,105,223]. Recognizing this broader perspective, in this dissertation, we focus on training dataset explanations [8] as a key form of transparency. By revealing how data shape model behavior and outcomes, these explanations aim to support more informed, responsible, and effective human–AI collaboration.

2.1.2. Goals of AI Explanations

Explanations can be designed to serve a range of goals. Prior work has derived nine common goals for explanations [80,237,308] that include: *Transparency* (explaining how

the system works); *Trust* (increasing users' confidence in the system); *Satisfaction* (increasing the ease of use or enjoyment); *Efficiency* (aiming to make decisions faster); *Effectiveness* (aiming to help users make good decisions); *Education* (enabling users to generalize and learn); *Persuasiveness* (convincing users to perform an action); *Scrutability* (allowing users to tell the system it is wrong); and *Debugging* (enabling users to identify defects in the system). When an explanation has multiple intended goals, their dependencies may be complementary, contradictory, or even unknown. For example, improved *transparency* often aligns with more informed *trust* [83,167,183,340], indicating that these goals usually complement each other. On the other hand, the goal of *persuasiveness* can contradict the goal of *effectiveness* in some settings [70,237].

In this dissertation, we emphasize the practical utility of explanations for users' immediate tasks and decisions [144,195], and within this emphasis, we introduce and examine *critiquing* as an explanation goal. By *critiquing*, we mean helping users identify a system's strengths and weaknesses for their own use cases and decision contexts, including where it is likely to work well and where it may fail. This goal addresses well-documented challenges, such as users resisting a tool due to a lack of knowledge about its capabilities, utility, intended usage, and potential algorithmic aversion [94,114]. We position critiquing within onboarding [55,190], where users reflect on the system before or alongside early interaction. We also consider how critiquing may relate to existing goals. For example, the capacity to *critique* through explanations may shape perceived *transparency* and inform *trust*, which can support calibrated reliance.

2.1.3. Explanation Approaches in AI Systems

A large body of prior research has explored how to improve system transparency and interpretability through various explanation approaches. Most of these existing approaches focus on providing information on the decision process and the decision factors. Some of these approaches are *local explanations* that focus on explaining individual decisions from a system (e.g., *counterfactual* [267,275] or *case-based* [54,236] explanations). Others are *global explanations* that focus on helping users understand the overall system [96], by describing how the model works (e.g., *feature importance* [38,88], *feature contribution* [25,72,326,340], and *demographic-based* [13,327] explanations). In terms of the degree of specificity, some of the approaches are *model-specific explanations* [63,179,193] that pertain to a particular model and can explain that model’s decisions. Others are *model-agnostic explanations* [267,268] that can explain the decisions from a range of models.

The majority of these explanation approaches are algorithm-centered [1,2,37,197,255]. Recognizing that this narrow focus has important limitations, researchers have called for a shift to broader contexts [103,105,106,196]. These calls have led to the development of two complementary human-centered directions: socio-technical approaches [103,104,106,108,223], which consider factors beyond the algorithmic “black box,” such as human, organizational, and ethical dimensions; and data-centric approaches [8,33,37,345], where the focus of the explanations is on the training dataset of the system.

Within data-centric approaches, there is ongoing work on documenting and presenting training data, particularly for individuals with specific expertise (e.g., AI developers, data specialists, and domain experts) [15,33,76,127,145,257]. This work

commonly employs Q&A approaches to organize training dataset documentation. For example, *datasheets for datasets* [127] utilize a Q&A approach to provide information on the motivation, creation, composition, intended use, and maintenance of a dataset, whereas *factsheets* [15] uses a Q&A format to provide information on the creation of a system, training, deployment, and performance. Another common approach is to use structured headings to organize the information. For example, *Data Nutrition Labels* [145] present dataset information similarly to the standardized “Nutrition Facts” seen on food packaging. Other work has also explored combined formats, including *Data cards* [257], which combines blocks of texts, open-ended answers, and examples to summarize essential facts about the data (e.g., the process and rationale that influence the data).

Although researchers have advocated extending these dataset explanations to a broader range of interested parties [76,257], little is known about how to make such explanations accessible and what effects they may have on users. In my M.Sc. thesis [12], we explored the feasibility of a model-agnostic, global explanation describing how an AI system was trained. Findings showed that users valued access to such explanations but raised open questions about how they process and use them. This Ph.D. dissertation builds on that foundation by examining how users use training dataset explanations to assess and understand AI systems, and how design choices in these explanations affect this process.

2.1.4. Summary

This section highlights the limitations of algorithm-centered explanations within a human-centered view of transparency and positions training dataset explanations as a complementary form of explanations. It also clarifies existing explanation goals and

introduces *critiquing* as a goal that emphasizes practical utility of explanations. Together, these points motivate our focus on designing training dataset explanations that help users understand and assess AI systems.

2.2. Impact of AI Explanations on Users

Numerous studies have investigated the impact of explanations on how users perceive and interact with AI systems. These studies have shown that explanations can promote transparency [83,208,258], impact users' acceptance of the systems [83,173,336], and influence fairness perceptions [8,36,38,96,191]. The presence of explanations leads to increased transparency, subsequently fostering users' acceptance of the system. However, there are mixed results regarding the impact of explanations on users' trust in AI systems. While some studies reported positive effects of explanations on users' satisfaction and trust [167,180,198], others found no impact on trust [72,83,255], suggesting gaps between the focus of explanations and user needs.

Beyond mixed effects on trust, prior studies have highlighted how existing XAI approaches can sometimes fail to meet users' expectations, and can even lead to unanticipated and unintended negative consequences that have adverse impacts on users [107,196]. Ehsan and Riedl defined these negative downstream effects as "explainability pitfalls" [107]. Prior research has discussed several ways in which pitfalls can appear in explanations [107,319]. For example, that the presence of explanations can sometimes hinder people's ability to identify system mistakes [255]. Additionally, Kaur et al. discovered that the existence of explanations can lead to over-confidence in a model [163]. Placebic explanations that do not provide any useful information can lead individuals to

inappropriately trust the system [109]. Moreover, users often overestimate their understanding of explanations, which can lead to illusionary understanding [78].

In light of these observations, we first conduct a scoping review [14] to identify the primary ways explainability pitfalls arise in explanations. Subsequently, we examine a common pitfall as a central factor and evaluate a mechanism to mitigate this issue.

2.3. Prominent Design Factors for Explanation Design

Multiple design dimensions within explanations shape how useful they are across users and contexts [36,72,96,181,278,302]. In this section, we focus on two factors that are prevalent in the literature and are central to this dissertation: presentation style [36,302] and information depth [72,181,278].

2.3.1. Presentation Style for AI Explanations

The way information is presented in an explanation can impact peoples' perceived comprehension of the explanation [181,282], their understanding of the system [72], and their perceptions of system fairness [36]. For example, hybrid explanations that make use of multiple modalities (e.g., text, visualization, audio) can lead to an improved understanding of the system [270,302], however, there can be tradeoffs between preference and performance (e.g., understanding). For example, receivers' performance can be comparatively worse with their preferred explanation [302].

While different presentation styles have been proposed for training dataset documentation, prior studies with end users have typically examined only a single Q&A format. Moreover, little is known about how users actually interpret and apply

information from these explanations. In this dissertation, we extend prior work by comparing two contrasting presentation styles and investigating how they support users in *critiquing* a system.

2.3.2. Information Depth of AI Explanations

Information depth, which refers to the level of detail and complexity included in the content of an explanation, stands out as a critical design factor as prior work has produced mixed results concerning the impact of information depth on user understanding [72,181,278,293], explanation preference [223], system reliance [50,247], and cognitive load [69,167,181].

These mixed results can be interpreted through Cognitive Load Theory [300], which explains how the amount and type of mental effort required by a task influences users' ability to process information. The theory distinguishes between intrinsic load, which comes from the inherent complexity of the material, extraneous load, which arises from how information is presented, and germane load, which reflects effort devoted to building understanding and mental models [299,301]. Increasing explanation depth can raise intrinsic load by introducing more concepts, dependencies, and relationships that users must process [301]. At the same time, when detailed explanations are well-structured and clearly articulated, they may reduce extraneous load by clarifying system behavior and minimizing ambiguity. Depth can also support germane load by encouraging users to reflect on system behavior and construct more coherent mental models. Since intrinsic, extraneous, and germane load are considered to be additive [242], reducing extraneous load while increasing germane load is only effective if the overall cognitive demands remain within users' processing limits [295,301]. When explanation depth exceeds

available cognitive resources, increases in intrinsic load may not be offset by reductions in extraneous load or gains in germane processing. In such cases, information depth can overwhelm users, increasing the risk of shallow understanding, misinterpretation, or reliance on cognitive shortcuts [242,248,313]. This perspective helps explain why information depth can both improve understanding and increase cognitive effort, and why its effects depend on the amount of information provided and the way it is structured for users.

The above theoretical tension is reflected in empirical findings and recommendations in the literature. For example, some researchers argue that AI explanations should be selective, drawing on properties of human explanation [204], because complete reasoning can overwhelm users [184,223]. Simple explanations are often preferred by receivers [203,223,264], but can lead to incomplete understanding [278]. Other researchers argue for prioritizing comprehensibility of explanations over their compactness (i.e., limiting information to avoid overwhelming users) [197]. In-depth information in explanations can help users to develop the most accurate mental models [181,278], but can also lead to higher cognitive load [69,167,181]. Moreover, when explanations are highly detailed, users may struggle to process all available information within their cognitive limits. As a result, they may rely on cognitive shortcuts [310], such as accepting explanation material without critical reflection or selectively focusing on details that confirm their pre-existing beliefs (i.e., confirmation bias [168]), contributing to overreliance [50,247]. Excessive detail can also lead to false beliefs about the system's accuracy [293]. These mixed results indicate a lack of clear guidelines on how much detail to include in explanations.

As discussed in Section 2.1, prior work on explanations is mostly algorithm-centered, and consequently, studies on information depth have largely focused on algorithm-centered explanations. By contrast, the training dataset explanations we study are descriptive accounts of the data used to train AI systems. Given this difference, it is unclear whether existing findings regarding information depth will generalize to this setting. We address this by comparing two levels of information depth and examining how they affect users’ understanding and use of training dataset explanations.

2.3.3. Summary

This section highlights presentation style and information depth as two key factors that shape users’ comprehension of explanations and their understanding of AI systems. It also demonstrates that much of the empirical evidence on these factors, particularly on information depth, comes from algorithm-centered explanations, leaving open questions about how presentation style and information depth should be designed for training dataset explanations.

2.4. Diversity and Assessment of Users’ AI Literacy

We begin this section by discussing the diversity of users’ AI understanding and experiences, and highlighting the importance of developing validated instruments for assessing AI literacy that can support equitable, effective, and user-centered AI design and evaluation. Next, we review existing assessment tools for users’ AI literacy, noting their strengths and limitations.

2.4.1. Diversity in Users' AI Literacy

The users of XAI systems are very diverse in that they encompass individuals with varying goals, responsibilities, and understanding of AI [27,91,256]. This diversity implies that users' explanation needs, expectations, and interpretations can differ substantially [16,288]. For example, users with formal AI training may focus on model logic and data provenance, while others without technical backgrounds may prioritize clarity, fairness, or decision justifiability [105]. Even within non-expert groups, prior work shows that professional background influences how people perceive and use explanations [31,155,234,283].

In the context of training dataset explanations, prior work suggests that participants who had formal training in AI perceived the explanation differently than people without a CS/Engineering background [8]. While prior studies also discussed the importance of considering the impact of users' AI knowledge and background on their interaction with AI explanations [105], the field still lacks systematic understanding of how users' AI literacy shape users' comprehension, trust, and critical engagement with AI systems and how explanations can support their engagement.

2.4.2. Application for Reliable Assessment Tools for AI Literacy

Given the diversity in users' AI literacy, reliable assessment is essential for both research and practice [188,207]. Such measurement can enable researchers to capture the diversity of users' knowledge [207], compare AI understanding across different populations and usage contexts [105], and evaluate the effectiveness of AI literacy interventions [18,220]. Reliable assessment also supports identifying equity concerns in technology and interaction design [156]. For example, users with limited exposure to AI technologies or

lower digital literacy may be underserved by AI-driven systems [312,338]. Without validated tools, such disparities may go undetected as the gaps remain invisible to researchers and practitioners. Reliable assessment makes such inequities visible, supporting the efforts to design equitable and inclusive technologies [188].

The need for reliable assessment also extends to XAI and broader human–AI interaction [1,2]. For example, prior work suggests that user-facing explanations may be more effective when they account for users’ prior knowledge and mental models [38,96,181], which in turn requires reliable measurement [124,162]. Beyond research and design, reliable assessments can also support a wide range of practical applications [60,285]. For example, in educational and onboarding contexts, they can help identify knowledge gaps [338], guide the design of scaffolding strategies, and monitor users’ progress over time. Our work contributes to these initiatives through a validated instrument that can enable researchers and practitioners to assess users’ AI literacy.

2.4.3. Existing Tools for Assessing Users’ AI Literacy

To support growing calls for validated tools to assess users’ AI literacy [60,172,199], researchers have investigated a wide range of instruments. These include informal collections of questions on AI knowledge and experiences [8,105,231,338], self-report questionnaires [60,187,229,252,323], and structured knowledge tests such as quizzes and multiple-choice assessments [74,147,290,329,339].

Informal instruments range from single-item questions about prior AI experiences [8,231] to a collection of questions about experiences and beliefs [105,338]. Such proxies offer several practical advantages as they are quick to administer and impose minimal

cognitive load on respondents, making them particularly useful in exploratory research, and large-scale surveys. However, while such measures can provide surface-level insights into users' familiarity with AI [8,105], they may not accurately reflect actual understanding or AI literacy.

Lintner's systematic review of AI literacy (i.e., ability to understand, interact with, and critically evaluate AI systems and AI outputs [207,228]) instruments found that most existing instruments rely on self-assessment [199]. However, these instruments vary both in the specific aspects of AI literacy dimensions they focus on and the populations they are designed for. Some focus on cognitive and ethical aspects such as awareness, usage, evaluation, and ethics [323], while others emphasize integrating psychological constructs such as self-efficacy and self-management [60]. In terms of target population, certain instruments were designed to assess students' AI literacy development [229], while others are targeted to non-experts more broadly [187] or for workplace settings [252]. While self-assessment instruments are particularly useful for capturing users' confidence and perceived understanding, these instruments also face well-documented challenges, including response clustering [9], ceiling effects [65], and susceptibility to biases like overconfidence and miscalibration [102,178,285].

To address these challenges, researchers have also developed knowledge-based assessments. For example, Hornberger et al. created a multiple-choice questionnaire (MCQ) to evaluate university students' AI literacy [147]. Other MCQ-based instruments have been created for middle school learners, targeting age-appropriate knowledge and comprehension of AI principles [75,339]. A true/false test has also been created to assess practical AI-related skills among adult learners [290]. While these tools provide more

grounded evaluations of factual knowledge, they often focus on technical knowledge or skills and are typically tailored to educational settings or student populations. As a result, they tend not to capture the broader sociotechnical and ethical aspects of AI literacy, and their applicability to non-student populations remains underexplored.

Prior work across various contexts (e.g., student performance [229], workplace decision-making [252]) has shown that self-assessments and objective knowledge tests are often weakly correlated [100,215]. When used together, self-assessment and objective approaches can offer complementary perspectives [100,287], with the former capturing confidence and perception, and the latter evaluating factual understanding. Combined assessments can also enable the detection of miscalibration, which is the discrepancy between perceived and actual knowledge [157,240], commonly associated with the Dunning–Kruger effect [178]. Despite these advantages, there is a lack of instruments that integrate both perspectives into a single, validated scale to assess users’ AI literacy. We address this gap by developing and validating a unified scale that assesses AI literacy through both self-reported and demonstrated measures. Our work can support diagnostic insight into how people’s self-perceived understanding relate to their demonstrated understanding [178], while also expanding the breadth of AI literacy assessment.

2.4.4. Summary

This section demonstrates that users vary widely in their AI literacy and outlines why reliable assessment of AI literacy is important for research and design. It also highlights the limitations of existing tools including narrow scope and heavy reliance on self-assessment, which constrain their usefulness for designing and studying XAI systems. These gaps motivate the dual-format assessment tool developed in this dissertation,

which combines self-assessed and demonstrated AI understanding in a single, psychometrically validated scale.

2.5. Summary of Related Work

This chapter situates the dissertation within the broader landscape of explainable AI and highlights several open challenges that motivate the research presented in this dissertation. First, although explanations are widely studied as mechanisms to support human-AI collaboration, most work focuses on algorithm-centered explanations, with comparatively limited attention to data-centric transparency. Consequently, there is limited understanding of how training dataset information is communicated to users, despite its importance for interpreting and assessing AI systems. Second, prior work demonstrates that explanation design influences users' understanding, trust, and decision-making, yet findings remain mixed and often context-dependent. In particular, design factors such as presentation style and information depth have been examined primarily in the context of algorithm-centered explanations, leaving open questions about their role in shaping users' interpretation and use of training dataset explanations. These considerations motivate the empirical studies presented in Chapters 3 and 4, which examine how explanation design choices affect users' understanding and assessment of AI systems. Third, the literature highlights diversity in users' AI knowledge and experiences, as well as how this diversity may impact explanation effectiveness. However, understanding these differences has been limited by the lack of robust and validated instruments for assessing AI literacy. This limitation motivates Chapter 5, which develops and validates a dual-format AI literacy scale designed to support more systematic investigation of user-level differences.

Chapter 3

Investigating Users' Usage of Training Dataset Explanations and Impact of Presentation Style

A growing body of research suggests that presentation choices in explanations shape how users interpret, comprehend, and ultimately apply those explanations [36,72,181,270,302]. Much of this work, however, focuses on the influence of presentation styles on subjective perceptions (e.g., clarity, trustworthiness, or user engagement) [36,270,302], while paying limited attention to how presentation affects comprehension, and what users actually *do* with explanations. This is especially true for training dataset explanations [8], which communicate information about the dataset used to train a system, such as its collection, composition, and properties. Understanding how presentation styles impact comprehension and users' ability to apply what they learn from explanations is critical for designing explanation interfaces that support effective human–AI collaboration. The same content may either support meaningful use or leave users uncertain, depending on how it is presented [36,90]. The question, therefore, is not

only whether presentation style matters, but also how it matters and what implications it has for users' understanding and use of explanations.

In this chapter, we examine the role of presentation style in supporting users' usage and comprehension of training dataset explanations and how this, in turn, influences their perceptions of the system that provides the explanations. We compare two contrasting approaches: a narrative-driven Data Story which presents training dataset explanations as a cohesive story, and a structured Q&A approach that organizes the same information as answers to specific questions. While both formats convey the same underlying information, they differ in how they present the information and support user navigation through it. Accordingly, we address the following research questions.

- 1) How does presentation style affect users' information usage and comprehension of training dataset explanations?
- 2) How does presentation style impact users' perceived trust, fairness, and understanding of the system?

To answer these questions, we conducted a user study where we asked participants to critique an automated hiring system based on one of the two presentation styles (Data Story, Q&A) for the training dataset explanation. We analyzed their critique to understand their usage of information, and we examined their subjective impressions of the system to understand how the task of critiquing influenced their perceived understanding, trust, and fairness of the system.

In this chapter, we first introduce the two presentation styles and how they were used to convey the training dataset explanations. We then describe the study and present

key findings. This study was presented at the 2024 IEEE VL/HCC conference and published in its proceedings [9].

3.1. Presentation Style Selection

We chose two different presentation formats to present training dataset explanations. One of these was a Q&A-based approach, which has been used in prior work regarding training dataset explanations and reflects an established structure for presenting such information [8,127]. The Q&A-based approach offers modularity and familiarity, allowing users to quickly locate specific information and engage with the explanation in a direct manner.

As a contrasting format to the Q&A-based approach, we selected a storytelling-based presentation. Our motivation to include such an approach was twofold. First, we aimed to explore a design choice that diverges from the Q&A approach, rather than iterating solely on its variations. This allowed us to avoid converging on a local design minimum [51] and instead examine how a distinctly different format might shape users' interpretation and use of training dataset explanations. Second, we considered storytelling's potential as an explanatory medium. Stories are a fundamental part of human experience [164]. They have long been used to help people make sense of their environment, experiences, ideas, and to communicate with others to create a shared understanding [335]. Storytelling can attract and sustain the interest of audiences and enable them to make meaningful connections [99]. The narrative aspect of stories can create a situated understanding [269], which helps make them accessible to [87] and engaging for a diverse audience [280].

We therefore compare Q&A and storytelling to isolate the role of presentation style in users' understanding of training dataset explanations and their assessment of an AI system.

3.2. Using Data Storytelling in Training Dataset Explanations

Data storytelling refers to the practice of communicating insights from data through a combination of data, narratives, and visuals arranged in a structured way [189]. The products of this process are *data stories*, which are backed by facts and supporting visualizations [189]. The goal of a data story can vary from *informing*, *educating*, *entertaining*, and *explaining*, to *persuading* the audience [164,189]. Data stories engage the audience by reflecting and promoting a point of view through a narrative, which is considered as an integral part of storytelling [298].

There are several potential benefits of using storytelling as a medium for communicating new information [131]. Stories are an effective way to engage the audience as they make people experience the information instead of just consuming it [304]. Narratives and storytelling have been used to communicate a range of scholarly information (e.g., cognitive science [45], healthcare [98,130,138], statistics [250]), with prior work finding that narratives are more engaging and easier to comprehend than traditional scientific communication [45,130]. This is particularly true for non-expert audiences who are already accustomed to narrative formats of communication in the daily life (e.g., mass media, news, interpersonal conversations) [87].

Information about the training *data* is a central component of training dataset explanations [8], where the primary goal is to increase system transparency by *informing*

the user about the formation of training datasets and the training process. The similarities in components, and shared goals of *explaining* and *informing* people about *data* suggest that presenting training dataset explanations as *data stories* could help create explanations that are educational and reflective.

Below we describe how we constructed our data story to present training dataset information. It is important to note that there are many ways to approach data storytelling. Rather than aiming to optimize the data story for effectiveness, our goal was to construct a reasonable starting point that would allow us to explore how a narrative format might impact how users experience and use the explanation content, and how this compares to the Q&A format proposed in prior work [8].

3.2.1. Designing the Data Story Format

To design a data story for presenting training dataset explanations, we built upon prior work on data-centric explanation approaches [8] that introduced a Q&A-based format for communicating training dataset information. This Q&A explanation was designed to operationalize the dataset documentation framework proposed by Gebru et al. [127] and presents key information on aspects of the training data and process as answers to specific questions, including: Collection (the collection process, data sources, and pre-processing of the data); Demographics (the demographic distribution of the data instances); Recommended Usage (list of recommended use cases for the data); Potential Issues (issues and considerations related to the dataset); and General Information (overview information about the dataset including release date, history of usage, and updates).

In developing the Data Story to present the above training dataset information, we first identified common data story patterns and components from the literature [164,189,239,280]. This literature suggests that most good data stories are designed with the goal of *informing*, *persuading*, or *explaining* to the audience [189,239]. The narrative in a data story can vary from having a linear path with no interactivity (author-driven), to a non-linear path with a high level of interactivity (reader-driven), to a hybrid approach. Further, most data stories use annotated graphs and visualizations to convey information.

We decided to have our Data Story follow an author-driven narrative style since prior work has described author-driven narratives as *explanatory* [307,317] and best suited for efficient communication [280]. A disadvantage of an author-driven narrative is that readers cannot ask their own questions about the data [307]. Reader-driven narratives, on the other hand, are more exploratory, but can also require greater literacy about visualizations and the presented topic, as well as greater time and effort to consume [307]. We used a textual narrative interspersed with supporting visualizations. We aimed to remain neutral with communication to allow the audience to draw their own conclusions about the data and system. We guided the reader in the story through scrolling and used annotations to direct readers' attention to aspects that are crucial in the visualizations present in the story.

Our Data story began by communicating the goals of the explanation to the reader to set an expectation of what information would be presented. Figure 1 provides a snippet of how our Data Story began with the context and overview.

Following industry trends, Company “A” is planning to use an automated system based on machine-learning algorithms to inform their hiring decisions.

To train their system (i.e., to teach the system to properly interpret data and learn from it), they used an internal company dataset. We looked at the documentation provided for the dataset and using that knowledge, we can talk about the important properties of the dataset. We discuss the types of data the dataset contains, the individuals in the data, and how the data was collected, processed, and have been used till now.

Figure 1: Setting up the context and overview in the Data Story

We used annotated images and visualizations based on the findings that these elements help comprehension and navigation [29,118,164,239,280,328]. For example, Figure 2 depicts the different visualizations used to present demographic information. We also complemented the textual information with slideshows of infographics in some cases as prior work has found that repetition can help emphasize importance [21]. For example, when describing how data was pre-processed, the Data story first uses a block of text, which is accompanied by a slideshow of infographics and annotations. Figure 3 provides a snippet from a slideshow. The detailed slideshow can be found in the appendix (Appendix A.1).

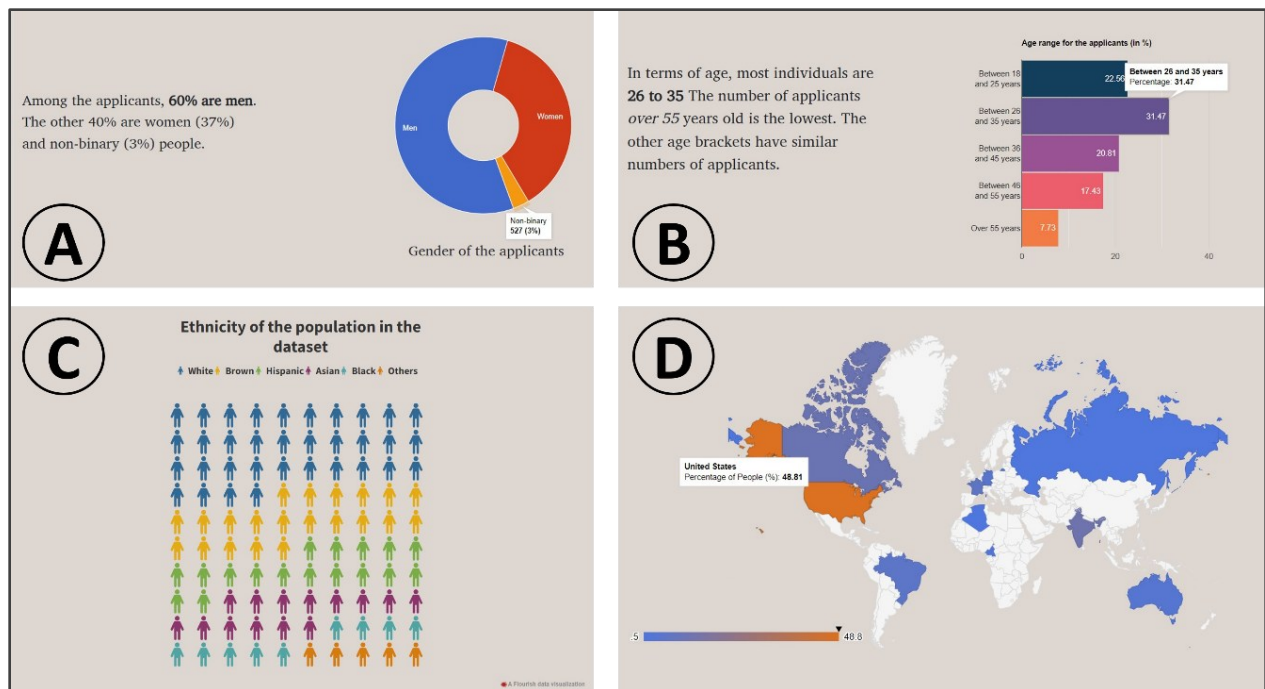


Figure 2: Visualizations used to present different demographic information in the Data Story.

(A) shows the gender distribution, (B) shows the age distribution, (C) shows the ethnicity distribution, and (D) shows the country distribution. The visualizations were accompanied by text annotations that highlighted takeaway information. For example, the accompanied annotation in (A) described that “Among the applicants, 60% are men. The other 40% are women (37%) and non-binary (3%) people”.

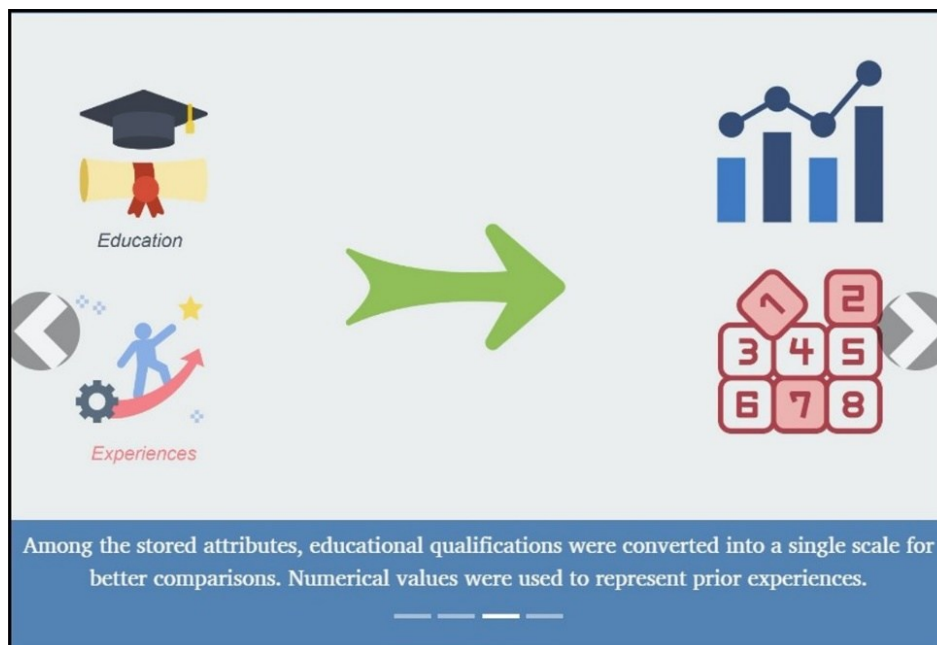


Figure 3: An example of a slideshow accompanying textual information blocks in the Data Story: using infographics to depict the data pre-processing stage.

While our design followed an author-driven structure, we incorporated some local interactivity (e.g., responsive figures and tooltips) to enhance user engagement without increasing complexity. For example, as depicted in Figure 4, the explanation interface provides more information on different data attributes upon hover. This local interactivity allowed users to explore details selectively while remaining guided by the narrative flow [280]. A complete version of our Data Story can be found in the appendix (Appendix A.1).

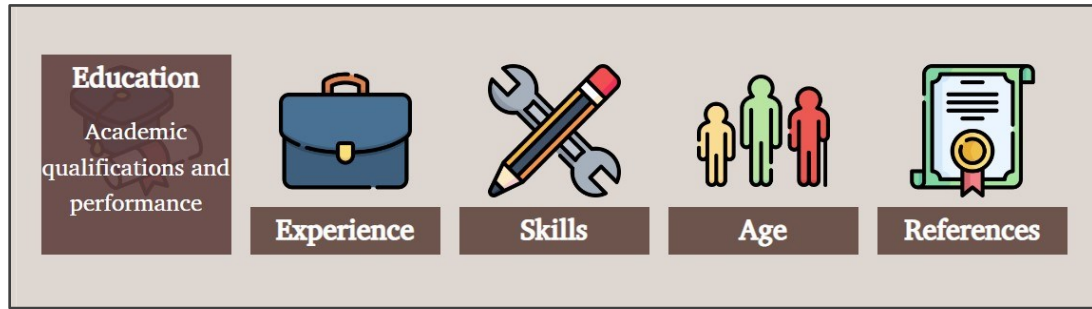


Figure 4: An example of local interactivity in the Data Story: hovering on an image provides a tooltip on the attributes of applicants in the dataset.

3.3. Designing Q&A-based Training Dataset Explanation

We designed the Q&A-based presentation for our training dataset explanation, following the approach previously explored in the literature [8]. The Q&A-based interface is depicted in Figure 5. It presents the same underlying facts about the dataset and training process as the Data Story but organizes the information as answers to specific questions. We, however, chose to omit the “Potential Issues” category from the explanation [8] for the fear that the category name would lead study participants to overemphasize the information in their critique. It should be noted that we also excluded the same underlying facts from the Data Story.

	Collection	Information about the amount of data, the source of the data, the collectors, and the labeling process	▼
	How many instances are in the dataset?	▼	
	The dataset contains data from resumes of 17562 applicants. For each applicant, the dataset contains information about 15 attributes. This includes attributes like - Education, Experience, Skills, Age, and References.		
	Who collected the data?	>	
	What was the data collection process?	>	
	What tools were used in data collection?	>	
	Was any pre-processing done on the data?	>	
	How were the data labeled?	>	
	Any other comments?	>	
	Demographics	Information on gender, ethnicity, age, and country distribution of the instances	>
	Recommended Usage	List of recommended use cases for the dataset	>
	General Information	Overview information about the dataset	>

Figure 5: Overview of the Q&A-based presentation for training dataset explanation [8]. Information about the training data is presented as answers to specific questions, which are organized according to four categories (Collection, Demographics, Recommended Usage and General Information).

The explanation interfaces were developed using HTML, CSS, and JavaScript. The complete explanations are provided in the appendix (Appendix A.1 and Appendix A.2).

3.4. Study

To investigate how presentation style influences users' comprehension and use of training dataset explanations, we conducted a user study where we asked participants to use either the Data Story explanation or the Q&A explanation to critique an automated hiring system. We aimed to observe what information participants use in their critiques, how they comprehend the presented information, and how presentation style impacts how participants use or perceive the information.

3.4.1. Participants

We aimed to include a diverse sample of participants in terms of AI/ML experiences (e.g., ranging from having no formal training or experience with ML to working with machine learning in a professional or academic capacity). Therefore, we recruited participants by posting advertisements throughout a university campus, on different online platforms (e.g., Twitter, LinkedIn), and through snowball sampling (see Appendix A.3 for the recruitment poster). We recruited 42 participants, of which we excluded two participants who did not complete the full study and one participant who relied on Machine translation software for all instructions and text in the explanation. This left us with 39 participants (21 men, 17 women, and 1 non-binary). Our number of participants was influenced by several considerations including data saturation and having a manageable volume of data for qualitative analysis [56]. This number of participants is also in line with other studies on explanations with qualitative data as a focal point of the analysis (e.g., [71,103,165,184,195,197]). Among our participants, 23 were from a CS / Engineering background, and 16 were from a non-CS / Engineering background. In terms of experience with ML and AI, 30 participants did not have any formal training (i.e.,

academic or professional experience) with ML and AI, while 9 did. We also administered an AI literacy scale [323], which consists of twelve 7-point Likert items that cover four core constructs of AI literacy: awareness, usage, evaluation, and ethics. The mean AI literacy score for our participant pool was 4.74 (SD = 0.93) on a scale of 7. The demographics and background questions can be found in the appendix (Appendix A.4).

3.4.2. Study Design and Explanations

Our study employed a between-subjects design with Presentation Style as the independent variable. We had two Presentation Styles: Data Story and Q&A Interface. The design process of both Presentation styles is outlined in Section 3.2 and Section 3.3 respectively. We followed a random assignment of participants to interfaces. After excluding the three participants (two in the Data Story interface, one in the Q&A interface) as described above, our sample included 19 participants in the Data Story interface and 20 in the Q&A-based interface.

3.4.3. Scenario, Task, and Explanation Content

Our study followed a scenario-based approach where we asked participants to use one of the explanation interfaces to critique a fictitious system. Participants were presented with the scenario of an automated hiring system that can shortlist candidates for a company. Hiring system scenarios have been used many times in the XAI literature (e.g., for studying explanations [174], algorithmic trust and fairness [191], and AI ethics [113]). We also chose the hiring domain for its familiarity and importance to a diverse set of people (to ensure access to a broad pool of potential participants).

With the above scenario in mind, participants were asked to imagine that they work in the HR department of a company and that they have been tasked with recommending whether to purchase the system after reading through the training dataset explanation. To inform this decision, participants were asked to use their assigned explanation interface to critique the training data/process and the system. To scaffold and guide participant critiques, we asked them to perform a type of SWOT (Strength, Weakness, Opportunity, Threat) analysis [140], a widely used analytical tool for evaluating organizations, projects, and activities. SWOT is recognized as a valuable decision-making tool for its ability to help people develop a full awareness of the factors involved in making strategic decisions [35]. The use of SWOT analysis in our context was further motivated by the fact that it does not require extensive training or technical skills and that it helps users organize information and stay engaged [334]. We gave participants the following SWOT definitions:

Strengths (S) – Things that are done well in the data/system

Weaknesses (W) – The lacking and flaws that exist in the data/system

Opportunities (O) – Potential ways to improve the data/system

Threats (T) – Elements that could introduce flaws in the data/system

To give participants a sense of critique scope and to encourage participants to consider both negatives and positives, we asked them to try to identify around 10 total strengths, weaknesses, opportunities, or threats and to try to cover each category. We did not impose a strict requirement for the number of comments as we wanted participants to critique the system based on the strengths and weaknesses they saw as

opposed to working to meet a quota. We also added a “General comments” category in case participants had comments that they did not feel belonged to one of the SWOT categories as defined above.

We populated the explanation interfaces with information about a fictitious hiring dataset that was used to train the hiring system. We generated the information based on studies of existing hiring and resume datasets on Kaggle [347]. We ensured that the fictitious dataset and training process included a mix of positives and negatives. The detailed scenario can be found in the appendix (Appendix A.5).

3.4.4. Study Procedure

Our study sessions were conducted either in-person (20 participants) or remotely (19 participants) based on participants’ preferences. The in-person sessions were conducted in a research lab, whereas the online sessions were conducted over Zoom. We leveraged both settings to broaden the participant pool. I was physically present in the lab for the in-person sessions or on Zoom for the online sessions in case participants needed any help. In both cases, I provided help only when a participant made an explicit request (e.g., requesting clarification about the study task or the scenario). Other than participant location, all the other components of the study were identical, including the data collected.

After introducing the study and obtaining informed consent (Appendix A.6), I asked participants to complete a background questionnaire (Appendix A.4) that collected information on participants’ demographics, AI/ML training, and AI literacy. After completing the questionnaire, I introduced the participant to the SWOT analysis and a feedback collection tool (Userback) [346] that allowed them to annotate parts of the

interface and provide SWOT comments. Participants were then presented with the hiring scenario described in Section 3.4.3 and one of the explanation interfaces. Participants were asked to complete the SWOT analysis based on the information in the explanation interface.

After participants interacted with their assigned explanation interface, I asked them to complete a post-study questionnaire (Appendix A.7) containing Likert-scale questions targeting the participant's trust in the system, their perception of fairness, and the degree to which they understood the explanation and felt that they learned about ML. I concluded the study session with a semi-structured interview with each participant to get more detailed insights into their general attitudes towards AI and ML systems, and their perceptions of and interactions with the explanation. The semi-structured interview session was audio recorded and later transcribed for data analysis.

The study sessions lasted 90 minutes on average. Participants were compensated with \$25, which was greater than the minimum hourly wage. The study was approved by University of Manitoba's Research Ethics Board. The approval can be found in the appendix (Appendix A.8 and Appendix A.9).

3.4.5. Data Collection and Analysis

We collected participants' SWOT analysis comments using the Userback tool (see above). We also collected data on how participants felt about the data, the system, and the explanation through a self-reported Likert-scale based post-study questionnaire. Finally, we collected qualitative data from the semi-structured interview sessions with participants, which we audio recorded in both the in-person and online sessions.

My advisor and I were involved in analyzing the SWOT comments. I first iteratively coded [294] the SWOT comments. My advisor and I then discussed the coded data over multiple meetings to reach an agreement on the codes. We used ANCOVA (Analysis of Covariance) to compare aspects of coded SWOT comments with Presentation Style as a between-subjects factor with two levels and participants' AI literacy score as a covariate. We included AI literacy as a covariate given prior work suggesting that people with differing levels of AI knowledge might perceive the explanations differently [8,155,302]. We ran Pearson's chi-squared test to compare the categorical distributions. We used $p = .05$ as the threshold for statistical significance. We reported results with p -values between 0.05 and 0.1 as indication of trend. This approach follows conventions in exploratory research, particularly in studies with limited sample sizes [125,305].

To analyze the Likert-scale data from the post-study questionnaire, we used Quade's non-parametric ANCOVA with Presentation Style as a between-subjects factor with two levels and AI literacy as the covariate.

We did a thematic analysis [41] of participants' quotes from the interview transcripts. I initially coded the transcripts using an open coding scheme. My advisor and I then collaboratively looked for themes in the coded data over several iterations.

Following guidelines on qualitative analysis in HCI research [217], we did not calculate Inter-rater Reliability score for any of our qualitative data coding processes (SWOT comments, interview data) as the primary goal of the coding process was to yield the concepts rather than reach an agreement.

3.5. Results

We first present our results from participants' SWOT analysis, where they critiqued the system based on the training dataset explanation. We present results on what components they focused on in their critique, how they comprehended the presented information, how accurate the critiques were, and how Presentation Style impacted these measures. We further discuss how Presentation Style impacted participants' subjective impressions of the system and the explanations. Finally, we present some additional insights from our interviews. In presenting our results, we refer to the Data Story presentation as "Story" and the Q&A-based presentation as "Q&A" for the sake of brevity.

3.5.1. Analysis of Participants' Critiques from the SWOT Data

We analyzed participants' critiques to see how they used the information presented in the explanation to assess the hiring system. Participants provided a total of 352 comments as part of their SWOT analysis, consisting of a range of feedback on the system, the data used to train the system, the explanation itself, and the high-level process of the study. In our analysis, we focused only on comments that discuss the system and/or its training (including the data and process). Of the 352 comments, 203 met these criteria (102 with the Story and 101 with the Q&A), including one comment that we split into two as it was covering two separate points.

To get an understanding of participants' critique volume, we begin by describing their number of provided comments as well as the comment length. All but 4 of the 39 participants (2 in the Story and 2 in the Q&A) contributed at least one comment about the data or system indicating that most participants tried to use the explanations to

critique the system at least minimally. On average, each participant provided 5.37 comments with the Story (SD = 3.74) and 5.05 comments with the Q&A (SD = 2.91), a difference that was not significant ($F_{1,36} = .113, p = .738$). However, there was a trend ($F_{1,36} = 3.769, p = .060$) suggesting that participants left longer comments with the Story (mean 26.28 words, SD = 15.14) than with the Q&A (mean 17.28 words, SD = 10.55). Interestingly, participants' comment verbosity appeared to match the style of the explanation interface they saw in that the Story interface contained 32% more words (total = 847) than the Q&A interface (total = 613).

Coverage of Explanations Topics in Participants' Critiques

To get a sense of participants' usage of the presented information, we looked at the coverage of explanation content in their critique. To do so, first, we coded the SWOT comments according to the topic (e.g., dataset collection, demographics, data pre-processing). We were interested in whether participants tended to focus on only specific, potentially more accessible parts of the explanations, or if they provided more comprehensive critiques. Figure 6 illustrates the distribution of the discussion topics discussed in the SWOT comments according to Presentation Style and Table 1 provides some illustrative examples of the topics discussed by the participants.

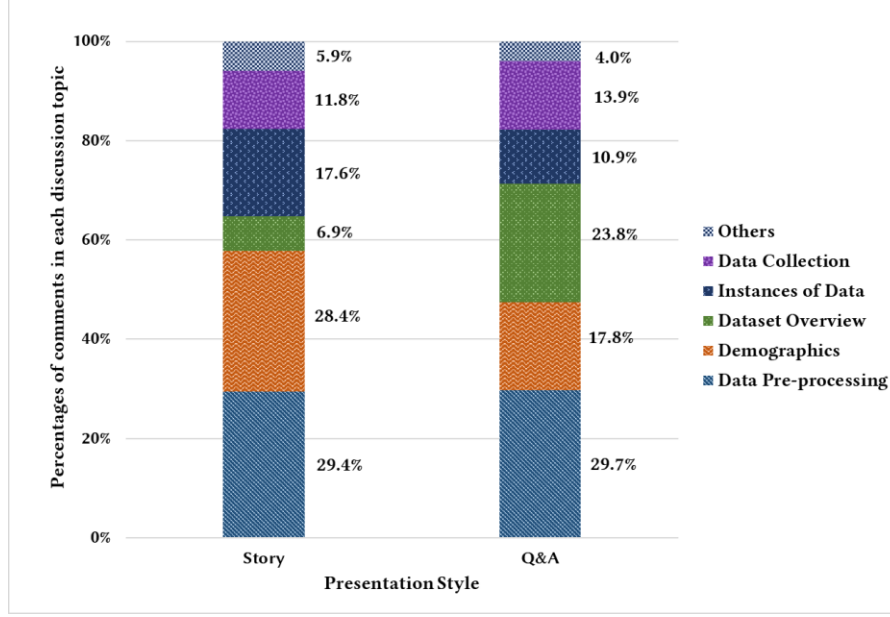


Figure 6: Coverage of explanation topics in the SWOT analysis comments according to Presentation Style.

At a group level, all the topics present in the explanations were covered in the critiques to at least some extent, however, on an individual level, most participants covered only a subset. On a per-participant basis, the mean number of different topics covered was slightly higher with the Q&A (mean 3.2, SD = 1.57) than with the Story (mean 2.84, SD = 1.77); however, the difference was not significant ($F_{1,36} = .449$, $p = .507$).

While presentation style did not impact the number of different topics covered, it did impact the distribution of topics covered in participants' critiques ($\chi^2(5, N = 203) = 14.136$, $p = .015$). In examining the differences in distributions, participants in the Story interface tended to focus more on the *Demographics* and information on the *Instances of the data* (e.g., available attributes, data labels), whereas in the Q&A, participants focused more on *Dataset Overview* information (e.g., dataset size, recommended usage, timeframe of data). When presenting information on *Instances of Data*, the Story used more visuals and interactivity than the Q&A, which might explain why participants focused more on these

topics in their comments. For the *Dataset Overview* information, both interfaces used plain text, however, the Story used a narrative whereas the Q&A provided increased structure by separating the information into answers to different questions. This additional structure might have helped participants in the Q&A use the information as part of their critiques. Participants also mentioned the benefit of the additional structure in the post-session interview.

Table 1: Example comments for each explanation topic discussed by participants in the SWOT analysis.

Discussion Topic	Comments Provided by Participants
Data Pre-processing	The assumption that any blank data points are average [<i>talking about handling missing values with averages</i>] is a dangerous assumption. It can lead to impossible information and may lead to some issues when it comes to real world implementation. (P5, Story)
Demographics	Likely would be considered discriminatory and illegal to use [age] information in a hiring decision, especially to narrow down candidates. (P13, Story)
Dataset Overview	Unless the data is updated there is a chance that [the dataset] could become outdated and not reflect current trends. (P30, Q&A)
Data Instances	I think this is a pretty decent number of metrics used [<i>talking about the number of attributes</i>]. I was expecting less, so 15 should make it better. (P33, Story)
Data Collection	Are we taking into account the bias associated with data collection by an internal committee member? (P10, Q&A)

Identifying Strengths, Problems, and Biases in Participants' Critiques

We also looked at whether participants were able to identify strengths, problems, and biases in the system. In designing the explanations, we tried to ensure that they covered both potentially positive aspects of the training data (e.g., size of the dataset, recommended usage) and aspects that were potentially more negative (e.g., how some of

the data pre-processing was done). In the SWOT analysis, participants identified strengths, weaknesses, opportunities, and threats separately, but we grouped the last three categories together as potential problems since all comments in these categories discussed flaws that already exist or could arise in the system.

Of the filtered 203 SWOT comments, 70 pertained to strengths, and 133 to potential problems or weaknesses. Having a higher number of weaknesses was expected as participants were encouraged to provide at least one comment in each category, and they used the three of the categories to describe problems or weaknesses. While the percentage of comments that were positive was higher with the Q&A interface (39%) than with the Data Story (30%), the difference between the two distributions was not statistically significant ($\chi^2(1, N = 203) = 1.518, p = .218$).

To get further insights into participants' understanding of the information, we analyzed whether participants identified any potential biases. Using types of data biases as described by Mehrabi et al. [218], we additionally labeled the 133 SWOT comments that pertained to problems in terms of whether they were explicitly talking about a bias or discrimination in the system. We found that 53 comments (39.8% of the overall problems identified) pertained to five different data biases: *Representation Biases* (biases arising from how Populations are sampled); *Measurement Biases* (biases arising from the measurement and choice of features); *Selection Biases* (biases arising from a non-random sampling of subgroups); *Temporal Biases* (biases arising from the differences between populations over time); and *Historical Biases* (biases that already exist in the data even given a perfect sampling and feature selection). Figure 7 provides a breakdown of the biases identified for each Presentation Style and Table 2 provides examples.

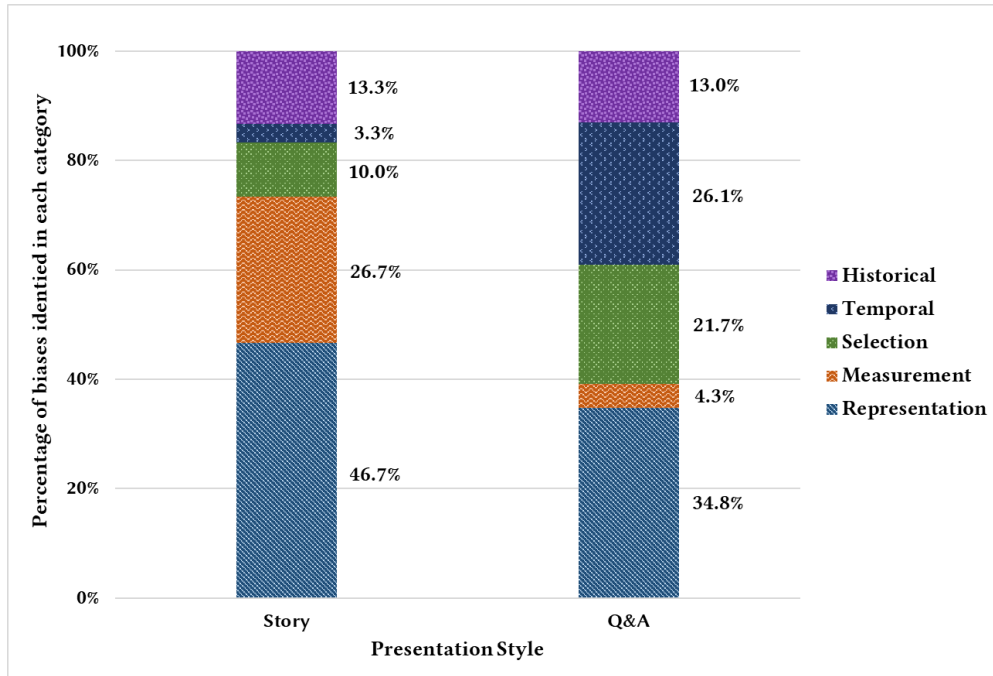


Figure 7: Type of biases identified by participants for each Presentation Style

Participants identified 30 potential biases using the Story (14/19 participants identified at least one bias) and 23 biases using the Q&A (13/20 identified at least one bias). We found a statistically significant difference between the distribution of biases identified in the two interfaces ($\chi^2(4, N = 53) = 10.555, p = .032$), which indicates that the presentation Styles impacted the nature of the topics and biases discussed in the critiques.

Looking at the distributions in Figure 7, we see that *Representation Biases* (in terms of different demographic variables like gender, age, ethnicity, and country) were the most identified bias by the participants. Participants seemed to identify the *Representation Biases* from the presented Demographic information. In both interfaces, all the demographic information was presented using the same visuals. However, the Story presented the information as part of a narrative, whereas the Q&A presented the information as answers to individual questions, requiring more interface operations to

access the same information (i.e., to view the answer to each question). This potentially explains why participants identified more *Representation Biases* in the Story interface. Participants in the Story also identified more *Measurement Biases*, whereas participants in the Q&A identified more *Temporal Biases*. In the Story interface, the pre-processing information was presented with complementary visuals and infographics, which might have helped with comprehension. However, the Story presented the temporal information as part of transitions in the narrative, which might have de-emphasized this information compared to the other presented information, and hence participants identified fewer *Temporal Biases* in the Story.

Table 2: Examples of biases identified by participants in their SWOT comments.

Bias Type	Comments Provided by Participants
Representation Bias	System may learn false 'preference' of applicants who are men instead due to the bias in training set. (<i>P3, Story</i>)
Measurement Bias	In the education attribute, the system may have learnt to prefer candidates with a high GPA. However, if an applicant's prior education did not use the GPA metric, it may cause an unfair averaging or rejection. (<i>P36, Q&A</i>)
Selection Bias	Because the sample will be primarily representative of similar phrasing or shared background, the program later might be discovered unable to cope with more varied information. (<i>P27, Story</i>)
Temporal Bias	This data is 'time-stamped'. Meaning it has been trained on resumes that are of a specific time frame. What about knowledge and experience by those that have worked before 2016. Say someone that had their training completed in 2005? (<i>P16, Q&A</i>)
Historical Bias	[Using previous hiring decision] means that the system is learning and could reproduce the biases of the hiring team at the company, which could skew the AI system to not necessarily return the objective best applicants, just the applicants that [the] group of hiring people liked. (<i>P31, Story</i>)

Critique Accuracy

We were further interested in whether participants' comments contained relatively accurate critiques of the system or alternatively pointed to aspects of the explanations that participants did not seem to understand. To this end, we coded the comments according to three rough accuracy levels: *High understanding*, *Somewhat accurate with misconceptions*, and *Inaccurate*. Table 3 provides some illustrative examples along with a discussion of aspects of the information that was or was not well understood. We found that 54% (109/203) of the comments fell into the category of *High understanding*, 41% (83/203) comments were coded as *Somewhat accurate with misconceptions*, and 5% (11/203) comments were coded as *Inaccurate*. The fact that only 5% of the comments demonstrated complete inaccuracy suggests that both Presentation Styles were at least moderately comprehensible. The difference between the distributions was found to be statistically significant ($\chi^2(2, N = 203) = 12.1937, p = .002$), indicating that at the sample population level, the Story (high understanding: 66%, somewhat accurate: 29%, inaccurate: 5%) contributed to more accurate critiques than the Q&A (high understanding: 42%, somewhat accurate: 53%, inaccurate: 6%).

Table 3: Examples of participants' SWOT comments that demonstrate different types of critique accuracy.

Discussion Topic	Example	Interpretation of What They comprehended
High understanding	<i>Ideally you would have the same number of representatives for each of the groups [of gender]. It's understandable that this is not always possible, but [these imbalances] will add bias in further steps. (P9, Story)</i>	Participants talked about how the imbalanced distribution of demographics might be problematic for the system.
	<i>Generalizing education and experience to single values means that the quality of that education and experience are totally disregarded. (P30, Q&A)</i>	The pre-processing of attributes left the participant wondering about oversimplification and how that can be problematic for the system.
Somewhat accurate with misconceptions	<i>[Converting education and experience to numbers] seems to be the most unbiased part as the applicants can be compared on a single scale. (P15, Story)</i>	Participants saw the pre-processing step as a positive approach while missing the context that this might result in losing the nuances of the quality of the education and experience.
	<i>This is a great value of dealing with missing values. Otherwise, simply discarding them would mean lesser training data. (P10, Q&A)</i>	The participant saw the step of handling missing attributes with average values as a great way to deal with it while not considering the fact that the average values can skew the results.
Inaccurate	<i>[The gender distribution] is almost perfectly balanced. (P12, Q&A)</i>	The participant mentioned the gender distribution was perfectly balanced, when it was skewed towards men.
	<i>Improve on dealing with all missing values. For example, if missing experience attributes causes data to be deleted from the dataset, the recommendation algorithm may automatically reject candidates with no experience. (P36, Q&A)</i>	The participant seemed to mix up the training data and the hiring system that would use the data in their interpretation.

3.5.2. Impact of Presentation Style on Subjective Impressions

After performing the SWOT analysis, participants completed a questionnaire eliciting their reactions towards the system and the explanation. As indicated in Table 4, both explanations were rated highly in terms of understandability, with no significant difference between the two styles. We also did not see any significant impact of presentation style on participants' trust, judgment of fairness, or perception of depth of explanation between presentation styles. There is, however, a potential trend suggesting that participants felt that they learned more about how ML and AI work using the Q&A than when using the Story (Q&A: Mdn = 10.5, IQR = 3.25; Story: Mdn = 8, IQR = 3; $F_{1,37} = 3.716$, $p = .062$).

Table 4: Median (IQR) values for the Likert-Scale questionnaire data for the two Presentation Styles. We report the Median and IQR values as we conducted non-parametric tests. We also provide scale ranges as some measures combine multiple questionnaire items. For every measure where we combined questionnaire items, we calculated the value for Cronbach's alpha and found high internal consistency (all the alpha values were greater than .75).

	Scale range	Q&A	Story	Z	Sig
trust	8-56	42.5 (10.5)	35 (10.5)	2.386	.131
fairness	4-28	18.5 (7.5)	15 (7.5)	2.490	.123
understanding the data	3-21	15 (4)	17 (8.5)	0.003	.960
presentation of information	2-14	12 (2.5)	12 (1.5)	0.384	.539
understanding the explanation	1-7	6 (1.25)	6 (1)	0.790	.380
perception of depth	3-21	16 (4)	15 (8.5)	2.019	.164
learning	2-14	10.5 (3.25)	8 (3)	3.716	.062

Overall, our findings demonstrated the benefits and drawbacks of each presentation style and the nuanced ways participants used the presented information in their critiques.

We summarize our key results in Table 5.

Table 5: Summary of the key results from the study.

Dimension	Presentation Style
Explanation Topic Coverage in Critiques	Participant critiques included more comments on the Demographics and instance level information in the Data Story, whereas the attention was more on Dataset overview information in the Q&A.
Identifying Strengths and Problems	Presentation style did not impact the distribution of strengths and weaknesses identified by participants in their critiques.
Identifying Biases	Representational and Measurement biases were easier for participants to identify with the Data Story, whereas Temporal Biases were easier to identify with the Q&A.
Critique Accuracy	Data Story helped participants create more accurate critiques than did the Q&A.
Learning about AI/ML	A trend suggests participants felt that they learned more about AI/ML with the Q&A than they did with the Data Story.

3.5.3. Interview Findings

The interviews provided additional context to the questionnaire data. We found that participants, in general, perceived both Presentation Styles to be helpful and informative for their purpose of critiquing the system, which at least partially explains the lack of significant results in the questionnaire data. Like prior work [8], the interview data showed strong perceived benefits of the explanations in that participants felt that the training dataset explanations helped them develop informed trust in the system and that such explanations would be particularly important for high-stakes systems. Below we discuss themes that emerged from this study’s interview data.

Training Dataset Explanations as Tools for Critiquing and Learning about AI systems

A large majority of participants reported that training dataset explanations were useful for critiquing the AI system. The explanations improved their understanding of the data and the system, which supported more informed decisions.

I had [some understanding] before, but I think I've gotten a very clear picture now of how exactly it goes from the training data set and then, you know, analyzes that and then it uses that to make decisions. Yeah, I think I've got a much clearer picture of how it works now. (P30, Q&A)

This also led participants to feel like they had more trust in the system.

It helps me trust the system more knowing kind of how it was designed, and I think it's good information to have even for someone who doesn't know really AI systems at all. I think it could help build trust with the system. (P31, Story)

Across both presentation styles, participants stated that the explanations supported learning. They described gaining a clearer sense of how a system's training data is collected, processed, and used to train the system. Participants with prior AI experience, however, reported limited new learning but still valued the explanations as a refresher of what they already knew.

In terms of machine learning, from this specific explanation, I don't think I necessarily learned anything, but it was a very good refresher for what I had already known, so I think if you didn't have a lot of machine learning AI knowledge you would learn a lot.

This was just a good refresher for me for someone with a bachelor's degree, which is good. (P7, Story)

Background and Attitudes to AI/ML Impact Perception of the Presented Information

To get insights into factors influencing participants' perceived trust of the hiring system, during the interviews, we probed into participants' existing levels of general trust or distrust of AI/ML. These discussions revealed a potential relationship between general levels of trust in AI/ML and how participants perceived the explanations. Participants who expressed that they are generally trustful of AI and ML seemed to look for more positives in the explanations, whereas participants who generally do not trust AI/ML tended to focus on identifying problems. Moreover, regardless of the number of strengths or problems they identified in their critique, their overall impression of the automated system seemed heavily influenced by their prior beliefs. For example, in the following quote, a participant who was generally trustful of AI expressed why they trust the system even though they identified more potential problems than strength.

I am going to trust the system [despite the flaws] because of the results [I see] and the new [ML] technologies that we see daily. [I] trust on the instincts of bigger companies and trusting my intuitions, my hands-on experience with ML and how it works, I believe [the system] works. You just need to be careful about the data. (P10, Q&A)

On the other hand, the following quote is from a participant who identified several strengths in the system, but still was skeptical about the system because of their inherent distrust in AI systems.

Yeah, I think I would [be in the] distrusting group but it came through the experience. [...] after learning and discussing different aspects of the challenges and disadvantages of machine learning models and AI systems, I understood okay, there are still questions even if they are smart. (P39, Story)

The above examples are potentially connected to confirmation bias, which leads people to interpret information in a way that is consistent with their existent beliefs and ignore information that is inconsistent with their beliefs [230]. Confirmation bias has also been observed in prior work on how people interact with AI explanations [46,260,302].

We further saw indications of the explanations having the most impact on participants who had a limited familiarity with ML and AI systems. For example, a few participants mentioned that they would blindly trust the system in the absence of an explanation, but that their trust was impacted notably after they interacted with the explanation. For example, in the following quotes, the participant expresses how they would not question the system had the explanation not been present.

I think I would say yes to the system in the case [of not having an explanation], because I would think like, oh gosh, we spend so much time hiring people. Wouldn't it be nice if we could filter a bunch of them out through a computer. (P13, Story)

I would have gone with the system without any thinking [if I did not see the explanation]. I would have bought the system, which is driven by AI, because if I don't have any information [on] what it's doing, I wouldn't care. [Because] anyhow it's making my life easier. (P22, Q&A)

This impact of the explanations seemed to be much more subtle for participants with high familiarity and experience with AI and ML.

Unpacking Users' Interest Towards more Explanation Depth

The depth of the explanations was informed by pilot feedback and prior work [181]. In general, the goal was to have the depth at a level where various interested parties could engage with the explanations, while keeping the explanations accurate and informative. Both presentation styles were perceived positively and none of the participants described the information as being difficult to understand or overwhelming.

While most of the participants were happy with the current depth of the explanation, a few showed interest in more in-depth information for a more comprehensive critique. Looking at the background/ training, we noticed that almost all of this interest for more information was coming from participants who were from a CS/Engineering background or had some prior experience in AI.

I definitely will [want] more depth. For example, I didn't know all of the attributes. It was very important to have all the attributes in this explanation. If someone doesn't want to go into that depth, they can skip the explanation, [...] but having all the attributes there and having detailed information about each of the attributes, I think was very important. (P21, Story)

This perceived lack of depth of information was related to participants having specific questions the explanations did not answer:

It did leave me with some questions about certain things. Like it would be great to see examples of things like when it talks about using averages to fill in missing data, things

like that would be nice to see examples of, because I think it could go maybe a little bit deeper as there were some questions that left me with. (P7, Story)

The data was collected by an “internal” committee within the company. What does it mean? It didn't say anything about how the data should be processed. It says it uses an “automated” process. What does it mean? What exactly were all the “relevant” attributes? I need more information on these. (P32, Q&A)

An interesting tension, however, is that most participants focused their critiques on a subset of the topics covered in the explanations (see Section 3.5.1). Therefore, how to increase depth to help create more accurate critique without providing an overwhelming amount of information is an area for future work. For example, one participant suggested being able to adjust depth on-demand.

I think there could be some sort of level [on] how much information you're putting. Maybe you will have three different explanations, and you will [present] based on their [background information]. For example, if I'm a regular user then you will show me [one] version. [But] if I say I'm a machine learning expert, you will show me more details. (P39, Story)

3.6. Discussion

Our study investigated how receivers of training dataset explanation use the information to critique an AI system and how presentation style impacts their usage of the information. Our findings showed that most participants were able to use the training dataset explanations to produce relatively accurate, but not necessarily comprehensive,

critiques of an automated hiring system. Below we elaborate on our findings and discuss implications for design.

Our findings showed that, at a group level, participants used all the information presented in the explanation in their critiques. However, at an individual level, each participant only used a subset, and this subset was influenced by the presentation style of the explanation. Participants who saw the Data Story covered the demographics and instance-level information (e.g., attributes, data labels) more in their critiques, whereas with the Q&A interface, they focused more on the dataset overview information (e.g., dataset size, usage of data). We also observed differences between the types of biases identified in each presentation style. Participants in the Story identified more *Representation Biases* and *Measurement Biases*, whereas participants in the Q&A identified more *Temporal Biases*. Examining the nature of these differences, we found that the emphasis on repetition, visuals, and infographics in the Data Story appeared to direct user attention, whereas the structured representation directed attention in the Q&A interface. Finally, we saw initial indications that the Data Story interface might have helped participants create more accurate critiques, even though we saw a potential trend that participants felt that they learned more about ML and AI with the Q&A interface.

Our results support existing literature that hybrid explanations (that combine both textual and visual information) result in greater user comprehension than a visual or textual explanation [302]. For example, our Data Story explanation provides an explicit narrative that interspersed text and visuals. The critiques from participants in this condition contained a greater number of comments that demonstrated an accurate interpretation of the explanation material. However, some of our findings also contradict

that literature in that the Data Story explanation might have resulted in a lower perception of learning about ML and AI in comparison to the Q&A explanation. It is possible that the enhanced structure of the Q&A contributed to this perception. It is also important to note that our Q&A was not purely textual – it included some visuals but to a much lesser extent than the Data Story.

Our interview findings suggest that participants' inherent trust or distrust in AI systems might have impacted how they approached the critique (e.g., people who reported inherent trust in the interview seemed to focus more on the strengths compared to participants who did not). More surprisingly, however, is that some participants with inherent trust seemed to feel positively about the system even when their own critique identified more problems than strengths. This suggests that the explanation was not strong enough to combat participants' confirmation bias.

To help combat participants' confirmation bias, it would be interesting to investigate cognitive forcing functions [185] as a tool. Prior work has demonstrated the benefits of introducing cognitive forcing functions in explanations to reduce overreliance [47] or to minimize cognitive biases (which include confirmation bias) [241]. It would be interesting to investigate whether similar mechanisms could be used selectively within longer-format explanations for content that is not naturally drawing users' attention and how different presentation styles can support the process. For example, with the Data Story format, one approach could be to slow down the reveal of new information to allow the readers to reflect on the information they already have viewed. Another approach could be to include persuasive communication techniques within the explanation. While we tried to keep our Data Story as neutral as possible, persuasiveness is generally considered

to be a data storytelling communication goal [189,239], which could be leveraged within explanations. In a more negative direction, such persuasion could also potentially be misused by a designer attempting to convince the user of the merits of an ML system without also considering its limitations [79].

Apart from a trend indicating that participants thought they had learned more about AI/ML with the Q&A interface than the Data Story, we saw a limited statistically robust impact of presentation style on subjective impressions of the system and explanation content. Our interview data, however, suggested that some participants with a stronger AI/ML background (e.g., those from CS/Engineering disciplines or with prior AI/ML experience) wanted more depth and detail to be present in the explanations. Further, there are indications from the literature that AI knowledge can impact participants' understanding of an explanation [155,302]. To account for this at least partially, we used participants' AI literacy score as a covariate in our analysis, however, we did not include AI literacy as an explicit factor in our study design. A challenge in doing so is that it requires effective pre-screening questionnaires since individuals can gain ML / AI experiences through a range of activities and sources (e.g., formal education, working in related disciplines, and media). We used a combination of questionnaire items informed by prior work to characterize our study participants (e.g., educational background [8], formal AI/ML training/experience [8,105], self-reported AI knowledge [105], and an AI literacy scale [323]). In examining participant responses, we found that for many participants, the relationships among these different measures were muddled. For example, some participants with a non-CS/Engineering background, no formal experience with AI/ML, and low self-reported knowledge of AI/ML scored highly on the

AI literacy scale (e.g., over 5.5 out of 7), whereas some participants with prior experience in AI/ML and moderate self-reported knowledge scored in the mid-range on AI literacy scale (between 4 to 5). These inconsistencies highlight the difficulty of reliably characterizing AI literacy, suggesting that future work should refine such measures to better capture how AI literacy influences explanation utility and comprehension.

Our findings do not point to a clear winner between a Data Story presentation or a Q&A-based presentation for explanations that target communicating information on training data but rather point to a range of considerations for designers when selecting a format. For example, while participants reported high levels of trust and sense of system fairness with both explanation formats, if the goal of the explanation is to help a user accurately critique an AI system, a Data Story presentation might be preferable. If the goal is to give people the sense that they are learning about AI/ML, then the Q&A format has advantages. Another consideration is the resources required to produce an effective explanation. Of the two, the Q&A format is likely to take considerably less time to produce, given that does not require developing a narrative or tailored visualizations, but rather well-worded, user-centric responses to specific questions. Further, producing good data stories would benefit from the involvement of people with specific expertise (e.g., storytellers, graphic designers), whereas the Q&A format might be more approachable for those working directly with the training data (e.g., data scientists, data analysts). Therefore, the choice of a presentation format should depend on the explanation goals and the resources and expertise available.

Our results also showed that presentation style impacted the information that participants used in their critiques of the hiring system and the types of biases they

identified. Looking at the reasons behind these differences, we found that these measures were influenced by how a particular piece of information was presented in the explanation (e.g., using text, visuals) and how the explanation supported it. For example, in the Data Story explanation, the pre-processing information was presented with complementary visuals and infographics and participants engaged with this information more in comparison to the Q&A. On the other hand, temporal information was presented as part of transitions in the narrative, which seemed to have the effect of this information being de-emphasized in participant critiques. Therefore, regardless of presentation style, designers can use these properties to prioritize aspects of the training process that they feel are most important for users.

There are limitations to our approach and findings that warrant discussion and further investigation. First, our findings might not be applicable to different AI system scenarios. For example, with lower-stakes systems, users might be less motivated to engage with the explanation [48]. The characteristics of the training dataset might also influence the nature of the findings, for example, some training datasets might be more difficult to describe (e.g., audio recordings or videos), leading to poorer comprehension than we observed. The types of information provided in the explanation might also influence the direction of users' critiques. For example, including a "Potential Issues" category of information [8] might cause users to focus solely on this information in their critiques. We also note that we might see more benefits to a Data Story authored by professional storytellers. Finally, we did not observe significant differences between the presentation styles for some measures (e.g., the questionnaire items). It is possible that more significant differences could emerge with a larger sample size.

3.7. Summary

In this chapter, we examined how the presentation style of training dataset explanations influences users' comprehension, usage of information, and perceptions of an automated hiring system. We compared a narrative-driven Data Story format with a Q&A-based format from prior work in a laboratory-style study with 39 participants. Our findings provide insights on how participants used explanation content to critique the system, and the nuanced ways presentation style shaped participants' usage of the information, the accuracy of their critique, and subjective impressions of an automated hiring system. Our interview findings also provide some insights into how the interpretation of these explanations can potentially be impacted by confirmation bias.

Chapter 4

Explainability Pitfalls: Investigating the Roles of Information Depth and Progressive Disclosure for Training Dataset Explanations

Explanations in AI systems are designed to enhance users' understanding, inform trust, and to support appropriate reliance. Consistent with these goals, findings from our first study (Chapter 3) showed that training dataset explanations helped users critique an AI system. However, prior work has also demonstrated that explanations can sometimes create unintended negative consequences that undermine these goals and can harm users, which they referred to as “Explainability Pitfalls” [107]. Such pitfalls include uncalibrated trust, overreliance, misinterpretation, cognitive overload, and frustration, often stemming from design choices that shape how explanations affect users. Recognizing these pitfalls is therefore essential for designing explanations that genuinely improve users' understanding of and interaction with AI systems.

To build an understanding of explainability pitfalls, in this chapter, we begin with a scoping review of prior literature to identify common pitfalls and their potential causes. Among the identified pitfalls, *Information Depth* (i.e., the level of detail and complexity in an explanation) emerged as a particularly important factor [72,181,278] as prior work has produced mixed results concerning the effects of explanations depth on user understanding [72,181,278,293], explanation preference [223], system reliance [50,247], and cognitive load [69,167,181]. While there are trade-offs for both concise and detailed explanations regarding comprehension and trust, little is known about how these effects manifest in *data-centric* approaches [8,127,257]. Given the longer length and the distinct nature of information in training dataset explanations compared to model-centric explanations, it remains unclear how these prior findings might generalize to data-centric approaches. For example, will in-depth training dataset explanations similarly enhance understanding despite higher cognitive demands, or will their length and complexity overwhelm users, making simpler explanations more practical and preferable?

This lack of clarity suggests that Information Depth may function as an explainability pitfall in training dataset explanations, but its effects and trade-offs remain underexplored. Motivated by these uncertainties, in this chapter, we investigate training dataset explanations that vary in Information Depth and the use of Progressive Disclosure to understand how they affect users' critique of an AI system, their understanding of it, and their cognitive load. Progressive disclosure is a concept from UI design [62,226] that involves hiding advanced interface controls, limiting initial user errors while they are learning the system. In our context, it refers to revealing

explanation components on-demand, while keeping track of what has already been explained. Specifically, we aim to answer the following research questions:

- 1) How does information depth in training dataset explanations affect participants' critiques of an AI system, their understanding of it, and their cognitive load?
- 2) How does Progressive Disclosure influence users' cognitive load and how they navigate and understand training dataset explanations?

To answer these research questions, we conducted a study with 32 participants where we asked participants to utilize the training dataset explanations to critique two automated systems (i.e., assess their potential advantages and disadvantages) with varying levels of Information Depth. Additionally, we examined how Progressive Disclosure of information may help mitigating the adverse effects of potential information overload. This study has been accepted for presentation at the 2026 ACM Intelligent User Interfaces (IUI) conference [11].

4.1. Explainability Pitfalls

Despite many benefits of explanations shown in numerous studies, the impact of AI explanations is not universally positive. Some studies reported no impact of explanations on trust [72,83,255], suggesting gaps between the focus of explanations and users' needs. In some cases, explanations can even prompt users to act against their own interests, leading to unanticipated and unintended negative consequences. Ehsan and Riedl defined these negative effects as “explainability pitfalls” [107]. Understanding these potential pitfalls can help researchers and designers to anticipate and address problems early in

the design process, which might reduce the risk of bias, cognitive burden, or misleading impressions [104,163].

To gain insights into explainability pitfalls, we conducted a scoping review [14] on the literature related to explanations. A scoping review aims to map the breadth of existing research on a broad topic to clarify key concepts, describe the available evidence, and highlight gaps in the literature [14]. We used search terms such as “negative consequences/outcomes/impact of AI/ML explanations”, “explainability pitfalls”, and “harms of AI explanations” on Google Scholar. We also made use of forward and backward searching on the papers we collected from the search. We included papers that explicitly talked about negative consequences of explanations. We also included literature that did not explicitly discuss negative consequences but described how an explanation could potentially result in negative consequences. From the collected papers, we extracted all documented negative consequences. We then grouped the consequences that were similar (e.g., overtrust, undertrust) under one label (e.g., uncalibrated trust). We categorize our list of negative consequences into two broad groups.

1. *Adverse Outcomes*: effects that arise directly from how users interpret and act on explanations, such as overreliance, uncalibrated trust, or cognitive overload.
2. *Design Drawbacks*: issues that originate in the design process and can lead to those outcomes, such as excessive information, lack of actionability, or ambiguity.

Table 6 presents the list of the negative consequences we identified in our scoping review. For the *Design Drawbacks*, we provide illustrative mappings to potential *Adverse Outcomes* that may follow from those design decisions.

Table 6: Potential Negative consequences of Explanations

Adverse Outcomes
Uncalibrated trust [107,163,196]: Receivers of the explanation trust the system without proper evidence of whether or not to trust the system.
Overreliance [25,46,196,255]: Receivers of the explanations tend to rely on the system without any critical reflections (even when the AI system is wrong) and just because of the presence of the explanations.
Cognitive overload [292,293,302]: Receivers of the explanations are cognitively overloaded by too much information and/or too complex information presented in an explanation.
Misinterpretation [9]: Receivers of the explanations interpret the explanations differently from the intended representations and/or meaning which leads to an incorrect understanding and evaluation of the system.
Frustration [107,196,197]: Explanations leave the receiver frustrated owing to a range of reasons (e.g., no clear actions, too much information, complexity of the information, interaction technique) which hurts the user experience.
Design Drawbacks
Distracting Interface [197]: The explanation interface interrupts a user's focus and draws attention away from relevant information in the explanation. This can potentially lead to cognitive overload, misinterpretation, and frustration.
Passive Consumption of information [197]: The explanation lacks the facility to actively engage with it (e.g., interactivity), which can lead to knowledge gaps and result in frustration and uncalibrated trust.
Lack of actionability [107,196,197]: Receivers of the explanations do not have any idea/actions to take to improve system behaviour based on the explanation which leads to frustrations.
Information Depth [197]: The explanation provides a high level of detail and complexity (e.g., lack of clarity, unintuitive information). This can overwhelm users' mental capacity, leading to cognitive overload and increasing the chance of misinterpretation.
Ambiguity [197] : The explanation does not provide enough context to process the information properly, leading to a higher chance of misinterpretation.

Beyond documenting outcomes and drawbacks, we also investigated underlying causes that can explain why these negative consequences occur and how design choices interact with such consequences [47,58,129,184,205,281,293]. Our review highlighted

several key mechanisms rooted in human cognition and information processing: *confirmation bias* [168] (people only focusing on information that supports their existing beliefs), *cognitive shortcuts* [310] (using automatic thinking to make decision-making more efficient), *limited working memory* [300] (the cognitive demand of the explanation exceeds mental capacity), *information or visual literacy* [19,279] (the ability to seek, evaluate, organize, use, and communicate information in various formats), and *affect heuristic* [286] (where people rely on emotions rather than concrete information when making decisions).

These causes are distinct from *Design Drawbacks* but often interact with them. For example, *limited working memory* makes users susceptible to cognitive overload, but this is exacerbated when explanations present excessive detail without sufficient structure or guidance. Prior research has explored ways to mitigate these issues. For example, prior work has found that cognitive forcing functions [47] can guard against cognitive shortcuts and significantly reduce overreliance. Additionally, providing background information (e.g., information required to assess the correctness of a prediction) [129] in explanations can combat overreliance and help calibrate users' trust. Other studies have found that selective explanations (i.e., explanations that selectively present a subset of model reasoning based on recipient's preferences) can also reduce overreliance and improve the subjective perception of the system [184]. However, regarding cognitive overload, prior work has mainly proposed progressive disclosure as a strategy for managing information depth, for example by revealing additional details only when needed [293], but there is a lack of empirical evidence on how progressive disclosure help users navigate explanations and manage cognitive load.

4.2. Studying Information Depth and Progressive Disclosure

From our review of explainability pitfalls (Section 4.1), *Information Depth* emerged as a design drawback with the potential to create adverse outcomes such as cognitive overload, misinterpretation, and disengagement. Information depth can significantly impact user comprehension [72,181], and perceptions of the system [278]. For example, simpler explanations are preferred by receivers [203,223,264], but can result in incomplete understanding [278]. In contrast, detailed explanations can provide a clear mental model of the system [72,181], but demand greater cognitive effort [181]. While prior work has shown that both concise and detailed explanations have trade-offs for comprehension and trust, little is known about how these effects might manifest in training dataset explanations owing the differences in length and content of the explanations. For example, prior work explored four different types of explanations [38,96] (input-influence, sensitivity, case-based, demographic), all of which contained fewer than 100 words. Schoeffer et al. explored different combinations of common explanation types (relevant factors, factor importance, and counterfactual scenarios) [278], with the most comprehensive explanation consisting of 172 words. In contrast, in prior work on training dataset explanations [8], the shortest explanation was over three times this length at 613 words. Given the complex relationship between information depth and the utility of explanations, as well as the typically long-format nature of training dataset explanations, it is unclear how prior findings on impact of information depth will generalize for training dataset explanations.

Beyond what information is presented, how explanations are presented to users plays a critical role in shaping their impact [36,270,302]. As discussed in existing explanation

presentation paradigms [17], including static explanations, where users are presented with all available explanatory information at once [17,38,77], can support rapid formation of initial mental models [181,198], but may also increase cognitive load and overreliance [25,181]. In contrast on-demand, interactive explanations, where users actively choose when and what explanatory content to reveal [37,77,179,330], can support intentional engagement and critical reflection [47], while risking underuse by users who lack motivation [302].

One interface mechanism for enabling on-demand access is Progressive Disclosure, a concept from UI design that involves hiding advanced interface controls, limiting initial user errors while they are learning the system [62,226]. In the context of AI explanations, Progressive Disclosure has been proposed as a mechanism for providing effective transparency by balancing transparency with cognitive manageability [293]. Effective Progressive Disclosure typically requires that (1) information is revealed on demand, (2) explanation content is organized hierarchically, and (3) the system keeps track of the context and information already given to users [293]. In practice, there is limited empirical evidence showing that Progressive Disclosure reliably reduces cognitive load or improves system perceptions. Moreover, presenting explanations on demand may shift cognitive effort from processing content to deciding when additional information is needed, which may introduce new forms of cognitive cost [232,248].

To understand how these effects play out in the context of training dataset explanations, we designed a study examining two factors:

1. ***Information Depth*** — the level of detail and complexity in the explanation.

2. ***Progressive Disclosure*** — an interface strategy that reveals explanation components on demand, potentially mitigating overload while preserving access to details.

Our study systematically varied these factors, with two levels of *Information Depth* (detailed vs. summary) and two levels of *Progressive Disclosure* (present vs. absent). In a laboratory-style controlled study with 32 participants, we asked participants to use training dataset explanations reflecting these variations to critique two fictitious systems. The critiquing task was similar to the one described in Chapter 3, applied to two system scenarios. This design allowed us to examine how these factors shaped participants' critiques, understanding, perceptions, and cognitive load, as well as whether Progressive Disclosure helped manage cognitive load.

4.2.1. Participants

We recruited participants through various channels, including advertisements across a university campus, on different online platforms such as X (formerly Twitter) and LinkedIn, and through snowball sampling (see Appendix B.1 for recruitment poster). To ensure diversity in participants in terms of AI knowledge and background, we used a questionnaire to pre-screen participants. The questionnaire included questions on participants' academic background (computer science/ engineering/ non-engineering) and formal AI training (formal courses taken on AI/ML). We also included the AI literacy scale [323], which consists of twelve 7-point Likert items that cover four core constructs of AI literacy: awareness, usage, evaluation, and ethics. We recruited 34 participants but excluded two who did not follow the study procedure or complete the full study. Therefore, our final sample included 32 participants (14 men, 17 women, 1 non-binary).

Regarding experience with ML and AI, 24 participants had no formal training (i.e., academic or professional experience) with ML and AI, while 8 did. In terms of academic background, 17 had a CS background, 6 had an Engineering background, and 9 came from non-CS / Engineering fields. The mean AI literacy score for our participant pool was 5.6 (SD = 0.73) out of a maximum of 7. The pre-screening questionnaire and the demographics questions asked to participants can be found in the appendix (Appendix B.2 and B.3).

4.2.2. Study Design

We included two main factors in our study.

- i) *Information Depth*: Summary vs. Detailed
- ii) *Progressive Disclosure*: Present vs. Absent

Information Depth refers to the volume of information provided in the explanation. We defined two levels for this factor. In the *summary* version, the explanation offers high-level, concise information about the training data, while in the *detailed* version, the information is presented in a more elaborate and precise manner. Section 4.2.3 describes the information present in the two levels of this factor. The second factor is the *Progressive Disclosure* of information [293], which is a mechanism that enables transparency “on demand”. We also defined two levels for this factor: the *presence* or *absence* of *Progressive Disclosure*.

Our study had a mixed 2 x 2 design with *Information Depth* (*summary*, *detailed*) as the within-subjects factor and *Progressive Disclosure* (*absent*, *present*) as the between-subjects factor. In other words, each participant interacted with explanations with both levels of

Information Depth and one level of *Progressive Disclosure*. The order of *Information Depth* was counterbalanced across participants. We chose *Information Depth* as a within-subjects factor to enable participants to directly compare *summary* and *detailed* explanations. In contrast, to keep the study length manageable for participants, we treated *Progressive Disclosure* as a between-subjects factor.

4.2.3. Explanation Interfaces and Content

The design of our training dataset explanation was consistent with the design of the Q&A-based approach described in Chapter 3. We organized the explanation into four distinct categories of information. These categories include: i) *Collection* (the collection process, data sources, and pre-processing of the data); ii) *Demographics* (the demographic distribution of the data instances); iii) *Recommended Usage* (list of recommended use cases); and iv) *General Information* (overview information about the dataset, including release date, history of usage, and updates).

For the *detailed* explanations, we used the same Q&A design and content as Chapter 3 (Section 3.3). The explanations are depicted in Figure 8(A) and Figure 8(B). The questions and answers include concrete information on sample size, attributes, data collection process, data pre-processing process, demographic information in multiple dimensions (e.g., age, gender), recommended use, release dates, history of usage, and updates.

The *summary* versions distilled each of the four categories into a short, focused paragraph containing only the core message (see Figure 8(C) and Figure 8(D)). For example, regarding the data collection process, the *detailed* explanation included how an

automated extraction tool was initially used to gather and pre-process the data, followed by manual evaluation of a subset to verify the accuracy. In contrast, the *summary* explanation simply stated that the data collection process combined automated and manual approaches.

Regarding *Progressive Disclosure*, when it was *present*, we applied the principles of Progressive Disclosure (as described in Springer and Whittaker [293]) within the interface. This was achieved by adding interactivity, such as keeping only one category open at a time and requiring users to click on questions to view the answers. Once opened, answers remained visible for reference and to help users to maintain context. This design aimed to reduce the amount of information visible at any moment and encourage intentional exploration. Our approach is consistent with existing literature on how to achieve progressive disclosure [291,348], prioritizing a design that allows users to selectively explore information aligned with their interests.

In the *absence* of progressive disclosure, all information was visible by default, without the additional interactivity required to view content. This allowed users to view the full information set without interaction costs.

Figure 8(A) and Figure 8(C) depict the two explanations where progressive disclosure was *present*, whereas Figure 8(B) and Figure 8(D) demonstrate versions where progressive disclosure was *absent*. Complete versions of the explanations can be found in the appendix (Appendix B.4).

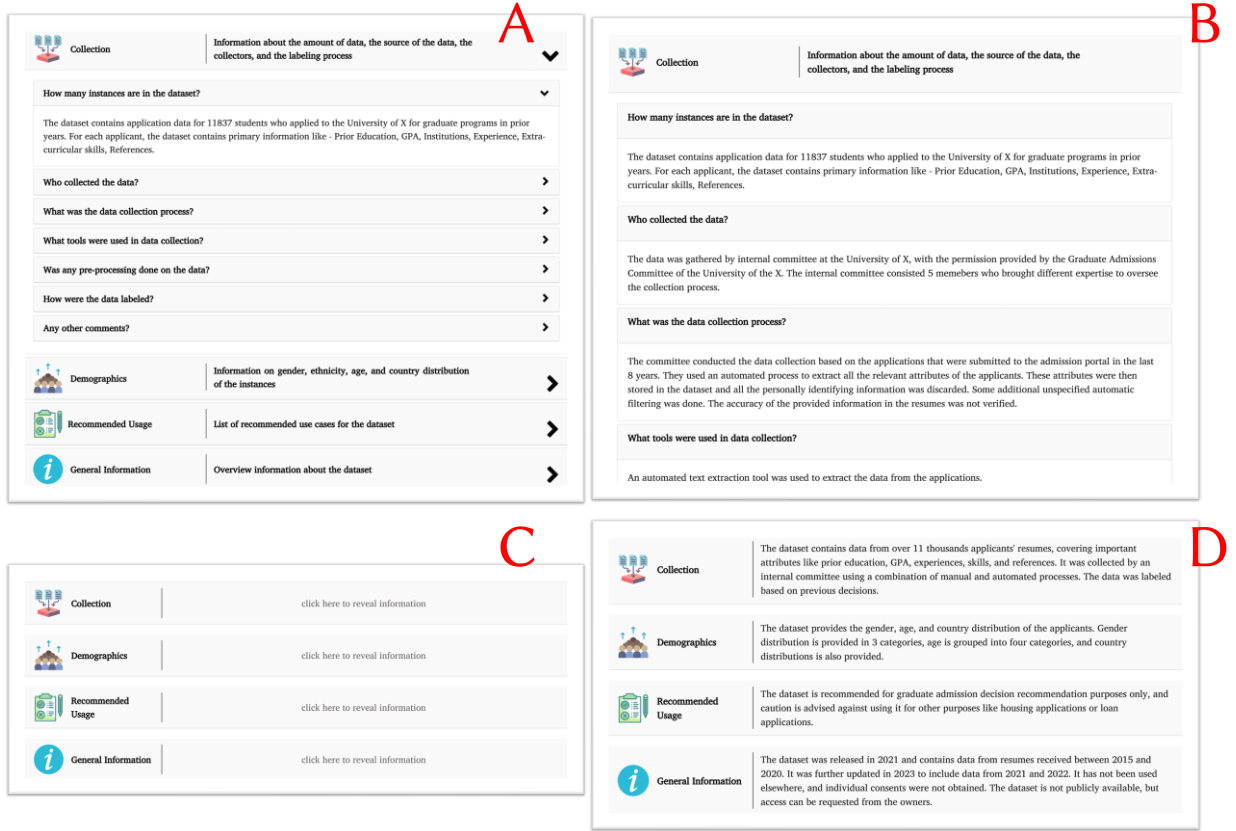


Figure 8: Snippets from the different versions of training dataset explanations. (A) demonstrates a *detailed* explanation with progressive disclosure *present* whereas (B) demonstrates a snippet from a *detailed* explanation with progressive disclosure *absent*. (C) demonstrates a *summary* explanation with progressive disclosure *present* and (D) demonstrates a *summary* explanation with progressive disclosure *absent*. All explanations can be found in full in the appendix (Appendix B.4).

4.2.4. Scenario and Task

Similar to Chapter 3, we used a scenario-based approach [273] where we asked participants to critique two AI systems based on training dataset explanations with each level of *Information Depth*. We used two scenarios from prior studies on training dataset explanations [8,9] with the aim of having scenarios of similar significance and familiarity to a diverse audience: an automated hiring system scenario and an automated admission system scenario.

Participants were asked to imagine that they work in the HR department of a company (for the automated hiring scenario) or are part of the recruitment committee at a university (for the admission system scenario) and they were tasked with utilizing the training dataset explanation to recommend whether to purchase the system. As part of this recommendation, participants were asked to use the information in the explanation to critique the training data and the system. To guide participants in the critique, we used a simplified version of the SWOT analysis technique (used in Chapter 3), where we asked participants to identify strengths and weaknesses of the data and the system. This adjustment was motivated by the observation in the first study (Section 3) that participants used all of the *Weakness*, *Threats*, and *Opportunities* categories of SWOT interchangeably to highlight potential problems in the system. Simplifying the framework therefore reduced redundancy and simplified the task for them. To encourage in-depth critiques, we asked participants to provide around eight comments with a minimum of one strength and one weakness, however, we did not impose a strict limit and told them to critique as they saw fit. We counterbalanced the scenario order across participants. The detailed scenarios can be found in the appendix (Appendix B.5).

4.2.5. Study Procedure

We conducted our study sessions remotely over Zoom, with me present to facilitate each session. Figure 9 provides an overview of our study procedure. Following the introduction of the study and obtaining informed consent (Appendix B.6), I asked participants to complete a brief demographics questionnaire. Subsequently, I provided them with the task description and demonstrated the tool (Userback) [346] that enabled them to annotate interface elements and enter critique comments. Participants were then

presented with one of the scenarios and an explanation interface (either *summary* or *detailed*), which they used to critique the system. Upon completing their first critique, participants completed a post-scenario questionnaire comprising Likert-scale questions assessing their trust in the system, perception of fairness, perceived understanding of the explanation, perceived learning about ML, and cognitive load. The post-scenario questionnaire also contained five comprehension questions (designed to gauge participants' understanding) about the data and the system based on information found in the explanations: two multiple-choice questions (e.g., Was there any update provided to the dataset?) and three open-ended questions (e.g., What kind of processes were used in data collection?). The post-scenario questionnaire can be found in the appendix (Appendix B.7)

After completing the first scenario, participants proceeded through the same process for the second scenario with an explanation interface with the other level of *Information Depth*. Upon completion of both scenarios, I conducted a semi-structured interview with each participant to gain more detailed insights into their critique and their perceptions of the differing levels of *Information Depth*. The study sessions lasted 2 hours on average and participants were compensated with \$30. The study was approved by University of Manitoba's Research Ethics Board (see Appendix B.8 for the approval).

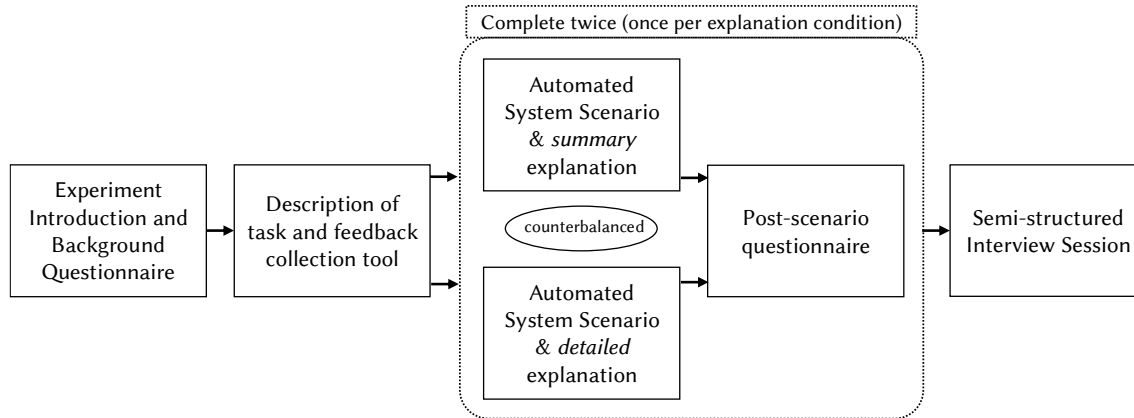


Figure 9: Overview of the study procedure. This procedure was the same for both levels of Progressive Disclosure (a between-subjects factor).

4.2.6. Data Collection and Analysis

We collected data from two primary sources. The first was questionnaire responses on how participants felt about the data, the system, and the explanation through a Likert-scale based post-scenario questionnaire. Additionally, we collected data on their cognitive load via the NASA TLX [137]. We used Wilcoxon Signed-Rank test to analyze the questionnaire data. We calculated Cronbach's α for measures with combined items to check for internal consistency, using a threshold of 0.70 to determine acceptable reliability. We used a non-parametric test since we did not assume a normal distribution.

The second primary data source was participants' critique comments. My advisor and I were involved in analyzing the critique data. I first coded the data following the coding process described in Chapter 3, categorizing each critique based on the specific information they referred to within the explanation and their accuracy. We then discussed the coded data over multiple meetings to reach an agreement on the codes. In line with qualitative analysis guidelines in HCI research [217], we did not calculate an

Inter-rater Reliability score for our qualitative data coding process, as the goal of was to generate meaningful concepts rather than to achieve consensus.

When comparing means, we used a repeated measures ANOVA with *Information Depth* as the within-subjects factor and *Progressive Disclosure* as the between-subjects factor. We ran Pearson's chi-squared test to compare the categorical distributions. We used $p = .05$ as the significance threshold and $p < .1$ as an indication of a trend.

As a supplement to these primary data sources, we collected qualitative data through semi-structured interviews, which we audio recorded and later transcribed for analysis. We conducted a directed qualitative content analysis [149] to examine participants' discussions of *Information Depth* and *Progressive Disclosure*. Following a deductive approach, I coded relevant comments in which participants explicitly talked about these factors. My advisor and I then discussed the coded data over multiple meetings to refine the findings. We chose this approach to provide additional qualitative context to our critique and questionnaire data while maintaining a focused exploration of the study variables.

4.3. Results

4.3.1. Impact of Information Depth and Progressive Disclosure on User Perceptions

As illustrated in Table 7, *Information Depth* impacted participants' subjective impressions of the utility of the explanation content and their impressions of the system. Participants felt they understood the data more with the *detailed* explanation (Mdn = 17, IQR = 3) than they did with the *summary* explanation (Mdn = 11.5, IQR = 10.75; $Z = 4.155$, $p < .001$).

Figure 10 shows that there was much greater variability for this measure with the *summary* explanation than there was with the *detailed* version. With the *detailed* explanation, participants further perceived the information as more sufficient to critique the system ($Z = 3.987, p < .001$) and felt that they learned more ($Z = 3.987, p < .001$). The *detailed* explanation also impacted participants' perceptions in the system, in that they reported higher levels of trust in the system ($Z = 3.091, p = .002$) and that the system was fairer ($Z = 3.913, p < .001$). Finally, unsurprisingly, participants reported higher perceptions of depth ($Z = 3.086, p = .002$) and cognitive load ($Z = 2.872, p = .004$) with the *detailed* explanation. That the medians are on the low end of the 7-point scale suggests that even the *detailed* explanation was not perceived as too detailed or overwhelming.

Table 7: Median (IQR) values for the Likert-scale Questionnaire data based on *Information Depth*. We report the Median and IQR values since we did not assume a normal distribution of the Likert-scale data and conducted non-parametric test. We also provide scale ranges as some measures combine multiple questionnaire items.

	Scale Range	Median (IQR) - Summary	Median (IQR) - Detailed	Z	Sig
Trust	8-56	31.5 (8.75)	36.5 (7.75)	3.091	.002
Fairness	4-28	13 (6.75)	16.5 (5.75)	3.913	<.001
Perceived understanding of the data	3-21	11.5 (10.75)	17 (3)	4.155	<.001
Perception of depth	1-7	1.5 (2)	3 (2)	3.086	.002
Sufficient information to critique	3-21	9 (6)	15 (6)	3.987	<.001
Perceived learning	2-14	6 (5.75)	9 (3.75)	3.355	<.001
Cognitive load	1-7	2.5 (3)	3.5 (2)	2.872	.004

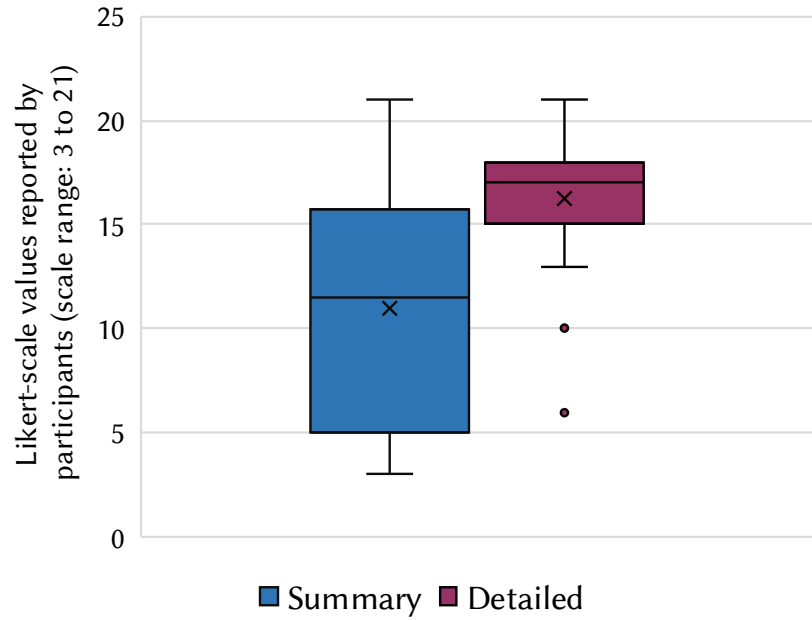


Figure 10: Participants' perceived understanding of the data across *Information Depth* (the sum of responses to three items)

In terms of *Progressive Disclosure*, as both conditions provided the same information, expectedly so, there were no differences in the first five measures in Table 7. However, as demonstrated in Table 8, there was a statistically significant impact on participants' perception of *learning* ($Z = 2.515$, $p = .012$), with participants' perceived learning being higher when *Progressive Disclosure* was *present* (Mdn = 8.75, IQR = 3.25) than when it was *absent* (Mdn = 6.75, IQR = 2.5). The increased perception of learning with the *presence* of *Progressive Disclosure* supports findings from earlier research which showed *Progressive Disclosure* affords more efficient learning progress [62]. For cognitive load, none of the explanations induced a high cognitive load on participants (each with a median of 3), and the use of *Progressive Disclosure* did not statistically significantly impact this measure.

Table 8: Median (IQR) values for the Likert-scale Questionnaire data based on *Progressive Disclosure*.

	Scale Range	Median (IQR) for PD: Absent	Median (IQR) for PD: Present	Z	sig
Trust	8-56	35 (5.385)	34 (7)	0.472	.637
Fairness	4-28	14.25 (5.63)	14.25 (3.5)	0.454	.650
Perceived understanding of the data	3-21	14.5 (4.38)	14 (6.88)	0.264	.809
Perception of depth	1-7	2 (2.25)	2.5 (2.75)	1.031	.303
Sufficient information to critique	3-21	11 (3.63)	13 (4.5)	1.569	.117
Perceived learning	2-14	6.75 (2.5)	8.75 (3.25)	2.515	.012
Cognitive load	1-7	3 (1.38)	3 (2)	0.609	.543

4.3.2. Impact of Information Depth and Progressive Disclosure on Participants' Critiques

To understand how *Information Depth* and *Progressive Disclosure* impacted the nature of participants' objective assessment of the system, we analyzed their critique comments across three dimensions: the volume of their critique comments, the breadth of their critique topics, and the accuracy of their critiques.

Critique Data Amount

Participants provided a total of 544 comments as part of their critiques. These comments included feedback on the system, the data used in the explanation, and the presentation of the explanation. To gain a sense if participants' critique amount would change between

explanations, we compared the number of comments they have made for each explanation they interacted with during the study.

Figure 11 provides the average number of comments according to *Information Depth*. We saw a main effect of *Information Depth* on the number of comments provided ($F_{1,30} = 14.951$, $p < .001$), where participants provided more comments in the *detailed* explanation (mean = 9.84, SD = 4.31) in comparison to the *summary* explanation (mean = 7.16, SD = 2.46). This suggests that participants provided more thorough feedback on the system when presented with *detailed* explanations. The effect of *Progressive Disclosure* on the number of comments was not significant ($F_{1,30} = 0.943$, $p = .339$) nor was the interaction effect of *Information Depth* $\times$ *Progressive Disclosure* ($F_{1,30} = 0.978$, $p = .33^2$). We further analyzed participants' average critique length (i.e., the average number of words in a comment) but we did not see any statistically significant effects ($p \geq .401$).

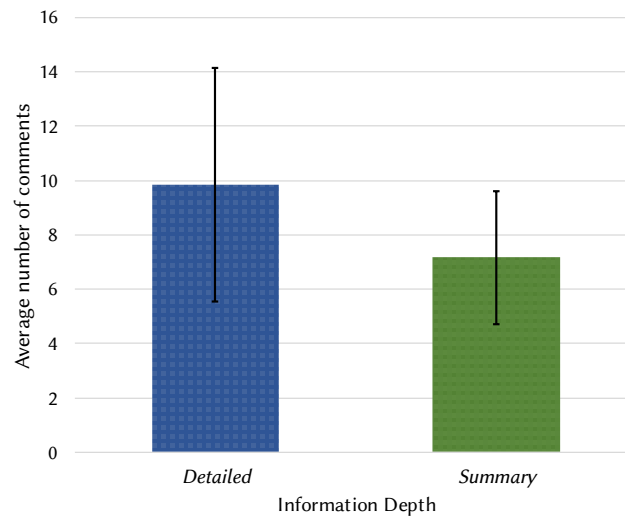


Figure 11: Average number of comments per participant across *Information Depth*. Error bars show standard deviation.

² For the cases where we used repeated measure ANOVA, we report the interaction effects as the test allows for this reporting.

Critique Data Coverage

We analyzed participants' critiques in terms of the coverage of topics presented in the explanation. For topic coverage, we label each comment in terms of the high-level categories presented in the explanation (Section 4.2.3). These categories include *Data Collection*, *Demographics*, *Usage*, and *General Information*. Some participants' comments focused on the overall dataset and did not fit under these categories. We labeled such comments under a "Overall" category which resulted in five comment topics (*Data Collection*, *Demographics*, *Usage*, *General Information*, *Overall*).

At a group level, all the topics were covered in participants' critiques for both levels of *Information Depth*, supporting previous findings that participants collectively find value in all the presented information [9]. On an individual level, we saw a trend concerning *Information Depth* ($F_{1,30} = 3.817$, $p = .06$) where participants covered a slightly higher number of topics with the *summary explanation* (mean = 3.78, SD = 1.18) than with the *detailed explanation* (mean = 3.28, SD = 0.99). There was no significant main effect of *Progressive Disclosure* ($F_{1,30} = 0.18$, $p = .674$) on the number of topics covered nor a significant *Information Depth* $\times$ *Progressive Disclosure* interaction ($F_{1,30} = 0.537$, $p = .469$).

We saw a statistically significant difference in the distribution of topics covered in participants' comments according to *Information Depth* ($\chi^2(4, N = 544) = 41.41$, $p < .001$). Figure 12 provides an overview of the distribution of topics and Table 9 provides some illustrative examples. In the *detailed explanation*, more than half of the comments focused on *Data Collection* whereas with the *summary explanation*, topic coverage was more balanced. This indicates that participants not only utilized the additional depth on the data collection process (e.g., specific details on collection tools and data pre-processing)

in the *detailed* explanation, they also prioritized this information over other topics. Although additional details were available across multiple topics, participants did not use this information as much. This could either be because they found the *Data Collection* aspect to be the most critical or because the *Data Collection* information was presented as the first category. In contrast, with the *summary* explanations, participants spread their critiques more evenly across multiple topics. *Progressive Disclosure*, on the other hand, did not significantly impact the distribution of topics covered ($\chi^2(4, N = 544) = 5.35$, $p = .253$).

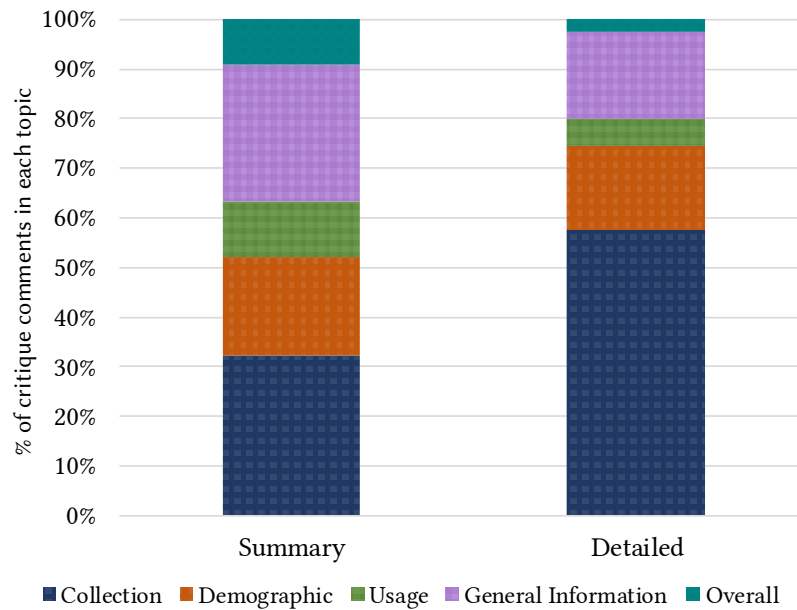


Figure 12: Distributions of topics in participants' critique across *Information Depth*.

Table 9: Illustrative examples of participants' critique across *Information Depth*.

	Summary	Detailed
Collection	This is a large dataset which is helpful in developing accurate models. (P25)	Combination of manual and automated processes for data collection indicates a balanced approach that leverages both human expertise and technological efficiency in data collection. (P24)
Demographics	This transparency from demographics information enables the system to monitor and potentially mitigate biases in hiring decisions. (P23)	This does not reflect the general population, so you are training your system on non-representative data. This could lead to bias in its learning. (P32)
Usage	Clear guidance for usage ensures appropriate and ethical utilization, and helps prevent harm. (P24)	Good explanation on what this should be used for. (P13)
General Information	It is up to date and contains resumes received between 2015 and 2022. (P18)	No individual consent raises ethical issues and invasion of privacy. (P16)
Overall	Good summary to quickly explain but could use a bit more information regarding the AI. (P21)	Overall, the dataset, while might have its flaws, is great in assessing qualifications, experience and skills. (P6)

Critique Data Accuracy

We also coded the accuracy of the critique data to investigate if participants seemed to understand the information they were using as part of their critiques. We followed the coding process described in Chapter 3, where we coded participants' critique comments on a 3-point scale (accurate, somewhat accurate, inaccurate). While we could label majority of the comments using this scale, some of the comments contained a question or stated uncertainty about the dataset. We labeled such comments as "Uncertain".

Figure 13 depicts a distribution of the critique accuracy across *Information Depth*. We found a statistically significant difference in the distribution of the critique accuracy

across *Information Depth* ($\chi^2 (3, N = 544) = 22.715, p < .001$). Looking at the differences, we noticed that the participants provided slightly more accurate and somewhat accurate comments with the *detailed* explanation (accurate: 241/315, somewhat accurate: 58/315) than with the *summary* explanation (accurate: 153/229, somewhat accurate: 36/229). We further observed that participants showed more uncertainty in the *summary* explanation (34/229) than in the *detailed* explanation (15/315). Finally, we found only one comment in the *detailed* explanation to be completely inaccurate in comparison to the six comments in the *summary* explanation. The fact that only 7 out of 544 comments were rated as to be completely inaccurate suggests that both the explanations were at least moderately comprehensible for participants. This further supports earlier findings in Chapter 3 that participants can productively use training dataset explanation to generate reasonably accurate critique comments. Here we demonstrated that participants can also achieve reasonably accurate critiques even with a summarized version. Table 10 provides some illustrative examples of critique comments in terms of their accuracy.

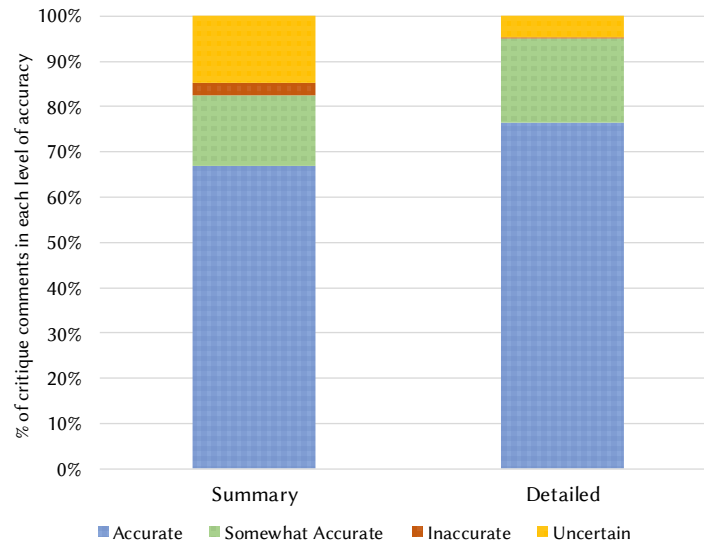


Figure 13: Distribution of the accuracy of participants' critique comments across *Information Depth*

Table 10: Illustrative examples for the critique accuracy ratings with researchers' interpretation of the critique comment in italics.

	Summary	Detailed
Accurate	This seems like a generally large sample, N=11k, which is good. (P33) <i>Comment on large sample size being helpful for training.</i>	Using average values for missing attributes might not provide accurate representation for individuals. (P20) <i>Comment on the concerns of using average values for missing data.</i>
Somewhat accurate	Restricting the dataset's usage to only one specific purpose may limit its potential applications and insights that could benefit other areas. (P24) <i>Comment is somewhat correct in that it seemed to ignore the risk of using the same data in improper contexts without the usage guidance.</i>	This information is unnecessary. Knowing one's ethnicity could potentially introduce bias (both positive and negative). (P8) <i>Comment is somewhat correct as it seemed to ignore the fact that having such demographic information in the dataset might be necessary to understand the representativeness of the dataset..</i>
Incorrect	How would the individuals know their data was used? This information does not serve any purpose. (P18) <i>Participant incorrectly assume that the information on previous use cases is for the individuals whose data are in the dataset and not for the potential consumer of the dataset.</i>	If the data from applicants below 10 years old were used to train decision making for adult applicants, that would be questionable. (P5) <i>Participant misunderstood a part of the explanation that mentioned that 9% of the applicants came from the age group of over 55 years as applicants being less than 10 years old.</i>
Uncertain	This lacks information. Which 3 genders? Which 5 categories? How does ethnicity and country affect the application? (P13)	How do we know that this tool is reliable and accurate? What if it messed up data during the extraction process? (P11)

In terms of *Progressive Disclosure*, Figure 14 provides an overview of critique accuracy across the conditions. Participants provided slightly more accurate critiques when *Progressive Disclosure* was *absent* (213/288) compared to when it was *present* (181/256). A similar pattern was observed for critiques that showed uncertainty (31/288 for *absent* and 18/256 for *present*). In contrast, somewhat accurate critiques were slightly more common

when *Progressive Disclosure* was *present* (52/256) compared to when it was *absent* (42/288). Finally, the number of incorrect critique was also higher when *Progressive Disclosure* was *present* (5/256) than when it was *absent* (2/288). Overall, the differences in the distribution of critique accuracy were not statistically significant across *Progressive Disclosure*, and we only observed a trend ($\chi^2(3, N = 544) = 6.538, p = .09$). Nevertheless, the observed trend points to a possible influence of *Progressive Disclosure* on critique accuracy, which warrants further investigation with a larger sample.

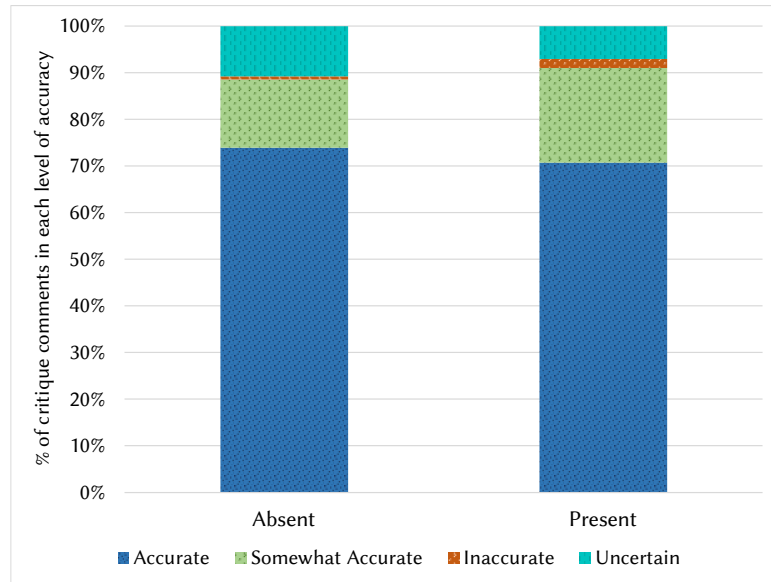


Figure 14: Distribution of the accuracy of participants' critique comments across *Progressive Disclosure*.

4.3.3. Impact of Information Depth and Progressive Disclosure on Comprehension Questionnaire

To gain an additional sense of participants' understanding of the explanation content, we analyzed responses to the two multiple choice questions and the three open-ended questions from the post-scenario comprehension questionnaire. We used a three-point scale to grade the open-ended questions (1 for correct answer, 0.5 for partially correct

answer, 0 for incorrect answer). The average scores for both the *detailed* explanation (mean = 4.27, SD = 0.61) and *summary* explanation (mean = 4.06, SD = 0.79) were high, with no significant main effects of *Information Depth* ($F_{1,30} = 1.664, p = .207$) or *Progressive Disclosure* ($F_{1,30} = 1.484, p = .233$), and no *Information Depth* $\times$ *Progressive Disclosure* interaction effect ($F_{1,30} = 0.798, p = .37$). These questions were designed to be high-level to ensure they were answerable with both explanations, which likely contributed to the lack of significant effects. The questions and sample graded answers can be found in the appendix (Appendix B.9).

4.3.4. Interview Findings

We additionally examined our interview data to explore if they support or contradict our results from participants' critiques and questionnaire responses. Below we present our key findings from the interviews.

Unequivocal Preference for Detailed Explanations

Our interview data suggest a clear preference for the *detailed* explanations, which supports our results from the questionnaire responses (Section 4.3.1) and participants' critique data (Section 0). Participants consistently emphasized the importance of detailed explanations to understand the systems' functionalities and implications.

I would say [I prefer the detailed] one because it gave me more information to make an informed decision if I'm going to purchase that system. But then the [summary] one, I would say, it was really vague. I didn't understand [it] fully, so I think I just felt more confident in the [detailed] one. (P5)

Most found the level of detail provided in the *detailed* explanation to be crucial for forming informed opinions about the AI systems and commented on how the lack of detail in the *summary* explanations led to frustration and uncertainty.

I really felt like with the [summary] one, I didn't have enough context or enough information to really know how the data [was] being used or where it is coming from. So, the lack of information impacted my trust in the system. The [detailed] one, I feel like there was a lot more background information. (P25)

I felt that the [summary] explanation was just missing a lot of information, and I didn't really know what I was getting myself into, and that it was frustrating because you were missing information. (P33)

We saw a statistically significant impact of *Information Depth* on participants' perceived cognitive load (Table 7) and some participants in the interview also acknowledged the potential for information overload with the *detailed* explanation. However, they still preferred having the additional information to make informed decisions.

To be honest, I didn't really have trouble, analyzing the [summary] one. I got confused at some point, but [that was] because there was not enough information. Even though I understood the [detailed one] and would prefer it, it was a bit mentally challenging to analyze [the detailed explanation]. (P14)

Progressive Disclosure of Information can be a Useful Addition to Explanations

Participants, regardless of the presence of *Progressive Disclosure*, appreciated the well-organized interfaces with the mix of text and infographics. The majority of the participants who interacted with the explanation with *Progressive Disclosure* further commented on the interactive nature of the explanations, citing that it allowed them to control what information they accessed. This might help explain why participants reported learning more with *Progressive Disclosure*.

I think [the interactivity] made it really exciting and made it easier for me to focus as opposed to if you give me a table with too much information. If you give user, the control where they can just click and see information that piques their interest, they'll be more invested into that activity. (P2)

When it comes to this [explanation] where you have to interact with the information, and you have to purposely [click] to look for the information, it kind of gives you a feeling of being in control of whatever information I'm seeing. This made me more interested in learning more in terms of the processes and everything and reading more in depth in all the categories. (P28)

Other participants, however, did not find any potential benefit of *Progressive Disclosure*, particularly with the summary explanation.

I felt like clicking sometimes was unnecessary because so little information [was revealed] in one click. (P1)

This indicates that Progressive Disclosure might be valued only when explanations contain substantial information.

4.4. Discussion

Our study revealed that *Information Depth* strongly shapes users' perceptions of training dataset explanations, their critiques of the AI system that provide the explanation, and their cognitive load. *Detailed* explanation significantly increased participants' perceived understanding, trust, and fairness judgments of the AI system, as well as their perceived learning, though they also required more cognitive effort. Participants' critiques also reflected a tradeoff: *detailed* explanations led to more accurate critiques but narrower topic coverage (i.e., with a heavy focus on data collection information), while *summary* explanations prompted shorter critiques with slightly more balanced coverage across the different topics present in the explanation. Participants expressed a consistent preference for *detailed* explanations despite the added effort.

Our results also showed that *Progressive Disclosure* is not sufficient to minimize the extra cognitive load introduced by *detailed* training dataset explanations, however, most participants commented on its utility and felt that they learned more when it was present. While the lack of impact on trust, fairness perception, and perceived understanding can be attributed to the fact that the same information was ultimately available across all conditions, it was surprising to see no differences in cognitive load based on suggestions in prior work [293,348]. One plausible explanation is that the task context imposed only low to moderate cognitive load. Even the detailed explanations yielded relatively modest load (e.g., mean value of 3.5 on a 7-point scale), which might have left limited room for

Progressive Disclosure to provide measurable relief. Additionally, given the absence of critical time pressure, participants may have been able to accommodate additional information without experiencing overload [133]. Progressive Disclosure might be more impactful in contexts characterized by time pressure or higher task complexity. It is also possible that alternative ways of achieving *Progressive Disclosure* might be more impactful, such as starting with a brief *summary* explanation and allowing users to expand it into a *detailed* version. Future research could also investigate other approaches to mitigating cognitive load, including cognitive forcing functions (e.g., mechanisms like checklists to slow down how people process information) [47,86,111], and alternative explanation formats (e.g., visuals, video, hybrid approaches) [302].

Despite the lack of significant results for *Progressive Disclosure*, our results provide a starting point for establishing the role of Progressive Disclosure in explanation design. The fact that *Progressive Disclosure* did not reduce cognitive load, but enhanced participants' perception of learning suggests that its strength might lie less in lowering cognitive burden and more in shaping the user experience, for example by supporting reflection and learning. Progressive Disclosure may be particularly useful in onboarding workflows, where supporting exploration, learning, and reflection is more important than immediate efficiency. Designers should therefore view Progressive Disclosure not as a universal solution for cognitive load management, but as one component in a broader design strategy that balances learning and usability.

Our findings highlight a potential tension between supporting both accurate and comprehensive critiques. Participants provided more critique comments with greater accuracy with the detailed explanation, but tended to focus predominantly on data

collection (the first category in the interface), whereas explanation topic coverage was relatively more balanced with the summary explanation. Therefore, designers may need to consider the goal of the explanation when deciding on how depth to provide. For example, if the goal is to encourage users to consider all aspects of a system, and improving subjective impressions is not important, summary explanations might be most effective. On the other hand, detailed explanation might be better suited for critique accuracy and enhancing subjective impressions. The observed trend that participants produced more critiques with greater accuracy when Progressive Disclosure was absent further highlights the complexity of balancing comprehensiveness with interface mechanisms. Future work should explore explanations that are not only detailed but also incorporate mechanisms that motivate users to engage with different types of information. For example, one approach could be to highlight key pieces of information to draw attention [59,122,333]. Another approach could be to prompt users [296] to consider all relevant information.

The increased trust and judgment of system fairness for *Detailed* explanations over *summary* versions raises important considerations regarding the potential for overreliance [25,247,255,326,340]. The higher perceived trust and judgment of fairness suggests that users felt more confident in the system, however, this confidence is only beneficial for human–AI collaboration if participants have engaged in critical thinking [158]. While the use of critiquing as a task in our study guarded against automatic thinking [158] and we saw positive impacts on critique accuracy with the *detailed* explanations, our findings do not speak to how this increased trust and understanding might ultimately impact reliance. An encouraging prior finding is that participants

showed lower trust with detailed training dataset explanations with “red flags” compared to those that described fewer potential causes for concern [8]. Future studies could explore how this calibration of trust might be influenced by the information depth. Field studies could also investigate how these types of onboarding explanation impact in-situ system use.

Importantly, the observed preference for *detailed* explanations emerged in a context where participants were presented with a reflective, high-stakes critique task. In such settings, users may be more willing to tolerate higher cognitive effort in exchange for a sense of completeness and understanding. This suggests that detailed explanations may be most appropriate during onboarding, auditing, or evaluation phases, when users are motivated to build a foundational mental model of the system. In contrast, users might benefit from *summary* explanations in lower-stakes or time constrained contexts. Designers should therefore treat explanation depth as context dependent choices, informed by users’ goals and needs rather than fixed defaults.

Despite their longer format than many AI explanations, the training dataset explanations as designed do not appear to suffer from the pitfall of high cognitive load. Training dataset explanations might still be susceptible to other explainability pitfalls outlined in Section 4.1, such as lack of actionability [107,197] and misinterpretation [9]. For example, future research could explore whether varying levels of *Information Depth* might lead to differences in the number of actionable insights users derive from the explanations. Future studies should also examine how such explanations can support actions during system onboarding, such as adjusting user expectations [173] and guiding early interactions.

We balanced participants' AI literacy [207,323] and backgrounds across study conditions to minimize potential confounds associated with differing levels of AI knowledge. Similar to Chapter 3, we used a combination of questionnaire items (e.g., educational background [8], formal AI/ML training/experience [8,105], self-reported AI knowledge [105], and an AI literacy scale [323]) to get an understanding of AI knowledge and experiences. Our participant pool nevertheless demonstrated a relatively high mean AI literacy score (mean = 5.6, SD = 0.73). While this suggests that many participants were generally knowledgeable about AI, the result may also reflect a ceiling effect in the AI literacy scale, as documented in prior work [65]. Interpreting AI knowledge was further complicated by the fact that participants with similar literacy scores varied in their educational background, training experiences, and self-reported AI knowledge. Given recent findings that AI knowledge can shape user perceptions of AI explanations [105,276,302], future research should investigate how participants of different levels of AI literacy experience and use explanations with varying depths of information. For example, participants with lower knowledge levels may feel burdened by detailed explanations and prefer summary versions, while those with higher knowledge levels may use detailed explanations to generate more nuanced critiques. Progressive disclosure might also benefit users with lower knowledge levels more, as they are more likely to learn from gradually revealed information. Advancing this line of work will require stronger and more comprehensive measures of AI literacy, which can support systematic investigation of how users of different levels of AI literacy interpret and apply training dataset explanations.

There are additional limitations to our work that warrant further investigation. Owing to feasibility, we used a mixed design in our study with *Information Depth* as a within-subjects factor and *Progressive Disclosure* as a between-subjects factor. This choice may have limited our ability to detect effects, particularly for *Progressive Disclosure*, as between-subjects comparisons typically offer lower statistical power than within-subjects designs. Future work could explore alternative study designs to increase the statistical power to detect differences across levels of *Progressive Disclosure* and to elicit contrastive comments on this factor. Additionally, our sample size was on the smaller side. Future studies with a larger participant pool would strengthen generalizability and support more detailed subgroup analysis, such as by AI literacy or domain background.

4.5. Summary

In this chapter, informed by our scoping review on explanation pitfalls, we examined how the depth of information in training dataset explanations and the use of Progressive Disclosure influence users' perceptions and understanding of an AI system and their cognitive load. In comparison to summary versions, detailed explanations improved users' understanding and subjective impressions of the system, but at the cost of higher cognitive load. Information depth further impacted the balance of topics that participants covered in their critiques of the system. While the use of Progressive Disclosure did not reduce cognitive load, it enhanced perceived learning. These findings highlight important design trade-offs in training dataset explanations by emphasizing the need to balance completeness, clarity, and effort.

Chapter 5

Developing a Psychometrically Validated Scale for Assessing AI Literacy

Chapter 3 and Chapter 4 of this dissertation examined how explanation design shapes user comprehension and use of the explanations and their assessment of AI systems. Specifically, in Chapter 3, we investigated how *Presentation Style* (Data Story vs. Q&A) influenced users' comprehension and usage of training dataset explanations, while in Chapter 4, we explored how *Information Depth* and *Progressive Disclosure* affected users' comprehension, usage of the explanations and their cognitive load. Together, these studies suggest that presentation and depth influence how people interpret and use training dataset explanations.

At the same time, these studies revealed a recurring challenge: users' familiarity, knowledge, and understanding of AI appeared to shape how they interpreted and used training dataset explanations. For example, Chapter 3 showed that some participants with

a stronger AI background (e.g., those from CS/Engineering disciplines or with prior AI/ML experiences) expressed a desire for greater depth in the explanations. It is possible that these participants would be able to extract greater value from *detailed* explanations (Chapter 4), not solely because of design choices but also because their prior AI experience would help them manage the complexity. Yet we found assessing users' AI understanding to be challenging across both studies. While we used a combination of questionnaire items (e.g., educational background [8], formal AI/ML training/experience [8,105], self-reported AI knowledge [105], and an AI literacy scale [323]) to gain a sense of users' AI understanding, these measures yielded inconsistent results. This highlights the need for more robust instruments that can comprehensively capture users' AI understanding.

In this chapter, we address this need by developing a psychometrically validated scale for assessing users' AI understanding. We first explored concepts that could comprehensively capture users' AI understanding. This process led us to adopt AI literacy [207] as our core construct, which encompasses an individual's ability to understand, evaluate, and interact with AI in everyday contexts. Following established scale development guidelines [61,92,143], we developed a dual-format instrument that includes Likert-scale items to assess self-perceived understanding and multiple-choice items to assess demonstrated understanding. Our approach moves beyond prior instruments that are often narrow in scope or tailored to specific contexts (e.g., student performance [229], workplace decision-making [252]), providing a comprehensive understanding of users' AI literacy. Beyond developing and psychometrically validating the scale, we also used it to investigate how users' perceived and demonstrated AI understanding align and discuss

how this relationship can potentially inform the design of AI systems that support improved human–AI collaboration.

We begin the chapter by first introducing the concept of AI literacy and then describing the scale development and validation process. We then analyze and present findings on the relationship between perceived and demonstrated AI understanding and their implications for human–AI interaction design.

5.1. Defining AI Literacy

Given the increasing prevalence of AI systems, researchers across disciplines have emphasized the importance of equipping individuals with the abilities needed to engage critically and responsibly with these technologies. These abilities are often conceptualized under the term *AI literacy* [74,207,228,323]. While definitions of AI literacy vary, there is broad agreement that it is a multidimensional concept encompassing technical, ethical, and social abilities [188]. Some definitions place stronger emphasis on ethical aspects such as fairness, accountability, and safety [142,207,228,323], while others focus more heavily on technical understanding [66,161,222]. However, most definitions converge on the idea that AI literacy is less about the technical development of AI applications and more about the ability to critically evaluate AI technologies, interact with them effectively, and understand their broader societal impacts.

Complementing this body of work is the concept of *AI competence* [60,74], which, while frequently used interchangeably with AI literacy, expands the notion by incorporating users’ confidence and ability to self-reflect on their AI understanding and focusing on how well individuals use AI in beneficial ways [74]. AI competence moves

beyond knowledge and understanding to include the ability to apply AI in task contexts ethically, responsibly, and productively [342]. The term AI ability also appears in this literature, usually referring to an individual's capacity to apply AI-related knowledge and skills in practice. However, AI ability is typically embedded within definitions of AI literacy and competence rather than being treated as a standalone construct.

A prerequisite for interpreting and evaluating how individuals apply AI in practice is having a foundational understanding of what people know about AI and how they reason about its impact. In light of this, we prioritize the latter rather than evaluating individuals' task performance. Accordingly, we adopt AI literacy as our core construct. This allows us to focus specifically on users' conceptual understanding of AI and their ability to reason about its societal and technical implications. We ground our operationalization in the widely cited AI literacy framework by Long and Magerko [207], which conceptualizes AI literacy as a knowledge-oriented construct and provides a well-established foundation for measuring AI understanding in alignment with our research goals.

5.2. Developing a Scale to Assess AI Literacy

We followed a multi-stage, theory-informed process consistent with established scale development guidelines [92,143]. In this section, we describe the process of building our scale, which is summarized in Figure 15.

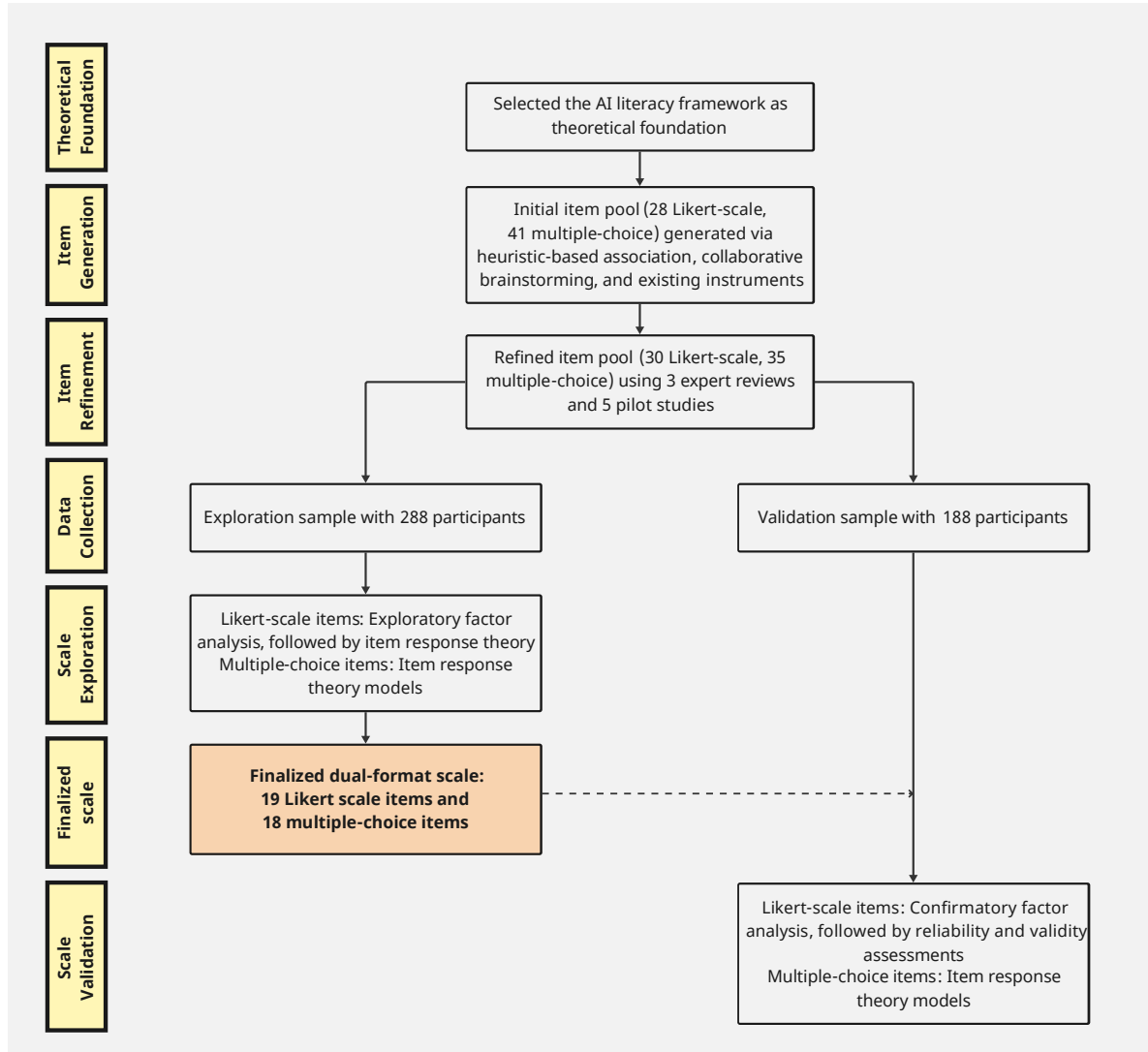


Figure 15: Overview of the process used to develop the dual-format AI literacy scale. The process includes identifying theoretical foundation, item generation, refinement, data collection, scale exploration, finalization, and validation.

5.2.1. Theoretical Foundation

Our scale development was guided by the AI literacy framework proposed by Long and Magerko [207]. This widely cited framework outlines 17 core abilities that help individuals evaluate AI technologies critically, communicate and collaborate with AI, and use AI effectively across online, home, and workplace contexts. These abilities encompass both technical and sociocultural dimensions, including the ability to recognize AI systems,

understand AI’s strengths and limitations, interpret decision-making processes, and reflect on public perceptions and ethical implications. Table 11 provides the list of abilities in the AI literacy framework. We excluded the ability related to “*Imagining Future AI*” from our scale development, as it has been characterized as speculative in prior work [147]. We used the remaining abilities as the basis for the item generation.

Table 11: Abilities in Long and Magerko’s AI literacy framework [207] and brief descriptions used in our operationalization.

Abilities in the AI Literacy Framework	Brief description of the ability
1. Recognize AI	Distinguish between technologies that use AI and those that do not.
2. Understanding Intelligence	Analyze what counts as “intelligence,” including differences across humans, animals, and machines.
3. Interdisciplinarity	Recognize there are multiple approaches to building “intelligent” systems and that AI draws on multiple disciplines.
4. General vs. Narrow	Distinguish between task-specific AI (narrow) and AI with broad, human-like capabilities (general).
5. AI’s Strength and Weaknesses	Identify problem types where AI performs well versus where it struggles, and use that to judge the appropriate use of AI.
6. Imagine Future AI	Envision future AI applications and consider their potential effects on the world
7. Representations	Understand what knowledge representations are and describe examples of how information can be represented for AI.
8. Decision-Making	Describe examples of how computers reason and make decisions.
9. ML Steps	Understand key steps in machine learning and the practices and challenges involved at each step.
10. Human Role in AI	Recognize how humans shape AI systems through programming, model choice, data curation, and fine-tuning.
11. Data Literacy	Understand core data skills like obtaining, interpreting, evaluating, managing, and using data.
12. Learning from Data	Recognize that AI systems often learn from data

13. Critically Interpreting Data	Understand that data needs interpretation and that training data choices influence model outputs and behavior.
14. Action and Reaction	Understand that some AI systems can act in the world, either via planned behavior or reactive responses.
15. Sensors	Understand what sensors are, how they are used by computers to perceive the world, and identify them in a variety of devices
16. Ethics	Identify and describe perspectives on key ethical issues around AI, including privacy, bias, transparency, and accountability
17. Programmability	Understand that AI systems are programmable and can be designed or taught to behave in particular ways.

5.2.2. Item Generation

To generate an initial pool of items, we used a combination of heuristic association (i.e., drawing on intuition and domain knowledge to generate items aligned with the abilities [143,323], a review of existing instruments [60,147,154,187,192,229,252,323], and collaborative brainstorming [39,143]. When appropriate, we adapted items from prior studies by selecting conceptually aligned AI literacy questions and rephrasing them to match the wording, scope, and dimensional structure of the Long and Magerko framework [207], while preserving their original conceptual nature.

We used 7-point Likert-scale items for self-assessment and multiple-choice questions (with one correct option out of four) for factual understanding assessment (also referred to as demonstrated understanding throughout this dissertation). We chose the Likert-scale format for its widespread use in self-efficacy and awareness measurement tools [23,92]. On the other hand, we used multiple-choice questions for factual assessment because of their ability to produce verifiable assessment, independent of self-perception [135]. While alternative formats such as open-ended or fill-in-the-blank questions can

also assess factual understanding, multiple-choice questions offer practical advantages in terms of objective scoring, response standardization, and scalability [253,271].

We ensured that each item explicitly reflected at least one dimension from Long and Magerko's AI literacy framework [207]. All items were reviewed and refined for technical accuracy and clarity by myself and my advisor, both of us having experience with questionnaire design and deployment. The iterative process resulted in an initial pool of 25 Likert-scale questions and 47 multiple-choice items. Although the items were conceptually informed by prior AI literacy scales, only a small subset of items could be directly used or adapted. Of the 25 Likert-scale items, 10 were adapted from existing instruments. Direct reuse was limited for Likert-scale items because most existing scales operationalize AI literacy using different frameworks or employ item wording that does not map cleanly onto the fine-grained dimensions targeted in our framework. The remaining Likert-scale items were newly constructed to provide more precise coverage of the targeted abilities. In contrast, the multiple-choice set included a mix of new items and 28 items adapted from earlier research [147]. The complete set of items, along with indications of which items were adapted and their original sources, is provided in the appendix (Appendix C.1).

5.2.3. Expert Reviews and Pilot Study

We refined the initial item set by conducting expert reviews and cognitive interviews with three domain experts: two additional HCI researchers and one social science researcher. Our expert sample size is consistent with recommendations for early-stage item development, where the focus is on refining clarity and identifying potential misunderstandings [143]. Following best practices, we used the think-aloud method [331]

during these interviews to identify comprehension issues. Based on expert feedback, we iteratively revised the items by replacing complex terms with simpler alternatives, removing unclear or redundant items, and splitting items that covered multiple constructs. This led to a refined item pool consisting of 30 Likert-scale items and 35 multiple-choice items.

We then conducted a pilot study with five participants to explore the usability of the questionnaire. The pilot aimed to estimate completion time, identify any remaining comprehension issues, and detect any formatting or grammatical errors. The average completion time was approximately 20 minutes. Participants generally found the items understandable, but we made minor refinements to improve clarity and consistency based on their feedback. All finalized items, along with their mapping to the primary dimensions of the AI literacy framework [207], are provided in the appendix (Appendix C.2).

5.2.4. Method

We conducted an online survey to investigate the structure and psychometric quality of the items. Below, we describe the participants and procedures used to evaluate the scale's structure and psychometric quality.

Participants

We recruited 300 participants via the Prolific platform [243]. Our sample size was guided by scale development literature [92,238], which suggests that 300 participants is adequate for scale development. We did not require participants to have prior experience with AI or Computer Science with the aim of including individuals with a broad range of AI

familiarity. We implemented attention checks to mitigate low-quality responses and excluded participants who failed these checks or left more than 10% of items unanswered. Based on these criteria, we removed 12 participants from the sample, resulting in a final sample of 288 participants (144 men, 143 women, 1 non-binary). Figure 16 demonstrates a breakdown of participants' ages, educational levels, and prior AI training, and we provide the full set of demographic questions and the details in the appendix (Appendix C.3 and C.4).

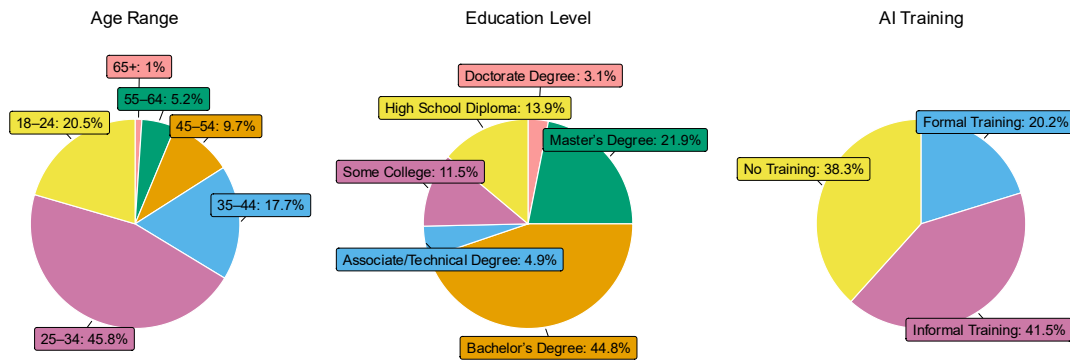


Figure 16: Demographic distribution of survey participants by age, education level, and AI training. The majority of participants were aged 25–34 and held a bachelor's degree, with AI training most commonly being informal or none.

Procedure

Participants first completed a brief demographics questionnaire, followed by 30 Likert-scale items and 35 multiple-choice items. To verify data quality, we included three attention-check items to each format (e.g., “please select X” for Likert-scale item, “how many sides does a triangle have?” for multiple-choice). We randomized the response options for all multiple-choice items. Participants took 17 minutes on average to complete the study and received six dollars in compensation. The study was approved by

University of Manitoba's Research Ethics Board. The approval and the study materials can be found in the appendix (Appendix C.5 and C.6).

5.2.5. Data Analysis

We conducted a multi-stage analysis to evaluate and improve the psychometric properties of the scale and to ensure that the scale reliably captures key dimensions of AI literacy. Specifically, we aimed to uncover the underlying factor structure, reduce redundancy, and retain items with strong measurement performance. To this end, we employed a combination of Exploratory Factor Analysis (EFA) [116] and Item Response Theory (IRT) [22]. We used EFA to uncover the underlying factor structure of the items and to guide the removal of poorly performing items. We subsequently applied IRT to evaluate measurement precision and discrimination power of the items across varying levels of AI literacy.

Self-Assessed Likert Items

We first examined the suitability of the Likert-scale data for factor analysis using the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy [160] and Bartlett's Test of Sphericity [28]. We applied EFA with Principal Axis Factoring given the ordinal nature of the data and the non-assumption of normal distributions [116]. We used oblique rotation due to the expected correlation between latent factors [92,116]. We determined the number of factors based on multiple criteria, including Kaiser criterion (eigenvalues > 1) [159], scree test [64], parallel analysis [139], and the Velicer's Minimum Average Partial (MAP) test [318]. Using multiple criteria provides a more reliable basis than relying on any single heuristic [82,92,116]. We considered items with low factor loadings (< 0.5), high cross loadings (> 0.3), or low communalities (< 0.2) for removal to improve

the clarity and interpretability of the scale. Factor loading represents how strongly an item relates to a factor, whereas cross loading indicates overlap with multiple factors, and communality captures the variance explained by the factor. Following EFA, we modeled items using the Graded Response Model (GRM) of IRT [274], which is suitable for Likert-scale data. IRT provides item-level insights into discrimination, and how response options functioned across varying levels of AI literacy. We used a discrimination threshold of 0.65 based on established psychometric guidelines to remove items [22]. We also used visual inspection of Item Characteristic Curves (ICC) to evaluate item performance [274]. ICCs graphically represent how the probability of responding to an item changes across levels of the underlying trait being measured. By examining the shape and position of these curves, we can assess how well items distinguish between participants and whether they function as expected. Finally, we assessed internal consistency of each factor using Cronbach's alpha [85], which estimates the degree to which items within a factor are correlated and thus measure the same construct. We considered 0.70 as the acceptable threshold.

Multiple-Choice Items

We scored multiple-choice items dichotomously (correct = 1, incorrect = 0). To examine the dimensional structure of the multiple-choice items, we computed tetrachoric correlations, which are appropriate for binary data and provide a more accurate estimation of item relationships than Pearson correlations [44]. We then examined the factor structure of the multiple-choice items using the same criteria as for the Likert-scale items. Given the absence of a clear factor structure from the analysis, we tested unidimensionality through Confirmatory Factor Analysis [44,262,277]. Following

confirmation of unidimensionality, we applied IRT using the 3-parameter logistic (3PL) model, which estimates item discrimination, difficulty, and guessing parameters [112,263]. We chose the 3PL model to account for guessing, which is especially relevant for multiple-choice items. *Discrimination* indicates how effectively an item can differentiate between ability levels, *difficulty* refers to the ability level needed for a correct response, and *guessing* estimates the probability that a low-ability participant answers correctly by chance. We considered items with poor model fit (e.g., RMSEA > 0.05, $S-\chi^2 p < 0.05$), low discrimination (< 0.35), or high guessing (> 0.4) for exclusion. We fit our IRT models using the *mirt* package in R [67].

5.3. Results

5.3.1. Likert-Scale Items

Our analysis indicates strong underlying structure and good psychometric quality of the Likert-scale items. The KMO value was 0.95, indicating excellent sampling adequacy, and Bartlett's test of sphericity was significant, $\chi^2 (435) = 7338.01$, $p < .001$, confirming that the data was suitable for factor analysis.

Using a combination of factor retention techniques (Kaiser criterion, scree plot, parallel analysis, and Velicer's MAP test), we identified a two-factor solution. We present the results of the parallel analysis in Figure 17, where the first two eigenvalues from the actual data exceeded those from the simulated data, supporting a two-factor solution. This suggests that AI literacy, as reflected in participants' responses, consist of multiple aspects.

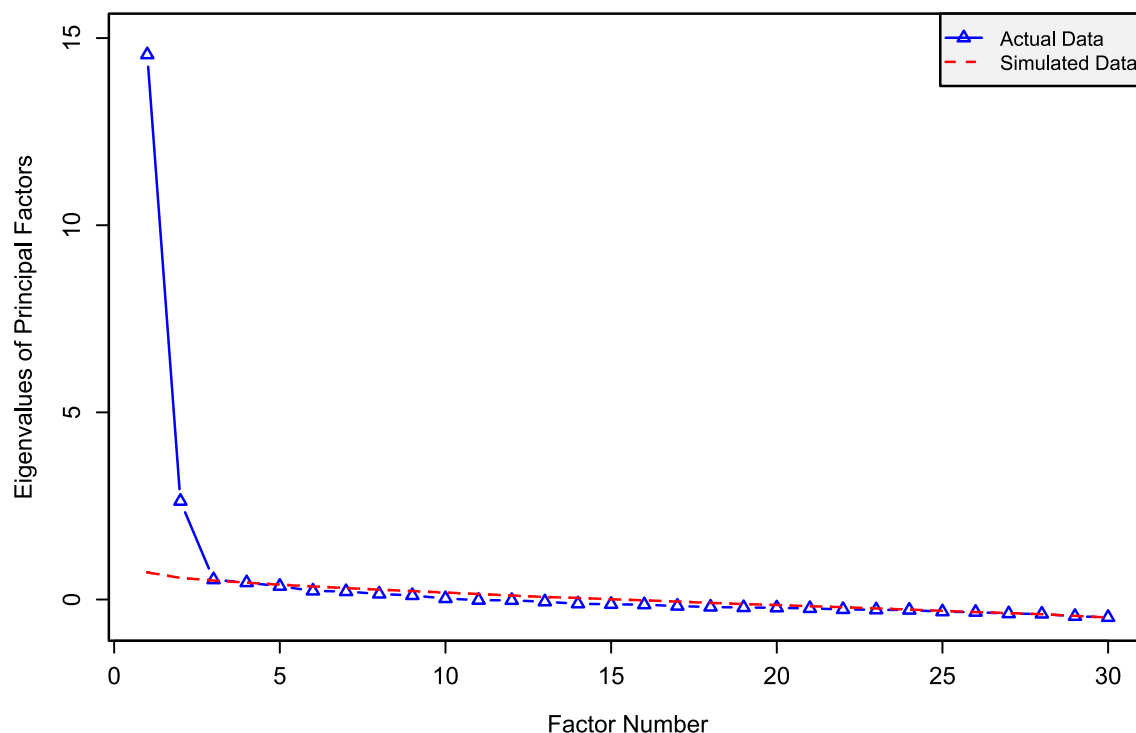


Figure 17: Parallel analysis to determine the number of factors for the Likert-scale items. The solid blue line represents the eigenvalues derived from the actual data, while the red dashed lines represent eigenvalues from simulated data. Factors are retained when the actual eigenvalues are larger than the simulated ones. Here, the first two eigenvalues meet this criterion, supporting the retention of a two-factor solution.

Following this, we conducted a two-factor EFA and explored the loadings for each item for both factors. Based on our predefined removal criteria (Section 5.2.5), we retained 24 out of original 30 items. Thirteen items loaded on the first factor and eleven loaded on the second factor. While the six items we removed did not align clearly with either factor, this does not necessarily mean the items were invalid, rather they contributed less to the emerging structure than other items and may require further refinement.

To reduce redundancy, we examined high inter-item correlations, lower item-total correlations, and conceptual overlap among items. We identified two pairs of items within Factor 1 and one pair within Factor 2 that exhibited high inter-item correlations ($\rho > .80$), suggesting redundancy. Additionally, one item in Factor 2 showed a low item-total correlation (< 0.6) and conceptual similarity with another item in the same factor. Based on these findings, we removed one item from each highly correlated pair and the low-performing item in Factor 2. This refinement resulted in a final structure of 11 items in Factor 1 and nine items in Factor 2.

We then applied IRT separately on each factor to further assess item quality and discrimination. As described in Section 5.3.5, we used both a discrimination threshold of 0.65 and visual inspection of ICCs to evaluate item performance. Figure 18 and Figure 19 respectively demonstrates the ICCs for all the items in Factor 1 and Factor 2. No items in Factor 1 were flagged during this phase, suggesting stable items within that factor. However, while all items met the discrimination threshold, we removed one item from Factor 2 (Item 8), as its ICC displayed overlapping curves, indicating that the item is a poor differentiator across levels of AI literacy.

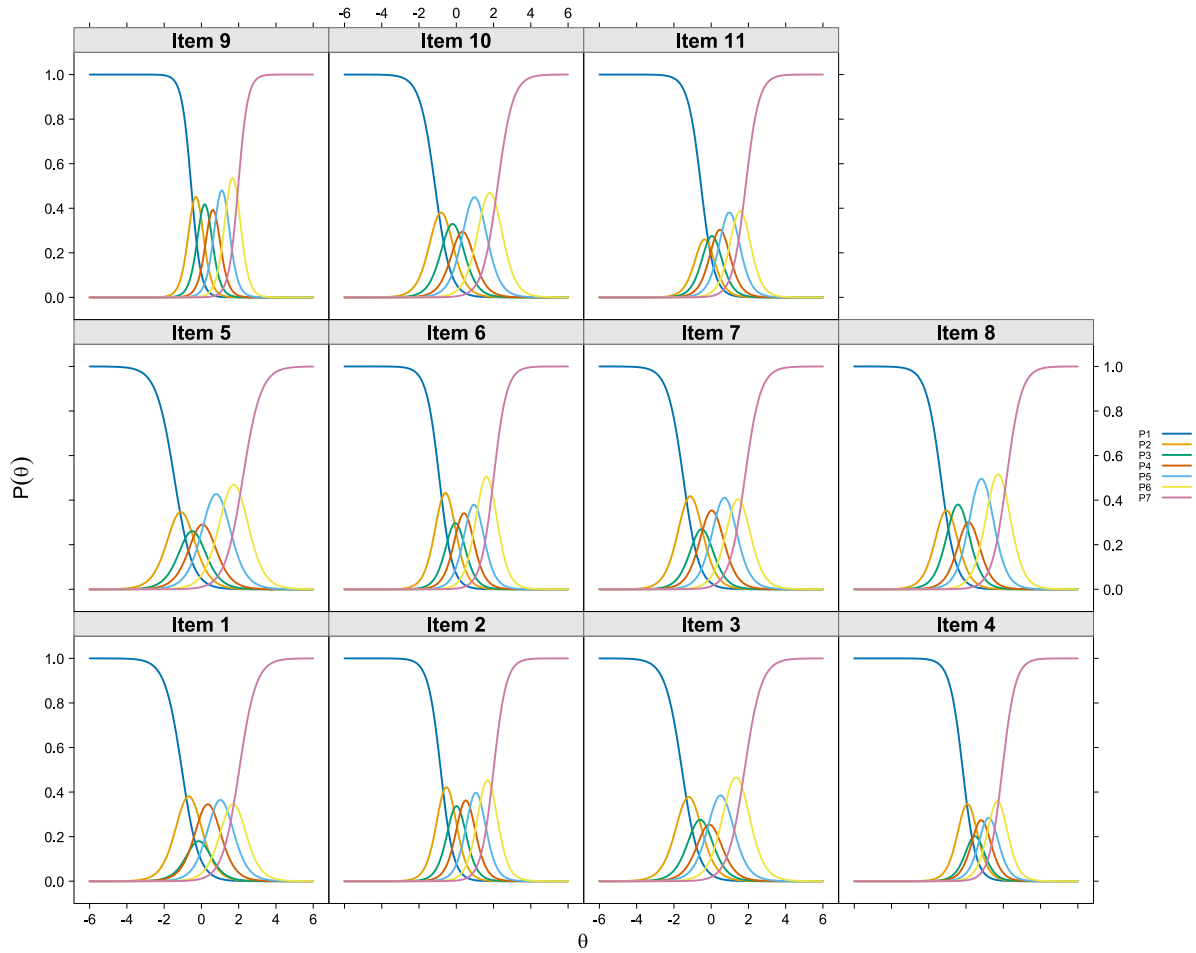


Figure 18: Item Characteristic Curves (ICCs) for Likert-scale items within Factor 1, estimated using a graded response IRT model. The curves illustrate the probability of selecting each response category (P1–P7) across different levels of AI literacy (θ). All items show clear and well-separated curves, indicating strong discrimination.

The final model consisted of two well-defined factors. Based on qualitative interpretation of item content, we labeled Factor 1 as “Functional Knowledge of AI Systems”, which consists of items reflecting an understanding of foundational AI concepts, including how systems are programmed, trained, and how they make decisions. Factor 2 was labeled as “AI in Society”, consisting of items about ethical considerations, societal impacts, and human–AI interaction. We present the retained items organized by factor in Table 12. Both factors demonstrated strong internal consistency, with

Cronbach's alpha greater than 0.92. Item-total correlations ranged from 0.71 to 0.85 for Factor 1 and 0.60 to 0.75 for Factor 2, indicating good item performance. The correlation between the two factors was 0.59, suggesting moderate overlap. More detailed analysis results for the Likert-scale items are available in the appendix (Appendix C.7).

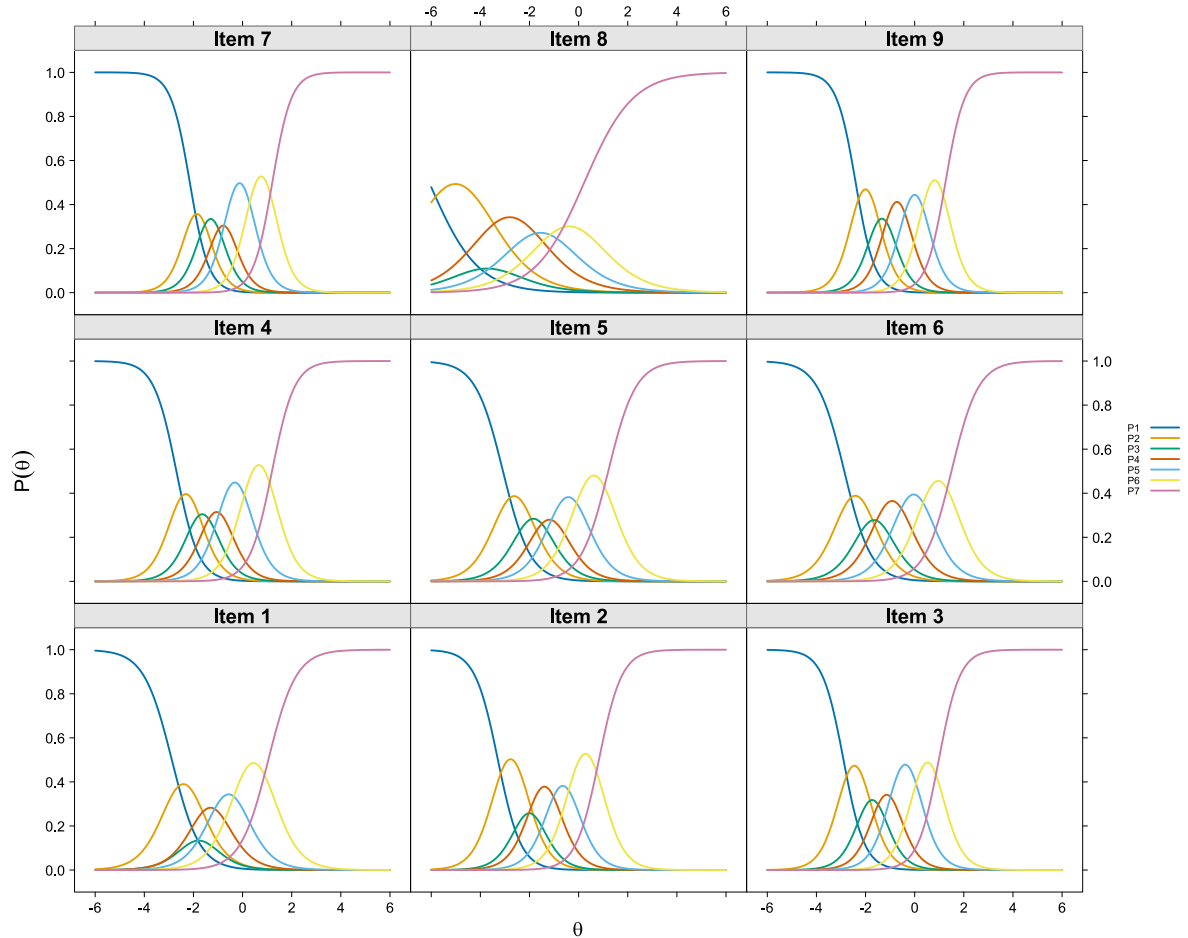


Figure 19: Item Characteristic Curves (ICCs) for Likert-scale items within Factor 2, estimated using a graded response IRT model. The curves illustrate the probability of selecting each response category (P1–P7) across different levels of AI literacy (θ). Most items show clear and well-separated curves, indicating strong discrimination. However, Item 8 displayed overlapping and flat curves, suggesting poor discrimination.

Table 12: Retained Likert-scale items grouped by factor. Items in Factor 1 (“Functional Knowledge of AI Systems”) reflect participants’ self-assessed understanding of how AI systems are built and function, including programming, training, data analysis, and machine learning processes. Items in Factor 2 (“AI in Society”) capture awareness of ethical challenges, regulatory concerns, and the broader societal implications of AI. All items are mapped to one of the dimensions in the AI literacy framework [207].

Retained Likert-Scale Items	Represented Ability from AI literacy Framework
Factor 1: Functional Knowledge of AI Systems	
I can differentiate between narrow AI and general AI.	General vs. Narrow AI
I understand the process of programming AI applications.	Programmability
I understand the process of training AI applications using data.	ML Steps
I am capable of programming a basic AI application.	Programmability
I can clearly articulate the strengths of current AI systems.	AI’s Strength and Weakness
I have a clear understanding of neural networks as a model in AI.	Representations
I can explain how AI systems make decisions based on learned patterns.	Decision-making
I am confident in my ability to analyze data used in AI training.	Data Literacy
I can thoroughly explain the machine learning process and its components.	ML Steps
I am capable of critically evaluating training data for inaccuracies.	Critically Interpreting Data
I understand the role of sensors in AI systems.	Sensors
Factor 2: AI in Society	
I am aware that AI decisions can reflect biases present in their training data.	Critically Interpreting Data
I recognize that the capabilities and limitations of AI systems are shaped by what they learn from data.	Learning from Data
I am aware that AI systems can effectively respond to human inputs in real-time.	Action and Reaction

I recognize the importance of collaboration across disciplines is necessary for improving human–AI interaction.	Interdisciplinarity
I am aware of the ethical challenges associated with AI development.	Ethics
I recognize that existing regulations play a role in addressing ethical concerns in AI development.	Ethics
I understand how AI systems adapt based on inputs from human.	Action and Reaction
I can describe the role of humans in influencing AI systems.	Human Role in AI

5.3.2. Multiple-Choice Items

Our preliminary analysis yielded mixed results regarding the suitability of the multiple-choice items data for factor analysis. While Bartlett’s test of sphericity was significant ($\chi^2(561) = 28718.75, p < 0.001$), the KMO measure was low (0.26), suggesting weak shared variance across items. The low KMO indicates that the data are not well-suited for identifying latent structures, a common limitation of dichotomously scored multiple-choice items due to reduced variance and sensitivity to item difficulty [89,119,262]. Our EFA confirmed the instability as parallel analysis suggested an overly complex 14-factor structure, and Velicer’s MAP indicated a two-factor solution, but both resulted in weak, uninterpretable structures. Given these results, we concluded that EFA was inappropriate for modeling the multiple-choice data and proceeded with IRT for item-level evaluation.

Before fitting the IRT model, we tested for unidimensionality of the multiple-choice data by fitting an unidimensional model using Confirmatory Factor Analysis with the package *lavaan* for R [272]. The one-factor model demonstrated good overall fit, $\chi^2(560) = 647.23, p = .006, CFI = .98, TLI = .98, RMSEA = .02, SRMR = .10$. While the SRMR value (.10) slightly exceeded the conventional cutoff of .08 [44], all other indices suggest strong

model fit. Thus, we deemed the model acceptable and supportive of an unidimensional structure.

We then fit a 3PL-IRT model to the initial 35 candidate multiple-choice items. Figure 20 displays the ICCs for the full set of multiple-choice items prior to item reduction, depicting how each item differentiates participants across levels of AI literacy (θ). These curves were used to evaluate item discrimination and guessing behavior. Each curve shows the probability of a correct response ($P(\theta)$) across different levels of participant ability (θ). Items with steeper curves exhibit greater discrimination, while curves that begin above zero on the y-axis indicate higher chances of guessing the correct answer.

We removed 13 items that had a low discrimination power and high guessing parameter (as defined in Section 5.3.5), as they offered limited value in distinguishing participants. We further removed 4 items based on poor model fit. This resulted in a final set of 18 multiple-choice items retained for the scale.

The retained items demonstrate strong psychometric properties with discrimination values ranging from 1.09 to 5.39, difficulty from -1.63 to 1.56 and guessing from 0 to 0.39. These values fall within acceptable ranges and cover a good spread [22], suggesting that the remaining items have the potential to effectively differentiate participants across a range of AI literacy.

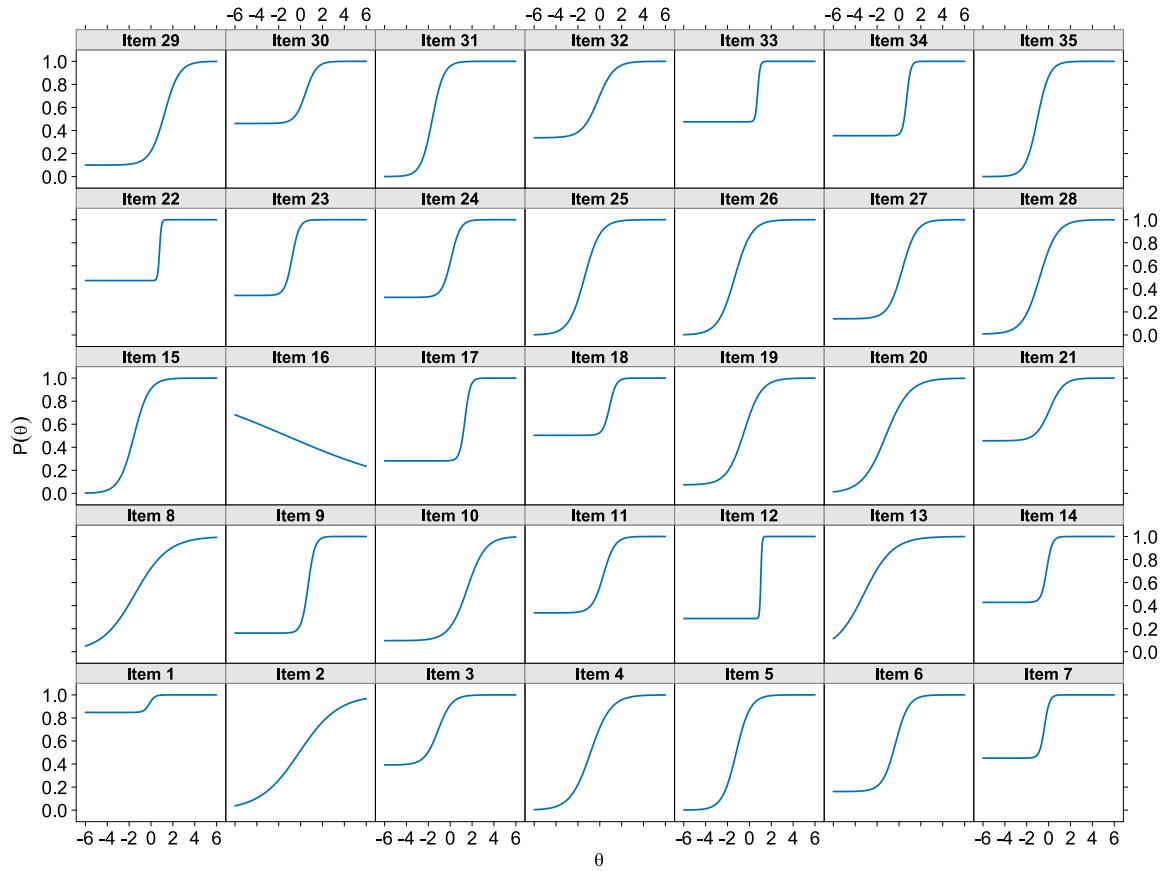


Figure 20: Item Characteristic Curves (ICCs) for the initial pool of 35 multiple-choice items prior to item reduction. These visualizations were used to assess item quality based on discrimination, difficulty, and guessing parameters.

Figure 21 presents the Test Information Function (TIF) for the multiple-choice items. The TIF illustrates the precision of the test across different levels of participant ability. The TIF peaked between $\theta \approx -1.5$ and 1.5 , indicating that the test is most precise at assessing individuals with below-average to slightly above-average ability levels of AI knowledge (with $\theta = 0$ representing average ability). Outside this central range, the amount of information declines sharply, especially for higher ability levels ($\theta > 2$), indicating reduced sensitivity for participants with very high AI knowledge. Similarly, precision diminishes for those with extremely low ability levels. These findings imply that while the scale is well-targeted for the general population, it may benefit from the

addition of higher-difficulty items to better capture differences in expert-level understanding.

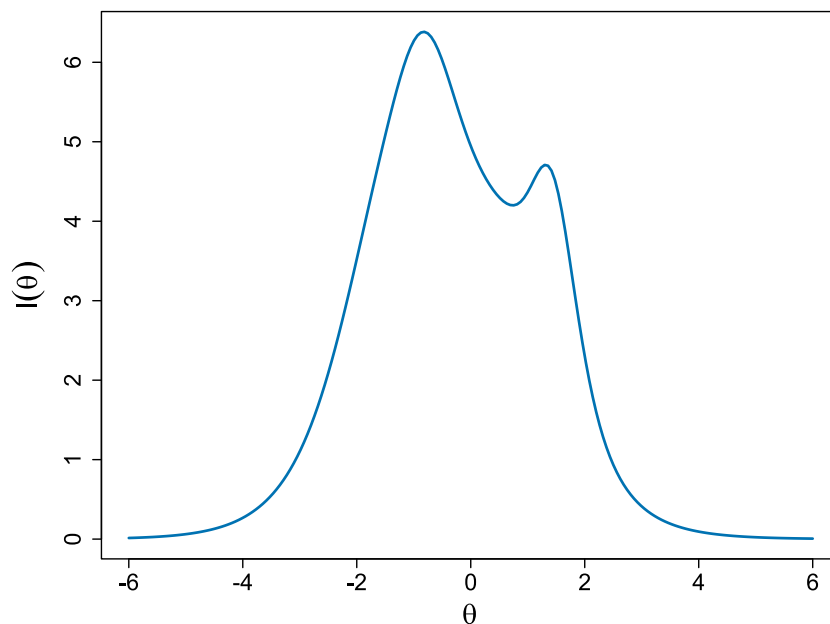


Figure 21: Test Information Function (TIF) plot for the multiple-choice item set with 18 items. It depicts the amount of information provided by a test across different levels of the latent trait (θ).

In summary, while a coherent factor structure could not be established through EFA, item-level analysis via IRT provided a robust alternative for refining the multiple-choice items of the scale. Internal consistency for the final set of multiple-choice items was acceptable (Cronbach's alpha = .80), supporting its reliability in measuring participants' factual knowledge of AI concepts. The final set of multiple-choice items are provided in Table 13 and more detailed analysis results for the multiple-choice items are available in the appendix (Appendix C.8).

Table 13: Final set of 18 multiple-choice items included in the AI Literacy Scale. Each item consists of a stem and four response options, with a single correct answer indicated in italics.

All items are mapped to one of the dimensions in the AI literacy framework [207].

Retained Multiple-choice Items	Represented Ability from AI literacy Framework
1. What is one of the primary characteristics of intelligence? a) Following preset commands; b) <i>The ability to learn and adapt</i> ; c) Memorizing large amounts of information; d) Moving objects with precision	Understanding Intelligence
2. What convinces us for an AI to be intelligent? a) The ability to walk and talk; b) The presence of an artificial brain; c) Being at least as intelligent as humans; d) <i>Acting rationally to achieve specific goals effectively</i>	Understanding Intelligence
3. Which statement best describes narrow AI? a) A system capable of solving diverse problems across multiple domains; b) <i>A system designed to address a single, specific problem</i> ; c) A speculative future AI that mimics human thought processes; d) An AI capable of performing any task as effectively as humans	General vs. Narrow AI
4. What is something narrow AI is unable to do? a) make decisions under uncertain conditions; b) <i>perform a wide range of tasks across different domains</i> ; c) solve a task with a greater efficiency than a human; d) learn from unstructured data	General vs. Narrow AI
5. In which task is AI already outperforming humans? a) <i>Accurately detecting tumors in medical imaging</i> ; b) Programming complex software; c) Translating lengthy novels; d) Designing innovative cancer treatments	AI's Strength and Weakness
6. In which domain do humans still outperform AI? a) Predicting extreme weather events from data; b) <i>Proving mathematical theorems</i> ; c) Answering factual quiz questions; d) Playing poker	AI's Strength and Weakness
7. Which data structure is most commonly used in AI systems? a) Flowcharts that map out decision-making processes; b) <i>Neural networks that simulate interconnected brain-like structures</i> ; c) Pie charts for visualizing datasets; d) Tables that organize raw data in rows and columns	Representations

8. How does AI decision-making differ from human decision-making? a) AI is entirely logic-based; b) <i>AI relies on learned patterns and lacks emotional influence in decision-making</i>; c) AI is capable of creativity and intuition; d) AI always makes faster decisions than humans	Decision-Making
9. What distinguishes supervised learning from unsupervised learning? a) <i>In supervised learning, the output values of the training data are known</i>; b) In supervised learning, humans continuously supervise the training process; c) In supervised learning, all steps are fully documented; d) Supervised learning is subject to stricter legal regulations	Decision-Making
10. At which stage can humans influence the outcomes of machine learning? a) By calculating prediction accuracy; b) By randomly splitting data into training and testing sets; c) <i>By selecting the model used in the learning process</i>; d) By abstracting the structure of the model	Human Role in AI
11. What is essential for training a machine learning model? a) <i>Large amounts of diverse data</i>; b) High-speed internet; c) Simple handcrafted rules; d) Constant human supervision	ML Steps
12. What is the primary benefit of using data visualizations? a) Storing data effectively; b) Preparing images for recognition tasks; c) <i>Conveying insights and clearly communicating results</i>; d) Performing statistical tests	Data Literacy
13. Why do machine learning systems produce strong results? a) They are extremely powerful in computation; b) They think like humans, but faster; c) <i>They analyze large datasets to identify patterns and improve models</i>; d) They are based on expert systems with stored knowledge	Learning from Data
14. You observe that a machine learning model is better at recognizing cats than dogs. What could explain this difference? a) There are fewer online images of dogs compared to cats, making them harder to identify; b) Smaller animals like cats are inherently easier for models to recognize than larger ones like dogs; c) Most machine learning models are naturally better at recognizing cats; d) <i>The training data for dogs was not diverse or representative of all dog breeds</i>	Critically Interpreting Data
15. Which statement accurately describes the machine learning process? a) It is grounded in behaviorist learning theories; b) <i>It is an iterative process that involves repeated refinement</i>; c) Supervised and unsupervised learning follow identical steps; d) The process can be reversed to generate artificial datasets	ML Steps

16. What should be considered when splitting data into training and test sets? a) Ensure the datasets are equal in size; b) <i>Randomly divide the data into training and test sets</i> ; c) Use higher-quality data for the test set than the training set; d) Ensure training and test data are as distinct as possible	ML Steps
17. Which is a common ethical concern related to AI? a) The speed of AI systems; b) <i>Bias and fairness in AI decision-making</i> ; c) AI's limitations in performing calculations; d) The programming languages used to develop AI	Ethics
18. Which societal challenge is frequently associated with AI? a) Lack of investment in education systems; b) Shortages of computer chips due to AI's computational demands; c) High error rates in AI-powered manufacturing processes; d) <i>Displacement of human workers by AI systems</i>	Ethics

5.4. Scale Validation

We conducted a psychometric validation of our finalized AI literacy scale using a second independent sample. This validation method establishes the scale as a dependable and well-structured measure [92]. We evaluated both the self-assessed (Likert-scale) and factual (multiple-choice) components, examining their validity, reliability, and psychometric quality to ensure the robustness of the scale.

5.4.1. Participants

We recruited 200 participants via the Prolific platform [243]. Our criteria for recruitment and data filtering were the same as in the first study (Section 5.2.4), with the additional requirement that participants could not have taken part in the first study. We excluded 12 participants in the filtering process, resulting in a final sample of 188 participants (95 men, 93 women). Figure 22 provides a breakdown of the sample according to age,

education level and prior AI training. Full demographic details are provided in the appendix (Appendix C.4).

The demographic composition of this sample was comparable to the first sample. Sample 2 had a lower proportion of participants with no AI experience (23% vs. 38%) and a slightly higher proportion of participants with a computer science background (31% vs. 23%). These differences suggest a modest shift toward greater AI familiarity but likely reflect natural variation in Prolific recruitment rather than major deviations in overall diversity.

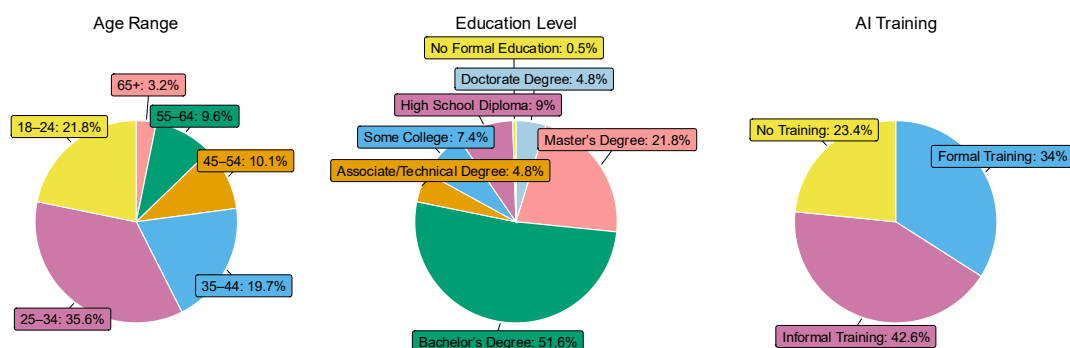


Figure 22: Demographic distribution of the validation sample by age range, education level, and prior AI training. Most participants were aged 25–34, commonly held a bachelor's degree, with mostly informal training. The distribution is closely comparable to the first sample.

5.4.2. Likert Items Validation

Confirmatory Factor Analysis

To evaluate the two-factor structure of the finalized self-assessment Likert-scale, we conducted a Confirmatory Factor Analysis (CFA) on the second sample (with 188 participants) using the *lavaan* package in *R* [272]. CFA is a statistical technique used to test whether the data fit a specified factor structure [44]. The CFA results indicated an

acceptable model fit, with the following indices: CFI = 0.99, TLI = 0.99, RMSEA = 0.088, SRMR = 0.072, and $\chi^2/df = 2.43$. While the RMSEA value exceeded the ideal range (0.05-0.08), it fell within the marginal fit range (0.08-0.10), suggesting a modest approximation error but still within acceptable bounds for practical use [150]. All other indices fell within commonly accepted thresholds for good model fit, collectively supporting the two-factor solution. Furthermore, all standardized factor loadings were above 0.60, indicating that each item was strongly associated with its intended latent construct. Overall, these results provide empirical support for the structural validity of the scale, confirming that the two-factor structure is a valid representation of self-assessed AI understanding.

Assessment of Scale Reliability and Validity

The overall scale demonstrated excellent reliability (Cronbach's alpha = 0.97). Within the two identified factors, the Cronbach's alpha values were 0.96 and 0.90, respectively, indicating strong internal consistency for both factors.

We assessed convergent validity using Composite Reliability (CR) [121] and Average Variance Extracted (AVE) [134]. CR reflects the internal consistency of the items within a factor, functioning similarly to Cronbach's alpha but accounting for factor loadings (which makes it a less biased estimate of reliability than Cronbach's alpha). On the other hand, AVE measures the proportion of variance captured by a factor relative to the variance due to measurement error, indicating how well the items represent the latent factor. Based on established thresholds in the literature (CR $\geq$ 0.70; AVE $\geq$ 0.50) [134], our results supported convergent validity. The CR values for the two factors were 0.96 and 0.89, and the corresponding AVE values were 0.72 and 0.57, all exceeding the

recommended cutoffs. These results suggest that the items within each factor are internally consistent and meaningfully represent the intended factor.

Finally, to evaluate discriminant validity, we used both the Heterotrait–Monotrait (HTMT) ratio of correlations [141] and Fornell–Larcker criterion [121]. HTMT is a widely recommended method for assessing whether factors that are supposed to be different are empirically distinct. HTMT estimates the ratio of between-factor item correlations to within-factor item correlations. A commonly accepted threshold for HTMT is 0.85 [171], with lower values indicating better discriminant validity. In our analysis, the HTMT ratio between the two factors was 0.71, providing evidence of good discriminant validity. We also applied the Fornell–Larcker criterion which evaluates whether each latent factor is empirically distinct from others by comparing the square root of each factor’s AVE to its correlations with other factors. The square roots of AVE were 0.848 for Factor 1 and 0.755 for Factor 2, both of which slightly exceeded the inter-factor correlation of 0.75. This satisfies the Fornell–Larcker criterion, providing additional evidence for discriminant validity. Together, these results suggested that the two factors represent conceptually distinct dimensions of AI literacy, and that participants’ responses to the items reflected this distinction.

5.4.3. Multiple-Choice Items Validation

We evaluated the performance of the 18 retained multiple-choice items using 3-PL IRT [22,89]. Figure 23 presents the ICCs for the multiple-choice items on the validation sample. Overall, the items demonstrated robust psychometric properties, with discrimination values ranging from 0.92 to 15.85, difficulty values from -2.54 to 0.72 , and guessing parameters from 0 to 0.57.

Most items continued to perform well, with some showing improved discrimination. However, a subset of items (i.e., Item 7, Item 10, and Item 11 in Table 13) showed elevated guessing parameters (above 0.40). This suggests that some responses may have been influenced by test-taking strategies (e.g., eliminating unlikely options or selecting the most technical-sounding choice) or item-level cues (e.g., patterns in phrasing or option length that inadvertently hint at the correct answer) rather than actual knowledge. These items may benefit from further refinement to reduce ambiguity and minimize the potential for guessing.

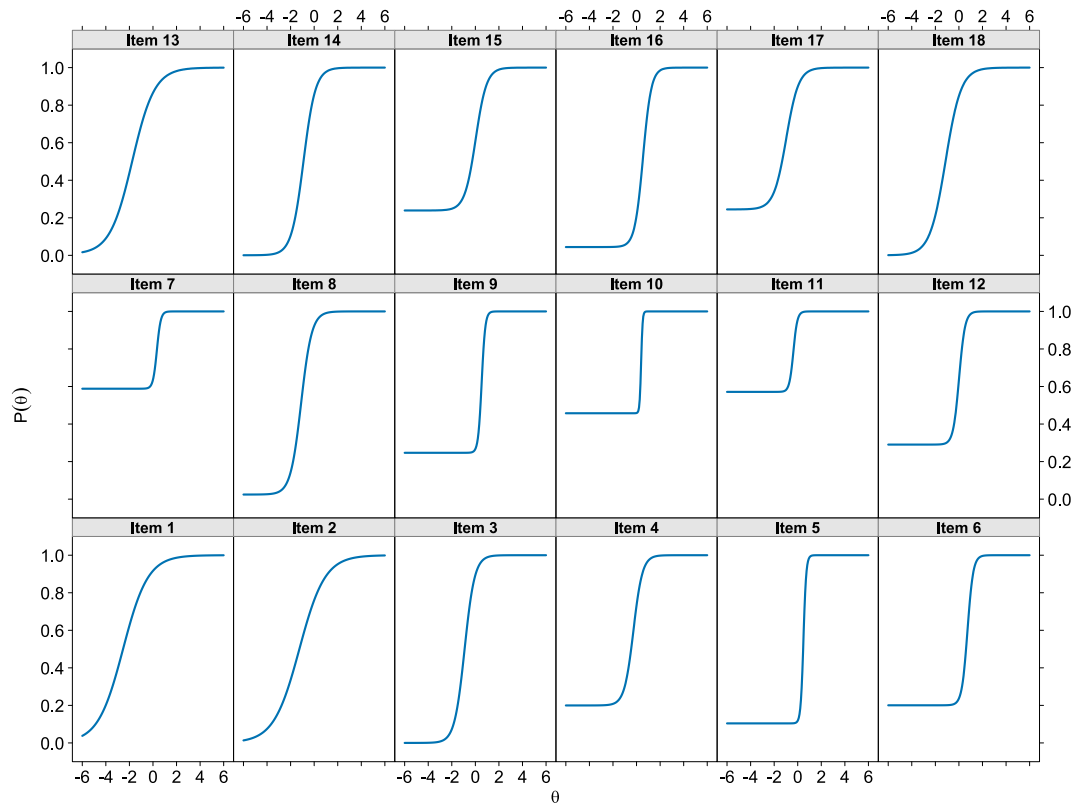


Figure 23: Item Characteristic Curves (ICCs) for the final 18 multiple-choice items. Each curve shows the probability of a correct response ($P(\theta)$) across different levels of participant ability (θ). Items with steeper curves exhibit greater discrimination, while curves that begin above zero on the y-axis indicate higher chances of guessing the correct answer. These visualizations were used to assess item quality based on discrimination, difficulty, and guessing parameters.

The overall scale demonstrated acceptable reliability and construct coverage. Internal consistency for the final multiple-items set improved in the validation sample ($\alpha = .85$), supporting the robustness of the scale. While some items showed increased guessing, the majority performed well across a range of ability levels, and the full item set provides a reasonably accurate and scalable measure of factual AI knowledge. We provided a summary of the scale development and validation results in Table 14.

Table 14: Summary of scale development and validation analyses.

Sample	Analysis	Likert-Scale (Self-Assessed)	Multiple-Choice (Factual)
Development Sample	Factorability	KMO = 0.95; Bartlett's χ^2 (435) = 7338.01, $p < .001$ (excellent factorability)	KMO = 0.26; Bartlett's χ^2 (561) = 28718.75, $p < .001$ (poor factorability)
	EFA	Two-factor solution; 24 of 30 items retained after EFA; further redundancy checks reduced the set to 20 items (11 in Factor 1 and 9 in Factor 2)	Unstable factor solutions; abandoned EFA
	IRT	Applied separately by factor; 1 item removed in Factor 2 (poor ICC)	18 out of 35 items retained, Discrimination: [1.09, 5.39], Difficulty: [-1.63, 1.56], Guessing: [0, 0.39].
Validation Sample	CFA	CFI = 0.99; TLI = 0.99; RMSEA = 0.088; SRMR = 0.072	N/A
	Reliability	Cronbach $\alpha = .96$ (Factor 1), .90 (Factor 2) CR: 0.96 (Factor 1), 0.89 (Factor 2) AVE: 0.72 (Factor 1), 0.57 (Factor 2)	Cronbach $\alpha = .85$
	Validity	HTMT = 0.71; Fornell-Larcker criterion satisfied	N/A
	IRT	All items showed good discrimination	Discrimination: [0.92, 15.85], Difficulty: [-2.54, 0.72], Guessing: [0, 0.57]

5.4.4. Measurement Consistency Across Subgroups

To enable subsequent analyses comparing AI literacy across gender and prior AI training subgroups, we evaluated measurement invariance [219] for the two-factor Likert structure factors and differential item functioning (DIF) [211] for the multiple-choice items. Establishing invariance ensures that observed group differences reflect genuine variation in AI literacy rather than measurement differences such as different groups interpreting certain items differently.

We evaluated model fit for the Likert-factors using the Comparative Fit Index (CFI) [73], where a change in this index (ΔCFI) of less than .01 is typically taken as evidence that measurement invariance holds. For the multiple-choice items, we tested DIF using likelihood ratio tests within the 3PL-IRT model [306] with Benjamini–Hochberg correction [34] for multiple comparisons, considering items to show DIF if flagged at $p < .05$ after adjustment.

For Likert-scale factors, measurement invariance across gender was fully supported. Configural, metric, and scalar models all met the criteria ($\Delta\text{CFI} < .01$), indicating that the scale worked the same way for men and women.

Across prior AI training groups, configural and metric invariance held ($\Delta\text{CFI} < .01$), but scalar invariance initially failed ($\Delta\text{CFI} = .012$), suggesting that some intercepts varied by training level. Following recommendations on partial invariance [52], we freed the intercept of one item identified through modification indices, which restored acceptable fit ($\Delta\text{CFI} = .008$). This result supports partial scalar invariance, indicating that the scale worked largely the same way across training groups, with only minor differences at the item-level.

For the multiple-choice items, DIF analyses under the 3PL-IRT model did not flag any items across gender or prior AI training. This suggests that all items function equivalently across subgroups.

Together, these results indicate that our scale is psychometrically stable across gender and prior AI training subgroups, supporting valid use of the scale for between-group comparisons. Full fit indices of the invariance tests are provided in the appendix (Appendix C.9 and C.10).

5.5. Relationship Between Self-Assessed and Factual AI Understanding

We examined how participants' self-assessed AI understanding aligned with their factual understanding of AI as measured by our AI literacy scale. Given that both participant pools answered the same set of questions, and the analyses did not involve re-validating the factor structure, we combined data from both participant samples ($N = 476$).

We operationalized *Self-Assessed Understanding* as the mean of all 19 items spanning two subscales: *Functional Knowledge of AI* (eleven items) and *AI in Society* (eight items), each rated on a 7-point Likert-scale. In contrast, we measured *Factual Understanding* using the 18 multiple-choice items, each scored binarily (0 = incorrect and 1 = correct) and summed to yield a total score. The *Functional Knowledge of AI* subscale had a mean of 3.66 ($SD = 1.54$, $Median = 3.91$, $IQR = 2.55$), while the *AI in Society* subscale had a mean of 5.27 ($SD = 1.09$, $Median = 5.38$, $IQR = 1.38$). The composite score representing *Self-Assessed Understanding* had a mean of 4.40 ($SD = 1.24$, $Median = 4.47$, $IQR = 1.79$). Finally, the mean *Factual Understanding* score was 11.97 ($SD = 3.88$, $Median = 12$, $IQR = 5$).

To explore alignment between self-assessed and demonstrated understanding, we computed Spearman's rank-order correlations given violations of normality in Shapiro-Wilk test ($p < .01$ for all). Our analysis revealed that *Factual Understanding* was weakly correlated with the *Functional Knowledge of AI* subscale ($r_s = 0.11$) but had a modest correlation with the *AI in Society* subscale ($r_s = 0.22$) and the *Self-Assessed Understanding* score ($r_s = 0.17$). These results suggest limited alignment between perceived and factual understanding of AI measured by the scale. However, the two self-assessment subscales, *Functional Knowledge of AI* and *AI in Society*, were highly correlated with each other ($r_s = 0.62$), indicating internal consistency in the perceived understanding dimensions.

While the above correlations indicated limited association between *Self-Assessed* and *Factual Understanding*, they do not reveal whether individuals tended to systematically overestimate or underestimate their understanding. To address this, we conducted a calibration analysis, inspired by the seminal work of Kruger and Dunning [178], which highlighted discrepancies between individuals' perceived and actual performance. We computed a calibration score (i.e., a bias index score [3]) by subtracting standardized *Factual Understanding* score (multiple-choice) from standardized *Self-Assessed Understanding* score (Likert-scale), a widely used approach to compute miscalibration between subjective confidence and objective performance [3,157,240]. Standardizing both variables (via z-scores) accounts for the differing scales (Likert-scale vs. multiple-choice) and enables direct comparison on a common metric [123,126,216]. A positive score indicates overestimation, a negative score points to underestimation, and a score of zero reflects perfect calibration (i.e., no difference between standardized self-assessed and factual scores).

The mean calibration score was essentially zero (2.53×10^{-16}), and a Wilcoxon signed-rank test revealed no significant deviation from perfect calibration ($Z = 1.07$, $p = .28$). These results suggest that, at the group level, participants were well-calibrated in their assessment of their AI understanding. Yet, this overall pattern may overlook important differences among subgroups, which we investigate in subsequent sections.

5.5.1. Calibration Differences by Self-Assessed AI Understanding

Prior research has shown that group-level calibration scores can obscure individual differences, especially when patterns of overestimation and underestimation offset each other [169,170,245,343]. To examine how calibration patterns varied based on self-assessed AI understanding, we divided participants into *high* and *low* self-assessment groups ($N = 245$ and $N = 231$, respectively) using a median split of their *Self-Assessed Understanding* scores ($Median = 4.47$). As shown in Figure 24, participants in the *high* self-assessment group had a mean calibration score of $+0.69$ ($SD = 1.21$), indicating overestimation. In contrast, the *low* self-assessment group had a mean calibration score of -0.73 ($SD = 0.99$), reflecting underestimation. A Mann-Whitney U test confirmed that the difference in calibration scores between the two groups was statistically significant ($U = 16121$, $Z = 8.12$, $p < .001$). These findings suggest that participants who perceived themselves as more competent were more likely to overestimate their understanding, whereas those with lower self-assessments tended to underestimate it. This resembles a reversal of the Dunning-Kruger effect [178], in which less competent individuals overestimate their abilities and more competent individuals underestimate their abilities.

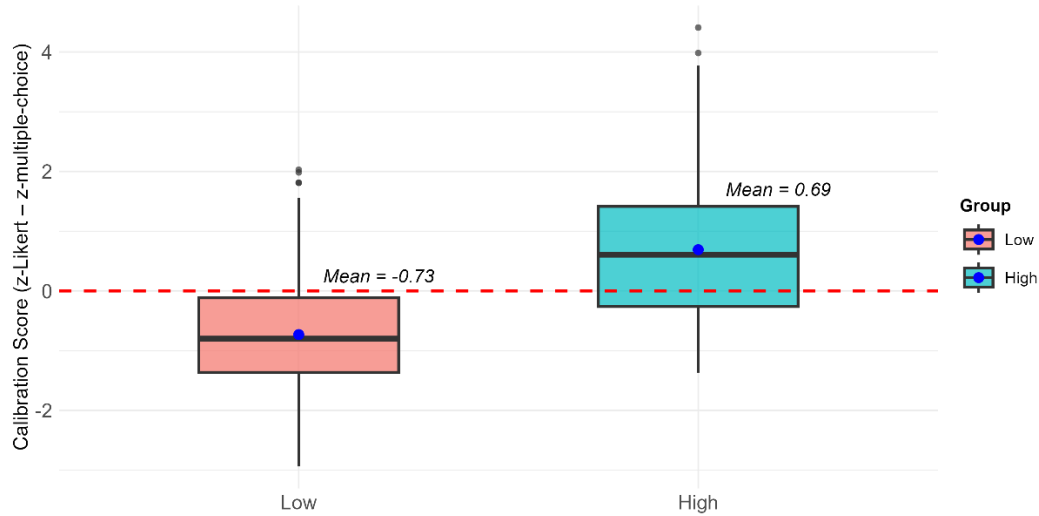


Figure 24: Distribution of standardized calibration scores (z-scored Likert – z-scored multiple-choice) by AI literacy group. The Low AI Literacy group (left) shows a negative mean score (-0.73), indicating underestimation, while the High AI Literacy group (right) shows a positive mean score ($+0.69$), indicating a tendency to overestimate their AI knowledge. The dashed red line represents perfect calibration (score = 0).

5.5.2. Calibration Differences by Gender

Motivated by prior research suggesting that gender can influence self-assessment tendencies, with men often showing greater overestimation than women [102,152,244], we conducted a gender-based analysis comparing knowledge calibration of participants who identified as either men or women. We did not include the one participant who selected a different gender identity due to the insufficient sample size. As shown in Figure 25, men had a slightly positive mean calibration score ($mean = 0.021$, $SD = 1.18$), suggesting a mild tendency to overestimate their understanding. In contrast, women had a slightly negative mean score ($mean = -0.012$, $SD = 1.44$), indicating a mild tendency toward underestimation. While the directional pattern is consistent with prior research on gender differences in self-assessment [102,152,244], the difference in calibration scores was not statistically significant ($U = 939.5$, $Z = 19.43$, $p = .48$).

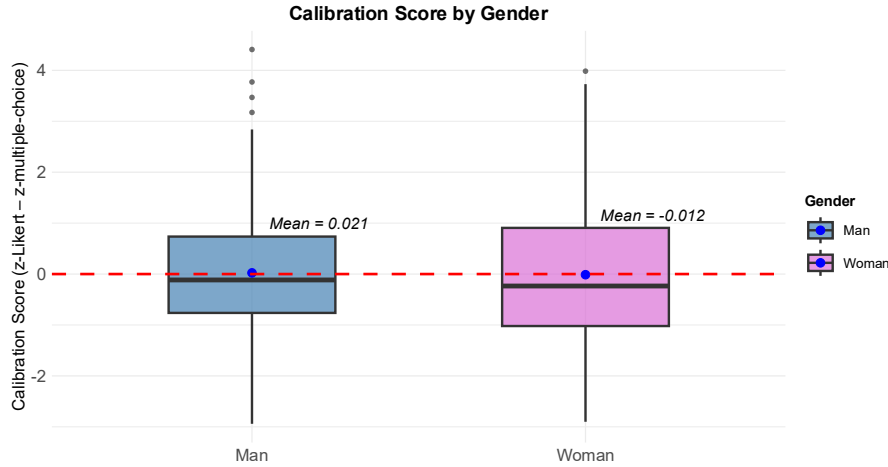


Figure 25: Boxplot of standardized calibration scores (z-scored Likert – z-scored multiple-choice) by gender. Men showed a slightly positive average calibration score ($M = 0.021$), indicating negligible overestimation of their AI knowledge, while women exhibited a slightly negative average score ($M = -0.012$), reflecting negligible underestimation. The dashed red line at zero represents perfect calibration.

5.5.3. Impact of Prior AI Training in Self-Assessed and Factual AI Understanding

We also examined how prior AI training influenced participants' *Self-Assessed* and *Factual Understanding* scores. We grouped participants into three groups based on their reported AI training: *No Training* (no formal education or training in AI), *Informal Training* (self-taught or gained through workshops, seminars, or on-the-job experience), and *Formal Training* (completion of formal courses or degrees specializing in AI). We conducted *Kruskal-Wallis H* tests to compare the three training groups across all self-assessed and factual understanding measures. As shown in Table 15, the results revealed significant group differences in all self-assessed measures. However, no significant difference was observed for *Factual Understanding*. Post-hoc Wilcoxon tests with Bonferroni correction confirmed significant pairwise differences across *all* AI training levels ($p < .001$) in *Functional Knowledge of AI* and *Self-Assessed Understanding*. For *AI in Society*, only the *No Training* group scored significantly lower than the other two.

Table 15: Median (IQR) values for Self-Assessed and Factual AI Understanding across AI training groups.

	Variable	No Training	Informal Training	Formal Training	H	Sig
Self-Assessed Understanding	Functional Knowledge of AI	2.18 (1.34)	4.09 (1.68)	5.18 (1.27)	230.69	<.001
	AI in Society	4.75 (1.47)	5.63 (1.31)	5.75 (1.25)	71.45	<.001
	Combined Score	3.24 (1.25)	4.68 (1.32)	5.37 (1.20)	201.50	<.001
Factual Understanding	--	12 (4)	12 (5)	14 (9)	5.47	.065

As demonstrated in Figure 26, median scores steadily increased with training for all the self-assessed measures. In contrast, median scores for *Factual Understanding* remained relatively stable, suggesting that increased AI training did not result in substantially higher *Factual Understanding* scores. These findings highlight the limitations of using self-reported AI training as an indicator of AI literacy and the importance of employing multifaceted assessment approaches.

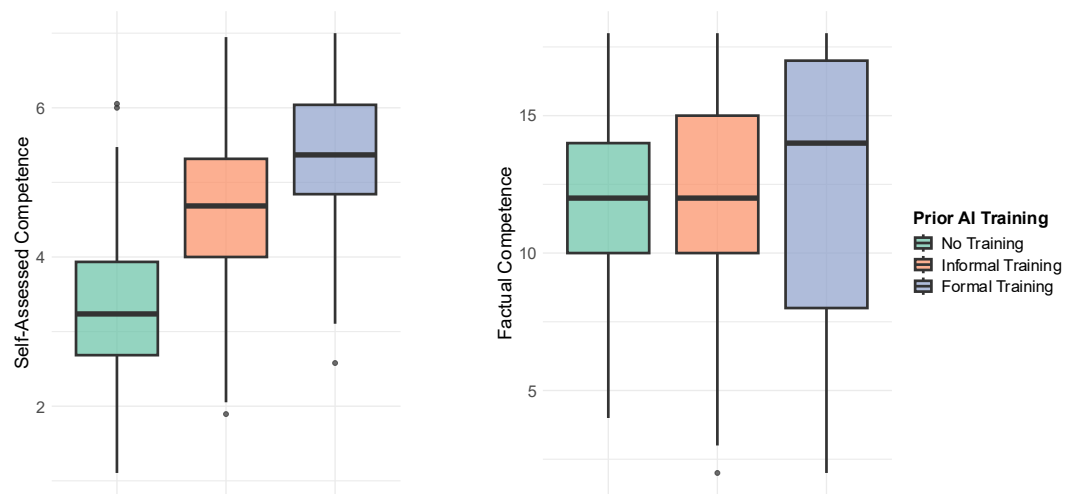


Figure 26: Medians for Self-assessed and Factual AI understanding across different AI training levels.

Figure 27 depicts calibration scores across AI training groups. Participants in the *No Training* group tended to underestimate their literacy ($mean = -0.79$, $SD = 1.07$), while those with *Formal Training* tended to overestimate their literacy ($mean = 0.70$, $SD = 1.43$). In contrast, participants with *Informal Training* were relatively well-calibrated ($mean = 0.18$, $SD = 1.09$). A Kruskal-Wallis test confirmed a statistically significant difference in calibration across groups ($H = 89.90$, $df = 2$, $p < .001$). Post-hoc pairwise Wilcoxon tests with Bonferroni correction revealed that all group differences were statistically significant (all $p < .01$).

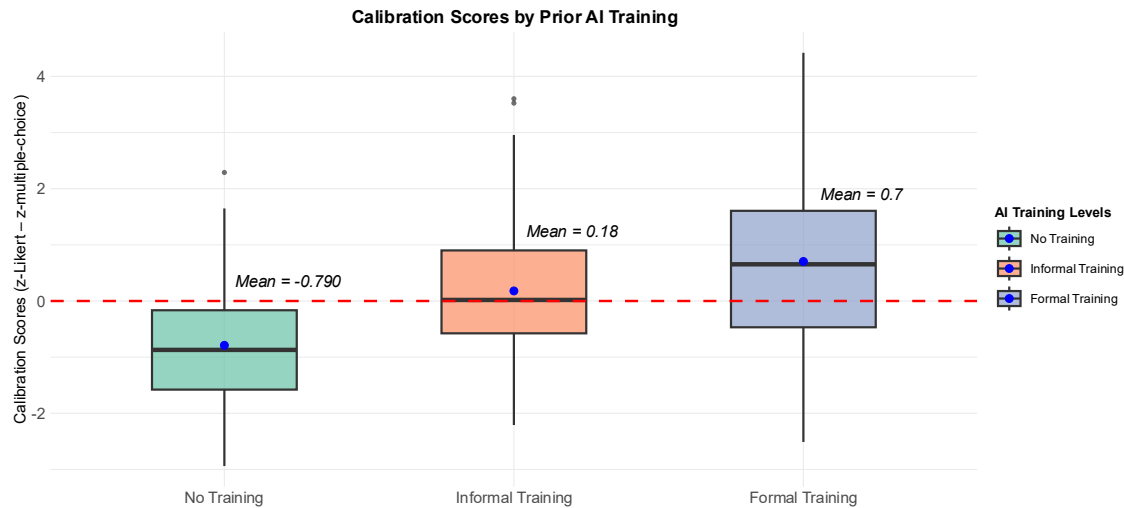


Figure 27: Boxplot showing calibration scores across AI training groups. Participants with no training had a negative mean calibration score, indicating underestimation, while those with formal training had a positive score, indicating overestimation. Participants with informal training were closer to zero on average (the red dashed line at zero denotes perfect calibration).

5.5.4. Robustness of the Reverse Dunning-Kruger Effect

To investigate whether the reverse Dunning-Kruger pattern was simply an outcome of dichotomizing self-assessment by median [209], we estimated regression models using self-assessment as a continuous predictor of calibration mismatch. Specifically, we regressed calibration score on *Self-Assessed Understanding* score while controlling for

prior AI training and gender. Although *Self-Assessed Understanding* score was based on ordinal data (i.e., Likert-scale), we treated it as continuous, which is consistent with common practice [233,266].

Self-assessment was a strong predictor of calibration mismatch ($b = 0.68$, 95% CI [0.58, 0.77], $p < .001$). The results were consistent when using z-scored *Self-Assessed Understanding* score as the predictor ($b = 0.84$, 95% CI [0.72, 0.96], $p < .001$). These findings indicate that participants with low self-assessed understanding tended to underestimate their understanding whereas those with higher self-assessments tended to overestimate. Figure 28 shows the calibration curve, where the regression line falls below zero at low self-assessment levels and rises above zero at higher levels, illustrating the reverse Dunning–Kruger effect. More detailed results are available in the appendix (Appendix C.11).



Figure 28: Calibration curve plotting Self-Assessed Understanding against Calibration Score. The upward-sloping regression line, crossing from negative calibration at low self-assessment to positive calibration at high self-assessment, illustrates the reverse Dunning–Kruger effect.

In contrast to the descriptive differences based on prior AI training observed in Section 5.5.3, neither *Informal* nor *Formal* training groups differed significantly from the *No Training* group once self-assessment was included in the model ($p > .05$). This indicates that prior AI training influences calibration indirectly by shaping where participants place themselves on the self-assessment scale rather than exerting a direct effect on mismatch. Gender, however, showed a small but significant effect ($b = -0.19$, $p = .039$), suggesting that, when holding self-assessment and training constant, men tended slightly more toward underestimation.

Together these results suggest that the reverse Dunning-Kruger effect is robust to alternate model specifications. The effect is primarily driven by self-assessment, with gender contributing a modest influence. However, this robustness check does not fully rule out the possibility of a broader mathematical artifact. Since we calculated calibration as the difference between self-assessed and demonstrated understanding, self-assessment was part of the outcome variable, while also being used as a predictor in the regression model. This creates a structural dependency between the predictor and the outcome, meaning that the calculation itself may produce part of the observed relationship rather than the relationship reflecting only a psychological effect. Researchers have raised similar concerns about the original Dunning-Kruger effect, suggesting that some apparent patterns may result from statistical artifacts [128,177]. Although our analysis shows the opposite direction, it involves a similar structural concern.

For this reason, we interpret the reverse Dunning-Kruger pattern cautiously. The findings show that participants' self-assessments did not always align with their demonstrated knowledge, but we do not treat this pattern as definitive evidence of a

stable cognitive bias. At the same time, the observed direction aligns with prior work on metacognitive monitoring in human–AI interaction, where higher AI literacy was associated with increased confidence but less accurate self-assessment of AI-assisted task performance [117]. We therefore treat this result as evidence of a calibration concern in AI literacy assessment, while acknowledging that the way we operationalized calibration may partly shape the pattern.

5.6. Discussion

Our work presents the development and validation of a dual-format scale for assessing AI literacy. Grounded in a widely cited AI literacy framework [207] and informed by established guidelines for scale development [39,61,92,143], our instrument integrates self-assessed (Likert-scale) and factual (multiple-choice) components to provide a comprehensive view of AI literacy. The scale demonstrates strong psychometric properties, with high internal consistency and evidence of construct validity.

5.6.1. Insights into Calibration and AI Understanding

Our findings demonstrate a weak correlation between participants’ self-assessed and factual understanding, suggesting that perceived and demonstrated AI understanding are not always closely aligned. Our findings also highlight calibration gaps when participants were grouped by both self-assessed understanding and prior AI training, resembling a reversal of the Dunning–Kruger effect [178]. Participants with low self-assessed AI understanding tended to underestimate their understanding while those with high self-assessed AI understanding overestimated theirs. Calibration patterns also varied across prior AI training levels; participants with no formal AI training generally underestimated

their understanding, those with informal training exhibited slight overestimation, and participants with formal AI training tended to overestimate their understanding more substantially. As shown in our regression analysis, the reverse Dunning–Kruger effect [178] remained robust when self-assessment was modeled as a continuous predictor. These calibration patterns challenge the common assumption that low-performers are more likely to overestimate their abilities and suggest a more complex interplay between confidence, AI literacy, and prior AI training.

Although these reversed patterns may seem counterintuitive, a closer reflection offers a plausible explanation. Given the perceived complexity and ubiquity of AI in everyday life, participants without formal training may lack confidence and undervalue their understanding of basic concepts, believing that they “do not know enough”. This aligns with prior work showing that self-perceptions are influenced not only by actual knowledge, but also by task familiarity, perceived domain difficulty, and self-efficacy beliefs [101,176]. With the landscape of AI changing rapidly, novices in AI may experience imposter-like effects, underestimating their understanding despite holding accurate or sufficient knowledge. Conversely, participants with formal exposure to AI might generalize their expertise beyond its actual scope, leading to inflated self-assessments. These results support the utility of dual-format assessment, which can reveal such discrepancies. Additionally, our findings indicate that prior AI training alone is not a reliable indicator of AI literacy, as training experience does not necessarily align well with self-perception or demonstrated understanding.

5.6.2. Conceptual and Measurement Challenges in Operationalizing AI Literacy

In defining the construct for our scale, we drew on a literature in which the terms AI literacy [74,207,228,323], AI competence [60,74], and AI ability are often used interchangeably and with considerable conceptual overlap. For example, AI literacy is frequently defined as the ability to clearly explain how AI technologies work and impact society, as well as to use them in an ethical and responsible manner and to effectively communicate and collaborate with them in any setting [188]. On the other hand, AI competence expands on these ideas by incorporating users' confidence, self reflection, and ability to apply their understanding in meaningful contexts [74]. Across both constructs, ability is referenced but not clearly delineated, which further complicates conceptual boundaries. Because our primary focus was to assess knowledge and understanding, we selected AI literacy as the foundational construct and grounded our work in one of the most widely cited definitions to preserve clarity and comparability with prior scales. At the same time, we acknowledge that our operationalization includes features that other researchers may associate with competence or ability, particularly in relation to confidence and the use of knowledge in practical decisions. Future work should examine how our scale aligns with AI competence-oriented frameworks and consider whether reframing or extending the instrument as an AI competence measure would offer a fuller account of users' capacity to apply AI related knowledge in practice.

Operationalizing Long and Magerko's AI literacy framework [207] involved trade-offs between conceptual clarity, item feasibility, and psychometric quality. Certain dimensions of the framework mapped more naturally to one format (e.g., ethics to Likert-scale, data literacy to multiple-choice), yet we included parallel items across both formats

to maintain conceptual alignment. Moreover, although we initially included items representing 16 of the 17 dimensions of the AI literacy framework [207], item reduction left *Recognizing AI* uncovered. The weak psychometric performance of our items related to this concept likely reflects the inherent ambiguity in identifying AI in practice [166], given the breadth and evolving nature of AI technologies.

5.6.3. Comparison to Existing Scales, Usage Guidelines, and Future Adaptability

Our scale differs from existing AI literacy assessment tools in several aspects. Most notably, it integrates self-assessment and factual understanding within a single validated scale, enabling triangulation between perceived and demonstrated understanding. Our scale also emphasizes breadth, covering 15 core dimensions of the AI literacy framework [207] with 37 items (19 Likert-scale, 18 multiple-choice), yet remains comparable in length to existing instruments, which typically range from 10 to 35 items [60,147,187,229,252,323]. Unlike many tools tailored to specific use cases, such as in education contexts [229] or the workplace [252], our scale uses general-purpose items (e.g., reasoning about algorithmic decisions, identifying risks in training datasets) for broader applicability across diverse population and contexts.

The factual understanding component enhances the value of our scale but presents challenges for long-term relevance given the rapid evolution of AI. As new AI paradigms emerge, what constitutes essential AI knowledge is likely to shift. For example, items related to the strengths and limitations of AI systems may need reinterpretation over time. We therefore advocate periodic reviews of AI literacy scales to maintain relevance and accuracy. Our IRT analysis further showed reduced sensitivity to differences among highly knowledgeable participants, suggesting a need for more challenging items to

enhance utility in expert-oriented contexts. A short-form version may also be useful where efficiency is critical, though this typically comes at some cost to reliability [148].

To guide future use, we recommend administering the instrument as a two-part scale that combines self-assessed and demonstrated AI literacy. The instrument consists of 19 Likert-scale items that yield two subscale scores corresponding to the two factors (Functional Knowledge of AI and AI in Society) as well as an overall score based on all items combined, and 18 MCQ items that form a single factual knowledge score. For the self-assessment part, Likert-scale items are coded from 1 to 7 and subscale scores are obtained by averaging items within each factor, and the overall self-assessment score can be calculated as the mean of the 19 items. For the multiple-choice section, each item is scored 1 for a correct answer and 0 for an incorrect response, then summed to get a factual knowledge score. Although we developed and validated the two sections jointly and encourage their combined use when researchers are interested in calibration between perceived and factual understanding, the structure of our scale also supports modular and longitudinal use. The two components do not assume a specific administration order, and each part can be administered and interpreted independently when only self assessed or only factual AI literacy is of interest. For example, the self-assessment (Likert-scale) component may be appropriate for large-scale surveys or educational contexts where quick insights into user confidence and perceptions are needed, whereas the factual (multiple-choice) component may be more suitable for knowledge-based evaluations. In longitudinal settings, the scale can allow researchers to track self-perceived and demonstrated understanding in parallel over time, offering insight into whether interventions improve knowledge, recalibrate confidence, or affect both simultaneously.

5.6.4. Limitations and Threats to Validity

While our scale offers a psychometrically validated approach to assessing both perceived and demonstrated AI understanding, there are limitations to our work that warrant further investigation. At the internal level, as with many survey-based instruments, participants' guessing, lapses in attention, or fatigue may have influenced responses despite the inclusion of quality checks. Regarding construct validity, some of the items may also overlap with related concepts such as data [95,332] or digital literacy [186,214], raising questions about the distinctiveness of AI literacy as captured by our scale. In addition, although we used expert reviews and cognitive interviews to refine items, we only relied on qualitative feedback from experts to iteratively improve items and did not compute a formal content validity index (CVI or iCVI) [254]. As a result, we cannot quantitatively report the degree of agreement on relevance across experts, which may leave gaps in how comprehensively the items represent the broader construct of AI literacy.

Beyond these construct-level consideration, our analytic choices also introduce limitations. Although we refined the self-assessed items using both factor analysis and IRT, we applied IRT alone to the multiple-choice components due to the unstable factor structure in the dataset. This limited our ability to identify latent dimensions within the factual items, constraining comparisons across formats at the construct level. Moreover, while multiple-choice items offer practical benefits for assessing factual understanding, they may be limited in their ability to capture the depth, nuance, and contextual reasoning essential for real-world AI understanding. Future work could explore complementary methods such as open-ended questions, scenario-based reasoning, or

behavioral tasks to better assess deep understanding and practical AI literacy. External validity is also a concern, since both of our samples were recruited through Prolific platform, which largely represent a digitally literate, global-north demographic [303]. Our samples also skewed towards adults aged 25 to 34 with informal AI exposure. Future validation should extend to different demographics and backgrounds, including older adults, workers with low AI exposure, and adults in the global south.

Finally, although we found a reverse Dunning-Kruger effect in the calibration of participants' self-assessed and demonstrated AI understanding, our study does not reveal the underlying causes of this effect. We discussed possible explanations for this effect in Section 5.5, however, we acknowledge that alternative explanations exist. For example, it is possible that the wording of the items resonated differently across participants with different AI backgrounds and training. It is also possible that the scale captures certain aspects of perceptions and attitudes toward AI more strongly than it captures the full breadth of AI literacy. Moreover, our operationalization of calibration as the difference between self-assessed and demonstrated understanding means that the observed pattern should be interpreted cautiously, as it may partly reflect the structure of the measure itself rather than a stable cognitive bias. Future work should systematically investigate these possibilities to produce a deeper understanding of the calibration patterns observed in this study.

5.6.5. Implications for Design

Our dual-format scale has practical applicability for the design and evaluation of AI systems. In XAI, tailoring explanations to users' prior knowledge is a recurring goal [1,110,179,180,223], yet most systems lack reliable ways to distinguish between users who

misunderstand, underappreciate, or lack confidence in AI output. By capturing both perceived and demonstrated understanding, our scale enables more targeted personalization. For example, users with high confidence but low factual understanding may benefit from guidance from the systems, while underconfident but knowledgeable users might need reassurance or opportunities to validate their reasoning. In adaptive systems, this distinction could guide how explanation detail [181] or modality [302] is adjusted based on individual profiles. As an illustrative example, a writing assistant could measure AI literacy during system onboarding and adapt interaction accordingly: novices might receive strong guardrails and brief rationales [6,179], intermediates could be offered comparison explanations that highlight alternatives [38,96], and advanced users could gain access to model [225] or data cards [127] alongside more direct prompt controls. Such tailoring could reduce AI error acceptance among novices, while enhancing task success and efficiency for advanced users. Similar personalization applies in educational [215,325] and system onboarding contexts [9,55], where the scale can differentiate between those who need conceptual instruction and those who may benefit more from confidence-building strategies.

Our AI literacy scale can also support identifying and addressing intervention-generated inequalities [316] by assisting in technologies audits [47,259]. It can reveal when an intervention increases users' confidence without improving their factual understanding, or benefits highly knowledgeable users disproportionately while leaving others underserved. Detecting such mismatches can inform redesign or instructional strategies to ensure more equitable impact. In this way, the scale can not only highlight

inequities but also guide more equitable educational and technological practices that foster inclusion and avoid reinforcing existing divides.

5.7. Summary

In this chapter, we introduced and validated a dual-format scale for assessing AI literacy that combines self-assessed and factual measures. Beyond establishing the psychometric validity of the scale, our work also revealed misalignment between people's perceived and factual understanding, with patterns that challenge common assumptions about overconfidence. By offering a comprehensive and diagnostic understanding of AI literacy, our scale supports identifying gaps between people's confidence and their actual understanding, evaluating educational and technological interventions, and advancing equitable system design.

Chapter 6

Reflections on Designing Effective (Training Dataset) Explanations for Diverse Users: Insights, Challenges, and Considerations

This dissertation investigates how to design effective training dataset explanations for diverse users through three complementary studies. Together, these studies reveal key opportunities and trade-offs in supporting a wide range of users. In this chapter, we first synthesize and reflect on the insights gathered from the three studies. Next, we discuss considerations and challenges for designing training dataset explanations for broad audiences. Finally, we reflect on the methodological challenges encountered in this dissertation.

6.1. Insights

6.1.1. Interaction Between Explanation Design Factors and Users' AI Knowledge

The first two studies of this dissertation explore how presentation style and information depth (i.e., design factors) influence users' understanding of training dataset explanations and their assessments of AI systems. A central insight from these studies is that the effectiveness of explanations is shaped not only by what information is included but also by how the information is communicated. In the presentation style study (Chapter 3), both the Data Story and Q&A-based approach conveyed the same underlying content, yet participants focused on different aspects of the explanation depending on the presentation format. A similar pattern emerged in the information depth study (Chapter 4), where participants' focus on the explanation content shifted with the level of detail. Summary explanations abstracted away finer-grained details, which limited users' ability to engage with specific aspects of the training data but tended to encourage a holistic interpretation of the explanation. In contrast, detailed explanations offered richer, more specific insight at the cost of increased cognitive effort and often drew attention to specific aspects of the explanation. These results illustrate how design choices can influence users' interpretation and usage of explanations.

While these patterns may appear somewhat intuitive, they are conceptually important because XAI research has often emphasized explanation content as one of the primary determinant of effectiveness [181,227,320], with explanations that share the same core information commonly treated as functionally comparable in evaluation practices. Our findings challenge this assumption by showing that even when

informational content is held constant, design choices can alter user understanding and system perceptions.

Across both of these studies, we also observed indications that participants with different AI backgrounds interpreted and evaluated the explanations in distinct ways. Although we did not examine these differences systematically (owing to challenges in categorizing participants according to their AI knowledge), this variability highlights that the effectiveness of explanation design does not operate independently of its audience. Rather, explanation effectiveness is multifaceted and emerges from the interaction between design features, users' AI knowledge and experiences, and explanation use cases. This motivates AI literacy-aware design and evaluation practices that move beyond one-size-fits-all transparency. Our third study (Chapter 5), which developed a scale to assess AI literacy, provides an initial step toward more systematic investigation of these user-level differences.

6.1.2. Calibration Insights and Design Implications from AI Literacy Scale

Our dual-format AI literacy scale revealed mismatches between users' perceived and demonstrated AI understanding. The reverse Dunning-Kruger pattern we observed indicates that some participants expressed high confidence in their AI understanding while scoring relatively low on the factual items, whereas others underestimated their competence despite strong performance. These patterns of overestimation and underestimation indicate that users differ not only in how much AI knowledge they possess, but also in how they judge their own level of understanding. Our findings point to calibration between perceived and demonstrated understanding as a potentially important user characteristic for future work on explanation design and evaluation. For

example, underestimation may warrant confidence building explanation strategies, such as clearer affirmations of what users already understand, whereas overestimation may call for explanation designs that surface potential gaps in understanding and direct users toward resources that strengthen their underlying knowledge.

The reverse Dunning-Kruger pattern we observed also complicates existing concerns that explanations can contribute to overreliance by increasing users' trust in AI systems without a corresponding increase in understanding [25,247,255,326,340]. Prior work has shown that explanations can raise users' confidence and willingness to rely on system outputs even when their mental models remain incomplete [25,163,255]. Our results shed light on how this effect might unfold in practice: when users overestimate their AI knowledge, they may also overestimate how well they have understood an explanation, resulting into placing unwarranted trust in the system. Conversely, users who underestimate their knowledge may be less inclined to question outputs or may defer more readily to the system, even when an explanation provides enough information to support a reasonable critique. More broadly, these miscalibration patterns highlight how poor calibration may shape how users interpret and rely on explanations, reinforcing the need for explanation designs that support both understanding and more accurate self-assessment.

Our scale development also highlighted practical challenges and opportunities in operationalizing AI literacy for diverse populations. Although the self-assessment component yielded two factors, the high correlation between them suggests either that users may experience technical and societal dimensions of AI knowledge as closely intertwined in practice, or that the scale does not fully disentangle these dimensions. This

overlap between technical and societal dimensions of self-assessed AI knowledge cautions against explanation strategies that assume clearly distinct “technical” and “social” user profiles and focus designs on only one dimension at a time. At the same time, the multiple-choice items formed a single dimension, which limits the scale’s ability to distinguish between types of objective AI knowledge and points to the need to expand and refine the item pool in future work.

6.1.3. Adaptive Explanation Interfaces and Onboarding

The findings from our explanation design studies and our AI literacy scale also bring forward a range of opportunities for flexible interaction and for developing transparency mechanisms that better support diverse users. In particular, these findings highlight the importance of flexibility during early, exploratory onboarding interactions, when users are still forming initial understandings of system behavior. The first two studies show that explanation design choices such as presentation style and information depth systematically shape what people notice and how they interpret system behavior. Given that these effects are not uniform across users or contexts, these results suggest a potential need for adaptive explanation interfaces rather than one-size-fits-all formats. Interfaces could offer flexible pathways that allow users to adjust the level of depth, move between narrative and modular structures, and access additional context on demand. Such adaptability can support a wide range of users by matching the complexity of explanations to their immediate goals, while still preserving opportunities for deeper exploration when desired.

6.2. Considerations

6.2.1. Balancing Transparency, Usability, and Learning

Our studies revealed that participants found training dataset explanations beneficial. However, across the explanation designs we examined, data-centric transparency surfaced a familiar tension between transparency and cognitive effort [181,278]. In particular, Chapter 4 showed that more detailed explanations were associated with higher perceived understanding and trust, but also with noticeably greater mental effort. Progressive disclosure improved perceived learning, but it provided only limited support to manage the cognitive demands introduced by detailed explanations.

Although participants preferred detailed explanations, this preference was observed within the range of explanation detail studied in this dissertation and may not extend to explanations that introduce substantially greater volume or complexity. Prior work has shown that excessive detail can leave users frustrated, hinder their ability to process information, and reduce the overall usability [69,167,181]. Designers must therefore consider not only what information to reveal but also how much information users can absorb at a time. Adaptive interfaces may provide a potential path forward by adjusting explanation detail to better align users' capacities and needs. The AI literacy scale developed in this dissertation provides one mechanism for supporting such adaptation, enabling systems to tailor explanation for a diverse set of users.

6.2.2. Designing for Actionability, Not Just Transparency

Across our studies, explanations improved participants' perceived understanding of the training data and shaped their perceptions of fairness and trust. However, they offered

limited support for translating that understanding into concrete actions. This pattern reflects a broader explainability challenge in which explanations succeed at informing users but fall short of enabling action [107,196,197]. When explanations stop at interpretation, they can increase users' confidence without providing a clear pathway for what to do next, which can contribute to passive acceptance [197] or inappropriate reliance [25,46,196,255]. This highlights an important distinction between transparency as information provision and transparency as a basis for action.

For data-centric explanations in particular, translating insight into action remains an open design problem because what users should do next depends on their role and the organization's workflow, not just the explanation itself. Our studies showed improved understanding and trust without corresponding support for determining what to do next with training dataset explanations. Accordingly, actionability emerges as a key next design challenge for data-centric explanations, one that can be explored by pairing explanatory insights with plausible next steps at the point of use [5,284,314]. In the context of training dataset explanations, this might include options to inspect relevant data subsets, compare groups or time periods, report concerns, view alternatives, or escalate issues to human oversight. Future work on training dataset explanations should examine how explanatory insights can be meaningfully integrated into users' existing roles, workflows, and decision-making practices.

6.2.3. Intervention-Generated Inequalities in Training Dataset Explanations

Prior work defines intervention-generated inequalities as well-intentioned design choices that unintentionally amplify inequalities among users [316]. Our findings offer initial indications that explanation designs can benefit users differently, pointing to

conditions under which training dataset explanations may give rise to intervention-generated inequalities. This aligns with prior research demonstrating that individual differences influence how users engage with and benefit from explanatory interventions. For example, prior work on individual traits such as Need for Cognition (i.e., a stable personality trait that reflects how much a person enjoys engaging in cognitively demanding activities [53]) has shown that users can receive uneven benefits from the same intervention.

Our findings point to the importance of considering additional, design-relevant user characteristics that shape how explanations are interpreted and used. In particular, our results highlight variability in how users benefit from different explanation designs, pointing to the potential role of differences in users' AI knowledge and experiences. One-size-fits-all explanation strategies therefore may risk unintentionally privileging users who are already better equipped to engage with training dataset explanations, while leaving others less supported in making sense of the information [249]. This highlights the need for explanation approaches that explicitly account for variation in users' backgrounds, rather than assuming uniform ability to engage with explanations.

The AI literacy scale introduced in this dissertation provides a practical tool for studying and operationalizing such user-level differences. By capturing variation in users' conceptual understanding of AI, as well as discrepancies between self-assessed and demonstrated knowledge, the scale can enable researchers to examine how different explanation designs interact with user capabilities. This, in turn, supports future work on literacy-aware explanation systems that adapt explanation depth, structure, or format to better support a diverse range of users.

6.3. Methodological Reflections and Challenges

6.3.1. Motivating Engagement Through Structured Critique Tasks

Motivating participants to critically reflect on explanation content in a scenario-based study presents a methodological challenge [46]. To address this, we simulated external motivation by asking participants to perform a SWOT analysis task. While this approach supported engagement, it may also have prompted deeper and more deliberate interaction with the explanations than what one might see in a more ecologically valid setting. Therefore, engaging users with the explanation content outside of study contexts might also require facilitating similar structured tasks (e.g., SWOT or SCORE [344] analyses) or providing other types of critiquing frameworks (e.g., Data Experience Critique framework [321]). Future work should conduct field studies to investigate how to engage participants with long-format explanations outside of a study context. Future research should also examine the generalizability of our findings to a broader variety of tasks, including those that do not inherently prompt critical thinking, such as simple information retrieval [213] or decision-support tasks [237].

6.3.2. Reflections on Hybrid and Remote Study Participation

For the studies in this dissertation, we recruited participants from a range of sources, including local participants (e.g., university students, snowball sampling), remote participants (e.g., recruited from online platforms such as Reddit, LinkedIn, and X), and crowdsourced participants (e.g., recruited from Prolific). The first study (Chapter 3) was conducted during a period of transition following the COVID-19 pandemic. With in-person studies previously being a norm in HCI research and respecting users' preferences,

we conducted the study in a hybrid setting that accommodated both remote and in-person participation. This approach increased flexibility and broadened the participant pool, but it also raised concerns during peer review regarding methodological consistency. In response to these concerns, we conducted our following study (Chapter 4) in a fully remote setting.

Conducting studies in different formats allowed us to reflect on the trade-offs between hybrid and remote participation. Hybrid studies broadened the participant pool and attracted more intrinsically motivated participants, particularly for in-person sessions. However, hybrid formats also introduced concerns related to consistency across study conditions, such as differences in participant environment and researcher presence. In contrast, remote studies expanded the participant pool and made participants feel more at ease in their own environments, but they also presented additional challenges, including reduced environmental control and technical difficulties. Remote studies were also more vulnerable to variability in participant effort and attention, which introduced additional noise into the data.

These observations motivate a set of methodological considerations for future studies. In our studies, we found pre-screening questionnaires helpful for identifying participants who were willing to engage more carefully with the tasks, particularly in remote settings. We also experimented with emphasizing both intrinsic motivations (e.g., framing the study as a learning opportunity) and extrinsic incentives (e.g., monetary compensation), which appeared to support engagement in our context. When the study context permits, our experience suggests that remote participation can be effective, as its flexibility and

broader reach may outweigh some of the limitations observed in hybrid formats. We presented these reflections and trade-offs in a workshop paper at IEEE VL/HCC 2024 [10].

6.3.3. Challenges to Measuring Users' Understanding of Explanations

Measuring users' understanding is a well-known challenge across many research areas. In our studies, we approached this challenge by foregrounding critique-based measures, complemented by comprehension questions and self-reported assessments, to capture different facets of users' understanding of both the explanations and the AI systems.

Our use of critique-based measures was motivated by prior work suggesting that comprehension is at the highest level when the reader can apply their understanding beyond the information explicitly presented [4]. Accordingly, we assessed participants' comprehension of the explanation through their critique coverage and a broad assessment of their critique accuracy. However, evaluating critique accuracy proved challenging due to variability in how participants interpreted the same information and the absence of a clear ground truth (i.e., an objective standard used to evaluate the correctness of participants' critiques) [337]. These challenges highlight the difficulty of operationalizing comprehension when explanations invite interpretation and contextual reasoning rather than fixed answers.

To complement critique-based measures, we also used comprehension questions as an objective assessment of understanding. While comprehension questions can provide a more direct assessment of understanding, their effectiveness depends heavily on how they are designed and scoped. For example, a larger number of questions or questions probing deeper reasoning could yield more fine-grained insights, but may also increase

participant burden and require participants to draw on background knowledge beyond what is provided in the explanation. In our case, the comprehension questions were intentionally designed at a surface level to ensure they could be answered based solely on the explanation, without requiring additional prior knowledge. As a result, these questions primarily captured basic recall and recognition rather than deeper reasoning about the system. Future work should explore alternative question designs that balance depth and coverage and are understandable to a broad range of users based on the information provided in the explanation.

We also included self-reported measures of understanding to capture participants' subjective perceptions. While such measures offer a straightforward and efficient way to collect data, prior work has questioned their reliability, as self-reported understanding may not accurately reflect actual comprehension [84,315]. This concern was also evident in our data, reinforcing the need to interpret subjective measures alongside other indicators of understanding.

More broadly, given that comprehension can be influenced by multiple factors [309], including text structure and vocabulary, reading skills, motivation and interest, and task context, developing methods to assess explanation comprehension in a rigorous, tractable, and accurate manner remains an important future research direction. These efforts may be complemented by additional data collection techniques, such as think-aloud protocols [289] or eye-tracking [32,221], to help interpret comprehension outcomes by revealing how users attend to and reason about explanation content.

6.3.4. Reflections on the Context and Task Specificity

Throughout this work, we defined AI systems and explanations at a relatively high level, with limited specification of domain constraints, task requirements, and decision consequences. We focused on high-stakes domains in our study scenarios, as earlier work on training dataset explanations suggested greater relevance in such contexts [8]. However, defining AI systems and explanations at this level of abstraction introduces an important consideration for interpreting the findings.

In practice, explanations are encountered within situated activities shaped by domain expectations, workflow constraints, and varying consequences of decision-making. As a result, what constitutes an effective explanation, or what qualifies as sufficient understanding, may vary substantially across contexts. This has direct implications for both the design and evaluation of explanations. For example, professionals reviewing AI-assisted recommendations in high-stakes environments may require different levels of detail, structure, and actionability compared with lay users interacting with AI in everyday settings. Treating explanations as context-independent constructs may therefore overlook important variations in how explanations are interpreted and evaluated.

This observation points to an important avenue for future work by encouraging more explicitly situated evaluations that examine how explanation design interacts with task goals, domain practices, and user roles. Such context-sensitive investigations would help clarify the generalizability of the findings presented here and support the development of explanation approaches that are aligned with real-world usage conditions.

6.4. Summary

This chapter reflected on the empirical findings and methodological experiences across the studies in this dissertation. Collectively, the studies highlight that the effectiveness of training dataset explanations depends on the interaction between explanation design choices, such as presentation style and information depth, and users' AI literacy, contexts of use, and cognitive resources. The chapter also emphasized calibration between perceived and demonstrated understanding as a key consideration for explanation design and evaluation. In addition, it surfaced broader design and methodological challenges, including supporting actionability, mitigating uneven benefits across users, and rigorously assessing explanation comprehension. These reflections motivate AI literacy-aware explanation approaches and more rigorous evaluation practices that move beyond one-size-fits-all transparency.

Chapter 7

Conclusion

Despite the central role of training data in shaping the behavior and limitations of AI systems, data-centric transparency remains limited in practice. As a result, users often lack critical context about the systems they rely on, which limits their ability to assess outputs, make informed decisions, and collaborate with AI effectively. To support more transparent and informed use of AI systems, we need to understand how to design training dataset explanations that can serve a wide range of users. With this motivation, this dissertation investigates how design choices in training dataset explanations shape users' understanding of explanations and their assessments of AI systems, as well as introduces a validated instrument for assessing users' AI literacy to support designing explanations for diverse users.

Across three complementary studies, this dissertation provides empirical evidence on how training dataset explanations shape users' understanding and assessments of AI systems, while identifying design trade-offs and offering guidance on using presentation

style and information depth when crafting explanations. It also delivers a dual-format AI literacy instrument that can enable future researchers to comprehensively assess users' AI literacy and design explanations that are responsive to users' needs. This chapter summarizes the dissertation's key contributions and limitations, as well as promising avenues for future work.

7.1. Thesis Contributions

At a broad level, this dissertation highlights that an effective explanation results from a successful interaction among explanation content, presentation, and the user. This perspective treats explanations as elements that become meaningful only through users' interactions with them. The studies reveal variation in how users make sense of and use training dataset explanations. This motivates future work to investigate whether adaptive explanation strategies that respond to users' AI literacy and related individual differences can better support comprehension and informed AI-supported decision making. Below we discuss the dissertation's primary contributions.

7.1.1. Contributions to Assessing Users' Understanding of Explanations

The dissertation provides empirical evidence on how presentation style and information depth of training dataset explanations shape users' understanding, perceptions, and assessments of AI systems. By comparing training dataset explanations presented as a Data Story and in a Q&A-based approach with the same underlying information, we show that structure and framing of information influence what users notice, how they interpret the explanation, and the types of concerns they raise. Similarly, the exploration of information depth and progressive disclosure reveals trade-offs between clarity,

cognitive load, and perceived learning. These findings move explainability research toward a more nuanced understanding of how specific design decisions impact user outcomes.

Beyond explanation design, this research also contributes insights into how users engage with training dataset explanations during onboarding settings. While onboarding has been identified as an important context for AI systems, this dissertation presents concrete findings on how training dataset explanations are used and interpreted to form initial impressions of a system's strengths and weaknesses. These findings support the importance of onboarding as a critical phase for explanation design and provide guidance for creating explanatory materials that support users who are encountering an AI system for the first time.

7.1.2. Contributions to User Modeling and AI Literacy Measurement

Another key contribution of the dissertation is the development and validation of a dual-format scale for assessing users' AI literacy. The scale combines self-assessed and demonstrated AI understanding into a dual-format structure that provides a richer and more generalizable measurement framework for capturing users' AI knowledge and experiences, enabling future research to examine calibration and miscalibration in explanation evaluation. The validation process establishes the scale as a psychometrically sound instrument with stable factor structure, strong reliability, and well-performing items across both its self-assessed and factual components. This instrument has the potential to support researchers and designers in characterizing user populations, tailoring explanation strategies, and identifying differences between self-assessed and demonstrated AI understanding.

The dissertation also provides early evidence for a reverse Dunning-Kruger pattern in the calibration of users self-assessed and demonstrated AI understanding. This finding highlights challenges in relying solely on subjective measures when designing explanations or evaluating user abilities. It underscores the need for multi-dimensional assessments that capture both confidence and knowledge to support equitable and informed human-centered AI design.

7.1.3. Methodological Contributions

From a methodological perspective, this work offers contributions that can support future human-centered XAI research. It introduces critiquing as an explanation goal. By asking users to critique AI systems based on the training dataset explanations, we gained insight into how users interpret, integrate, and apply explanatory information and how this, in turn, shaped their assessment of the systems. This approach complements traditional comprehension measures by capturing the analytical and evaluative reasoning that users perform when interacting with AI systems.

7.1.4. Contributions to Human-Centered Explainable AI

By foregrounding training dataset as a critical aspect of transparency, this research contributes to the field of human-centered explainable AI. Through empirical examination of how people use training dataset explanations and how design factors shape impact that usage, this work expands the scope of human-centered XAI research and highlights the value of data-centric transparency in supporting users' assessment of AI systems. It contributes to the reframing of transparency from model-centric transparency to a more socio-technical artifacts that data used to train the system, the

design decisions that shaped the system, and the humans behind it. In doing so, it highlights that explanation design must account not only for what the model does, but also for how the data was collected, curated, and used, and how users make sense of this information. This supports a broader shift toward explanation practices that are aligned with the needs and expectations of users.

7.2. Limitations

7.2.1. Scenario-Based Study

Both our explanation design studies (Chapter 3 and 4) employed a scenario-based approach. While this aligns with many prior studies on AI explainability [8,38,103–105,191,297] and allowed us to isolate the effects of presentation style and information depth, it may not fully capture the complexities and nuances of real-world applications. Future studies should validate our findings in more real-world settings that involve users in deployed or domain-specific contexts, domain knowledge, and situated decision-making.

7.2.2. Generalizability Across Contexts

Our explanation design studies were conducted in a controlled setting with task-specific contexts (e.g., critiquing training dataset explanations). While this provided experimental control, it may limit the transferability of findings to other forms of AI explanations such as decision-level, model-level, or counterfactual explanations. Our findings might not be applicable to different AI system scenarios. For example, with lower-stakes systems, users might be less motivated to engage with the explanation [48]. The characteristics of

the training dataset might also influence the nature of the findings, for example, some training datasets might be more difficult to describe (e.g., audio recordings or videos), leading to poorer comprehension than we observed. The types of information provided in the explanation might also influence the direction of users' critiques. Future work should explore how these findings hold across domains, systems, and usage contexts.

Our studies focused on onboarding scenarios, which represent only one stage of the AI use cycle. Future research should investigate other stages of the AI workflow and examine these dynamics in real deployment settings across varied domains. Additionally, since this dissertation focused on immediate comprehension and critique, longitudinal studies are also needed to understand how explanations affect learning, calibration, trust, and decision quality over time.

7.2.3. Sampling Bias and Participant Diversity

Participants in our studies were recruited from a wide range of channels (e.g., university students, snowball sampling, online platforms such as Reddit, X and LinkedIn, and the Prolific platform). This recruitment strategy introduced diversity in age, gender, professional roles, and geographic backgrounds, as well as variation in participants' prior experience with AI. While the sample was diverse in these respects, it may not fully reflect broader populations with respect to AI literacy, cultural expectations, educational backgrounds, or linguistic diversity. Across studies, our participant pool tended to be on the higher end of both education and AI literacy, which means the findings may not capture the needs and behaviors of individuals who are underrepresented in online research or who have limited digital access. In addition, while this dissertation acknowledges that users vary widely in their knowledge and experiences, the studies did

not deeply examine how intersectional factors such as gender, language, or socioeconomic status shape AI literacy or explanation effectiveness. Addressing these dimensions in future work would support the creation of explanations that are more inclusive, equitable, and sensitive to varied user experiences.

7.3. Future Research Directions

7.3.1. Advancing Adaptive and User-Aware Explanation Systems

A key direction for future work is leveraging the AI literacy scale to categorize users and investigate how individuals with different levels of AI literacy use training dataset explanations. Such studies can help advance explanations that adapt to users in real time. Future research can also pursue sociotechnical co-design of explanations by collaborating with actual system users, domain experts, and marginalized groups. This would enable researchers to identify which forms of dataset knowledge are most relevant to different interested parties and use these insights to inform explanation design. Finally, training datasets shape not only AI outputs but also institutional decision-making practices. Future work should examine how training dataset explanations relate to practices in auditing, regulation, and public communication, as well as how this transparency influences broader accountability practices.

7.3.2. Operationalizing Richer Models of AI Literacy and Competence

Although we operationalized AI literacy through Long and Magerko's widely cited framework [207], both AI literacy and AI competence are evolving concepts and multiple overlapping definitions and constructs exist. The dual-format AI literacy scale introduced

here opens opportunities for richer theoretical development and more refined measurement of AI literacy and competence. Future research can refine the construct, improve measurement models, and examine how AI literacy develops over time. There is also value in studying how competence relates to real-world behaviors, such as appropriate reliance, misuse, overreliance, or avoidance of AI systems. These directions can help establish AI literacy as a practical and predictive framework for guiding explanation design and evaluating user interactions with AI systems.

7.3.3. Extending Transparency Beyond Training Datasets

While we broaden system transparency by promoting and investigating data-centric transparency, future work can further extend transparency beyond the training dataset to illuminate other stages of the AI pipeline that also shape system behavior. For example, explanations can describe elements of the training process such as objectives, tuning procedures, and evaluation practices in accessible terms that help users build a grounded understanding of how the system learned from the data. Further, explanations can illustrate how model behavior is refined through human oversight during training and validation, including how researchers interpret errors, adjust model settings, and decide when a model is considered acceptable for deployment. These forms of process-level transparency would complement training dataset explanations by showing not only what the data contains but also how it is used to shape system behavior, supporting more informed reasoning about system capabilities and limitations.

Transparency could also extend into deployment contexts where systems encounter real world constraints, distribution shifts, and evolving data. Users may benefit from explanations about system boundaries, known failure patterns, and how updates affect

behavior over time. These forms of contextual and dynamic transparency could help users understand when to trust the system, when to intervene, and how to adapt their expectations as the system changes. Transparency is also social in that it makes visible the socio-organizational context around AI mediated decisions, including who designs and operates the system, what they do, and when and why key decisions are made. Future work could design and evaluate transparency mechanisms that jointly convey data practices, model behavior, deployment conditions, and the socio-organizational context of AI decisions [103]. Together, these directions shift transparency toward a more holistic approach that supports informed, responsible, and context-aware use of AI systems.

7.3.4. Exploring Other Explainability Pitfalls

We explored information depth as both a design factor and explainability pitfall. This leaves opportunities to examine other pitfalls. Future work should investigate additional pitfalls (outlined in Chapter 4) such as misinterpretation, frustration, lack of actionability to understand how they arise in the design and deployment of explanations and how they influence users' understanding, trust, and decision making. Such work can inform the development of pitfall-aware explanation strategies that explicitly anticipate, surface, and mitigate these issues rather than treating them as afterthoughts. By integrating knowledge about explanation pitfalls into the design process, future work can help ensure that explanations remain effective, support critical engagement, and provide as much benefit as possible to the people who rely on AI systems.

7.3.5. Long-Term Vision

This dissertation lays the foundation for a broader agenda that advances user-centered transparency in AI systems by integrating design, evaluation, and measurement. Across the three studies, we showed that specific design factors shape how users comprehend training dataset explanations and, in turn, how they assess the AI systems that rely on those datasets. The findings also underline that there is no one-size-fits-all solutions for effective explanations. A long-term goal is to build explanation ecosystems that adjust their depth, structure, and delivery to match the needs, abilities, and contexts of diverse users. The AI literacy scale we developed (and related tools) provides paths toward such adaptation by offering a way to assess users' AI understanding and to use this information to guide explanation design. Doing so will require continued work on transparent data practices, more robust measurement of AI literacy, and richer understanding of how users reason about AI systems' behavior in real settings. By continuing to refine explanation designs, evaluate their impact in real-world settings, and improve the tools for assessing users' AI literacy, the field can move toward AI systems that are more equitable, trustworthy, and better support informed human–AI collaboration.

Bibliography

1. Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 1–18. <https://doi.org/10.1145/3173574.3174156>
2. Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6: 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
3. Mirian Agus, Maribel Peró-Cebollero, Joan Guàrdia-Olmos, Igor Portoghese, Maria Lidia Mascia, and Maria Pietronilla Penna. 2019. What's about the Calibration between Confidence and Accuracy? Findings in Probabilistic Problems from Italy and Spain. *EURASIA Journal of Mathematics, Science and Technology Education* 16, 2. <https://doi.org/10.29333/ejmste/113111>
4. Julie Alonzo, Deni Basaraba, Gerald Tindal, and Ronald S. Carriveau. 2009. They Read, but How Well Do They Understand?: An Empirical Look at the Nuances of Measuring Reading Comprehension. *Assessment for Effective Intervention* 35, 1: 34–44. <https://doi.org/10.1177/1534508408330082>
5. Hissah Alotaibi and Ronal Singh. 2025. Explain to Whom? From User Feedback to Actionable Explanations. *International Journal of Human-Computer Interaction*: 1–13. <https://doi.org/10.1080/10447318.2025.2510824>
6. Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3290605.3300233>
7. Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine Bias *. In *Ethics of Data and Analytics*. Auerbach Publications.
8. Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21). <https://doi.org/10.1145/3411764.3445736>
9. Ariful Islam Anik and Andrea Bunt. 2024. Supporting User Critiques of AI Systems via Training Dataset Explanations: Investigating Critique Properties and the Impact of

- Presentation Style. In *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. <https://doi.org/10.1109/VL/HCC60511.2024.00024>
10. Ariful Islam Anik and Andrea Bunt. 2024. Diversity Challenges in Recruiting for Human-Centered Explainable AI Studies. In *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 429–431. <https://doi.org/10.1109/VL/HCC60511.2024.00067>
11. Ariful Islam Anik and Andrea Bunt. 2026. Designing Effective Training Dataset Explanations: The Impact of Information Depth and Progressive Disclosure. In *Proceedings of the 31st International Conference on Intelligent User Interfaces (IUI '26)*. Association for Computing Machinery, New York, NY, USA, 426–441. <https://doi.org/10.1145/3742413.3789087>
12. Md Ariful Islam Anik. 2020. Data-centric explanations: Explaining training data of machine learning systems to promote transparency. Retrieved October 24, 2025 from <http://hdl.handle.net/1993/35244>
13. Liliana Ardissono, Anna Goy, Giovanna Petrone, Marino Segnan, and Pietro Torasso. 2003. Intrigue: Personalized recommendation of tourist attractions for desktop and hand held devices. *Applied Artificial Intelligence* 17, 8–9: 687–714. <https://doi.org/10.1080/713827254>
14. Hilary Arksey and Lisa O'Malley. 2005. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* 8, 1: 19–32. <https://doi.org/10.1080/1364557032000119616>
15. M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. Natesan Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5: 6:1-6:13. <https://doi.org/10.1147/JRD.2019.2942288>
16. Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2021. AI Explainability 360 Toolkit. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, 376–379. <https://doi.org/10.1145/3430984.3430987>
17. Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, and others. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*.
18. Blair Attard-Frost, Ana Brandusescu, and Kelly Lyons. 2024. The governance of artificial intelligence in Canada: Findings and opportunities from a review of 84 AI governance initiatives. *Government Information Quarterly* 41, 2: 101929. <https://doi.org/10.1016/j.giq.2024.101929>
19. Maria Avgerinou and John Ericson. 1997. A Review of the Concept of Visual Literacy. *British Journal of Educational Technology* 28, 4: 280–291. <https://doi.org/10.1111/1467-8535.00035>
20. Noor A. Aziz, Awais Manzoor, Muhammad Deedahwar Mazhar Qureshi, M. Atif Qureshi, and Wael Rashwan. 2024. Unveiling Explainable AI in Healthcare: Current Trends, Challenges, and Future Directions. <https://doi.org/10.1101/2024.08.10.24311735>

21. Benjamin Bach, Moritz Stefaner, Jeremy Boy, Steven Drucker, Lyn Bartram, Jo Wood, Paolo Ciuccarelli, Yuri Engelhardt, Ulrike Köppen, and Barbara Tversky. 2018. Narrative Design Patterns for Data-Driven Storytelling. In *Data-Driven Storytelling*. A K Peters/CRC Press.
22. Frank B. Baker. 2001. *The Basics of Item Response Theory. Second Edition*. For full text: <http://ericae>. Retrieved May 27, 2025 from <https://eric.ed.gov/?id=ED458219>
23. Albert Bandura. 2006. Guide for constructing self-efficacy scales. *Self-efficacy beliefs of adolescents* 5, 1: 307–337.
24. Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7: 2–11. <https://doi.org/10.1609/hcomp.v7i1.5285>
25. Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. 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* (CHI '21), 1–16. <https://doi.org/10.1145/3411764.3445717>
26. Solon Barocas and Andrew D. Selbst. 2016. Big Data’s Disparate Impact. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2477899>
27. Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58: 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
28. M. S. Bartlett. 1950. TESTS OF SIGNIFICANCE IN FACTOR ANALYSIS. *British Journal of Statistical Psychology* 3, 2: 77–85. <https://doi.org/10.1111/j.2044-8317.1950.tb00285.x>
29. Scott Bateman, Regan L. Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. 2010. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2573–2582. <https://doi.org/10.1145/1753326.1753716>
30. Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.* 6, 1: 20–29. <https://doi.org/10.1145/1007730.1007735>
31. Sarah Bayer, Henner Gimpel, and Moritz Markgraf. 2022. The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems* 32, 1: 110–138. <https://doi.org/10.1080/12460125.2021.1958505>
32. Roman Bednarik and Markku Tukiainen. 2006. An eye-tracking methodology for characterizing program comprehension processes. In *Proceedings of the 2006 symposium on Eye tracking research & applications* (ETRA '06), 125–132. <https://doi.org/10.1145/1117309.1117356>
33. Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6: 587–604. https://doi.org/10.1162/tacl_a_00041

34. Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 57, 1: 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
35. Mostafa Ali Benzaghta, Abdulaziz Elwalda, Mousa Mousa, Ismail Erkan, and Mushfiquir Rahman. 2021. SWOT analysis applications: An integrative literature review. *Journal of Global Business Insights* 6, 1: 55–73. <https://doi.org/10.5038/2640-6489.6.1.1148>
36. Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21), 1–13. <https://doi.org/10.1145/3411764.3445365>
37. Aditya Bhattacharya, Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. 2023. 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* (IUI '23), 204–219. <https://doi.org/10.1145/3581641.3584075>
38. Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. “It’s Reducing a Human Being to a Percentage”: Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3173574.3173951>
39. Godfred O. Boateng, Torsten B. Neilands, Edward A. Frongillo, Hugo R. Melgar-Quinonez, and Sera L. Young. 2018. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health* 6: 149. <https://doi.org/10.3389/fpubh.2018.00149>
40. Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (NIPS'16), 4356–4364.
41. Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, Harris Cooper, Paul M. Camic, Debra L. Long, A. T. Panter, David Rindskopf and Kenneth J. Sher (eds.). American Psychological Association, Washington, 57–71. <https://doi.org/10.1037/13620-004>
42. Sarah Brayne. 2020. *Predict and Surveil: Data, Discretion, and the Future of Policing*. Oxford University Press.
43. Sarah Brayne and Angèle Christin. 2021. Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems* 68, 3: 608–624. <https://doi.org/10.1093/socpro/spaa004>
44. Timothy A. Brown. 2015. *Confirmatory factor analysis for applied research*. The Guilford Press, New York; London.
45. Jerome S. Bruner (ed.). 2011. *Actual minds, possible worlds*. Harvard University Press, Cambridge, Mass.
46. Zana Bućinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings*

- of the 25th International Conference on Intelligent User Interfaces (IUI '20), 454–464. <https://doi.org/10.1145/3377325.3377498>
47. Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1: 188:1-188:21. <https://doi.org/10.1145/3449287>
 48. Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important?: a study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, 169–178. <https://doi.org/10.1145/2166966.2166996>
 49. Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3411764.3445372>
 50. Adrian Bussone, Simone Stumpf, and Dympna O’Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*, 160–169. <https://doi.org/10.1109/ICHI.2015.26>
 51. William Buxton. 2007. *Sketching user experiences: getting the design right and the right design*. Elsevier/Morgan Kaufmann, Amsterdam Boston.
 52. Barbara M. Byrne, Richard J. Shavelson, and Bengt Muthén. 1989. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin* 105, 3: 456–466. <https://doi.org/10.1037/0033-2909.105.3.456>
 53. John T. Cacioppo and Richard E. Petty. 1982. The need for cognition. *Journal of Personality and Social Psychology* 42, 1: 116–131. <https://doi.org/10.1037/0022-3514.42.1.116>
 54. Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 258–262. <https://doi.org/10.1145/3301275.3302289>
 55. Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. “Hello AI”: Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW: 104:1-104:24. <https://doi.org/10.1145/3359206>
 56. Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 981–992. <https://doi.org/10.1145/2858036.2858498>
 57. Toon Calders and Indrė Žliobaitė. 2013. Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures. In *Discrimination and Privacy in the Information Society*, Bart Custers, Toon Calders, Bart Schermer and Tal Zarsky (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 43–57. https://doi.org/10.1007/978-3-642-30487-3_3
 58. Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. <https://doi.org/10.1145/3290605.3300733>

59. Giuseppe Carenini, Cristina Conati, Enamul Hoque, Ben Steichen, Dereck Toker, and James Enns. 2014. Highlighting interventions and user differences: informing adaptive information visualization support. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1835–1844. <https://doi.org/10.1145/2556288.2557141>
60. Astrid Carolus, Martin J. Koch, Samantha Straka, Marc Erich Latoschik, and Carolin Wienrich. 2023. MAILS - Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans* 1, 2: 100014. <https://doi.org/10.1016/j.chbah.2023.100014>
61. Serena Carpenter. 2018. Ten Steps in Scale Development and Reporting: A Guide for Researchers. *Communication Methods and Measures* 12, 1: 25–44. <https://doi.org/10.1080/19312458.2017.1396583>
62. John M. Carroll and Caroline Carrithers. 1984. Training wheels in a user interface. *Communications of the ACM* 27, 8: 800–806. <https://doi.org/10.1145/358198.358218>
63. Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. <https://doi.org/10.1145/2783258.2788613>
64. Raymond B. Cattell. 1966. The Scree Test For The Number Of Factors. *Multivariate Behavioral Research* 1, 2: 245–276. https://doi.org/10.1207/s15327906mbr0102_10
65. Celalettin Çelebi, Fatih Yilmaz, Uğur DemiR, and Ferhat Karakuş. 2023. Artificial Intelligence, AI literacy, Digital literacy, AI literacy scale. *Öğretim Teknolojisi ve Hayat Boyu Öğrenme Dergisi - Instructional Technology and Lifelong Learning*. <https://doi.org/10.52911/itall.1401740>
66. Dilek Cetindamar, Kirsty Kitto, Mengjia Wu, Yi Zhang, Babak Abedin, and Simon Knight. 2024. Explicating AI Literacy of Employees at Digital Workplaces. *IEEE Transactions on Engineering Management* 71: 810–823. <https://doi.org/10.1109/TEM.2021.3138503>
67. R. Philip Chalmers. 2012. **mirt**: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software* 48, 6. <https://doi.org/10.18637/jss.v048.i06>
68. Cecilia Ka Yuk Chan and Wenxin Zhou. 2023. An expectancy value theory (EVT) based instrument for measuring student perceptions of generative AI. *Smart Learning Environments* 10, 1: 64. <https://doi.org/10.1186/s40561-023-00284-4>
69. Mohamed Amine Chatti, Mouadh Guesmi, Laura Vorgerd, Thao Ngo, Shoeb Joarder, Qurat Ul Ain, and Arham Muslim. 2022. Is More Always Better? The Effects of Personal Characteristics and Level of Detail on the Perception of Explanations in a Recommender System. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22)*, 254–264. <https://doi.org/10.1145/3503252.3531304>
70. Li Chen and Feng Wang. 2014. Sentiment-enhanced explanation of product recommendations. In *Proceedings of the 23rd International Conference on World Wide Web*, 239–240. <https://doi.org/10.1145/2567948.2577276>
71. Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. 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. <https://doi.org/10.1145/3610219>

72. Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O’Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI ’19), 1–12. <https://doi.org/10.1145/3290605.3300789>
73. Gordon W. Cheung and Roger B. Rensvold. 2002. Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal* 9, 2: 233–255. https://doi.org/10.1207/S15328007SEM0902_5
74. Thomas K.F. Chiu, Zubair Ahmad, Murod Ismailov, and Ismaila Temitayo Sanusi. 2024. What are artificial intelligence literacy and competency? A comprehensive framework to support them. *Computers and Education Open* 6: 100171. <https://doi.org/10.1016/j.caeo.2024.100171>
75. Thomas K.F. Chiu, Yifan Chen, King Woon Yau, Ching-sing Chai, Helen Meng, Irwin King, Savio Wong, and Yeung Yam. 2024. Developing and validating measures for AI literacy tests: From self-reported to objective measures. *Computers and Education: Artificial Intelligence* 7: 100282. <https://doi.org/10.1016/j.caeai.2024.100282>
76. Kasia S. Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2022. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. <https://doi.org/10.48550/arXiv.2201.03954>
77. Michael Chromik and Andreas Butz. 2021. Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces. In *Human-Computer Interaction – INTERACT 2021*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda and Kori Inkpen (eds.). Springer International Publishing, Cham, 619–640. https://doi.org/10.1007/978-3-030-85616-8_36
78. Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. 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), 307–317. <https://doi.org/10.1145/3397481.3450644>
79. Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems. In *IUI workshops*.
80. Michael Chromik and M. Schuessler. 2020. A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. In *ExSS-ATEC@IUI*. Retrieved from <https://api.semanticscholar.org/CorpusID:214730454>
81. Danielle Keats Citron and Frank Pasquale. 2014. The Scored Society: Due Process for Automated Predictions. Retrieved October 23, 2025 from <https://papers.ssrn.com/abstract=2376209>
82. Anna B. Costello and Jason Osborne. 2005. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. <https://doi.org/10.7275/JYJ1-4868>
83. Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5: 455–496. <https://doi.org/10.1007/s11257-008-9051-3>

84. Jennifer G. Cromley and Roger Azevedo. 2007. Self-report of reading comprehension strategies: What are we measuring? *Metacognition and Learning* 1, 3: 229–247. <https://doi.org/10.1007/s11409-006-9002-5>
85. Lee J. Cronbach. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16, 3: 297–334. <https://doi.org/10.1007/BF02310555>
86. Pat Croskerry, Geeta Singhal, and Silvia Mamede. 2013. Cognitive debiasing 2: impediments to and strategies for change. *BMJ quality & safety* 22 Suppl 2, Suppl 2: ii65–ii72. <https://doi.org/10.1136/bmjqs-2012-001713>
87. Michael F. Dahlstrom. 2014. Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences* 111, supplement_4: 13614–13620. <https://doi.org/10.1073/pnas.1320645111>
88. Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, 598–617. <https://doi.org/10.1109/SP.2016.42>
89. Rafael J. De Ayala. 2022. *The theory and practice of item response theory*. The Guilford Press, New York London.
90. Julien Delaunay, Luis Galárraga, Christine Largouet, and Niels Van Berkel. 2025. Impact of Explanation Techniques and Representations on Users’ Comprehension and Confidence in Explainable AI. *Proceedings of the ACM on Human-Computer Interaction* 9, 2: 1–28. <https://doi.org/10.1145/3711011>
91. Advait Deshpande and Helen Sharp. 2022. Responsible AI Systems: Who are the Stakeholders? In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 227–236. <https://doi.org/10.1145/3514094.3534187>
92. Robert F. DeVellis and Carolyn T. Thorpe. 2022. *Scale development: theory and applications*. SAGE Publications, Inc, Thousand Oaks, California.
93. Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Communications of the ACM* 59, 2: 56–62. <https://doi.org/10.1145/2844110>
94. Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1: 114–126. <https://doi.org/10.1037/xge0000033>
95. Catherine D’Ignazio. 2022. Creative data literacy: Bridging the gap between the data-haves and data-have nots. *Information Design Journal*: 6–18. <https://doi.org/10.1075/idj.23.1.03dig>
96. Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI ’19)*, 275–285. <https://doi.org/10.1145/3301275.3302310>
97. Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. <https://doi.org/10.48550/arXiv.1702.08608>
98. Julie S. Downs. 2014. Prescriptive scientific narratives for communicating usable science. *Proceedings of the National Academy of Sciences* 111, supplement_4: 13627–13633. <https://doi.org/10.1073/pnas.1317502111>

99. Nancy Duarte. 2010. *Resonate: present visual stories that transform audiences*. Wiley, Hoboken, NJ.
100. David Dunning, Chip Heath, and Jerry M. Suls. 2004. Flawed Self-Assessment: Implications for Health, Education, and the Workplace. *Psychological Science in the Public Interest* 5, 3: 69–106. <https://doi.org/10.1111/j.1529-1006.2004.00018.x>
101. Anastasia Efklides. 2006. Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review* 1, 1: 3–14. <https://doi.org/10.1016/j.edurev.2005.11.001>
102. Joyce Ehrlinger and David Dunning. 2003. How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology* 84, 1: 5–17.
103. Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21), 1–19. <https://doi.org/10.1145/3411764.3445188>
104. Upol Ehsan, Q. Vera Liao, Samir Passi, Mark O. Riedl, and Hal Daumé. 2024. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1: 1–29. <https://doi.org/10.1145/3637396>
105. Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24), 1–32. <https://doi.org/10.1145/3613904.3642474>
106. Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, 449–466. https://doi.org/10.1007/978-3-030-60117-1_33
107. Upol Ehsan and Mark O. Riedl. 2024. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns* 5, 6. <https://doi.org/10.1016/j.patter.2024.100971>
108. Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O. Riedl. 2023. 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. <https://doi.org/10.1145/3579467>
109. Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI EA '19), 1–6. <https://doi.org/10.1145/3290607.3312787>
110. Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces*, 211–223. <https://doi.org/10.1145/3172944.3172961>
111. John W. Ely, Mark L. Graber, and Pat Croskerry. 2011. Checklists to Reduce Diagnostic Errors. *Academic Medicine* 86, 3: 307. <https://doi.org/10.1097/ACM.0b013e31820824cd>
112. Susan E. Embretson and Steven P. Reise. 2013. *Item Response Theory*. Psychology Press. <https://doi.org/10.4324/9781410605269>

113. Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags. 2022. What People Think AI Should Infer From Faces. In *2022 ACM Conference on Fairness Accountability and Transparency*, 128–141. <https://doi.org/10.1145/3531146.3533080>
114. Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3290605.3300724>
115. Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639: 115–118. <https://doi.org/10.1038/nature21056>
116. Leandre R. Fabrigar, Duane T. Wegener, Robert C. MacCallum, and Erin J. Strahan. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods* 4, 3: 272–299. <https://doi.org/10.1037/1082-989X.4.3.272>
117. Daniela Fernandes, Steeven Villa, Salla Nicholls, Otso Haavisto, Daniel Buschek, Albrecht Schmidt, Thomas Kosch, Chenxinran Shen, and Robin Welsch. 2026. AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior* 175: 108779. <https://doi.org/10.1016/j.chb.2025.108779>
118. Ana Figueiras. 2014. How to Tell Stories Using Visualization. In *2014 18th International Conference on Information Visualisation*, 18–18. <https://doi.org/10.1109/IV.2014.78>
119. W. Holmes Finch and Brian F. French. 2015. *Latent Variable Modeling with R*. Routledge. <https://doi.org/10.4324/9781315869797>
120. Mikkel Flyverbom. 2016. Transparency: Mediation and the management of visibilities. *International Journal of Communication* 10, 1: 110–122.
121. Claes Fornell and David F. Larcker. 1981. Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research* 18, 1: 39–50. <https://doi.org/10.1177/002224378101800104>
122. Robert L. Fowler and Anne S. Barker. 1974. Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology* 59, 3: 358–364. <https://doi.org/10.1037/h0036750>
123. Colleen C Frank, Gary R Mottola, Meiru Chen, Lei Yu, Patricia A Boyle, Gregory R Samanez-Larkin, and Kendra L Seaman. 2025. Overconfidence and Financial Risk Tolerance in Older Age. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences* 80, 6: gbaf032. <https://doi.org/10.1093/geronb/gbaf032>
124. Krzysztof Z. Gajos, Mary Czerwinski, Desney S. Tan, and Daniel S. Weld. 2006. Exploring the design space for adaptive graphical user interfaces. In *Proceedings of the working conference on Advanced visual interfaces - AVI '06*, 201. <https://doi.org/10.1145/1133265.1133306>
125. Siva Ganesh and Vanessa Cave. 2018. P-values, p-values everywhere! *New Zealand Veterinary Journal* 66, 2: 55–56. <https://doi.org/10.1080/00480169.2018.1415604>
126. Luciana S. Garbayo, David M. Harris, Stephen M. Fiore, Matthew Robinson, and Jonathan D. Kibble. 2023. A metacognitive confidence calibration (MCC) tool to help medical students scaffold diagnostic reasoning in decision-making during high-fidelity patient simulations. *Advances in Physiology Education* 47, 1: 71–81. <https://doi.org/10.1152/advan.00156.2021>

127. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12: 86–92. <https://doi.org/10.1145/3458723>
128. Gilles E. Gignac and Marcin Zajenkowski. 2020. The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence* 80: 101449. <https://doi.org/10.1016/j.intell.2020.101449>
129. Navita Goyal, Eleftheria Briakou, Amanda Liu, Connor Baumler, Claire Bonial, Jeffrey Micher, Clare Voss, Marine Carpuat, and Hal Daumé Iii. 2023. What Else Do I Need to Know? The Effect of Background Information on Users’ Reliance on QA Systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3313–3330. <https://doi.org/10.18653/v1/2023.emnlp-main.201>
130. Melanie C. Green. 2006. Narratives and Cancer Communication. *Journal of Communication* 56, suppl_1: S163–S183. <https://doi.org/10.1111/j.1460-2466.2006.00288.x>
131. Melanie C. Green, Jeffrey J. Strange, and Timothy C. Brock (eds.). 2003. *Narrative Impact: Social and Cognitive Foundations*. Psychology Press. <https://doi.org/10.4324/9781410606648>
132. David Gunning and David Aha. 2019. DARPA’s explainable artificial intelligence (XAI) program. *AI magazine* 40, 2: 44–58. <https://doi.org/10.1609/aimag.v40i2.2850>
133. Minhi Hahn, Robert Lawson, and Young Gyu Lee. 1992. The effects of time pressure and information load on decision quality. *Psychology & Marketing* 9, 5: 365–378. <https://doi.org/10.1002/mar.4220090503>
134. Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson. 2019. *Multivariate data analysis*. Cengage, Andover, Hampshire.
135. Thomas M. Haladyna and Michael C. Rodriguez. 2013. *Developing and Validating Test Items*. Routledge. <https://doi.org/10.4324/9780203850381>
136. Negar Hariri, Bamshad Mobasher, and Robin Burke. 2012. Context-aware music recommendation based on latenttopic sequential patterns. In *Proceedings of the sixth ACM conference on Recommender systems*, 131–138. <https://doi.org/10.1145/2365952.2365979>
137. Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Elsevier, 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
138. Lisa Hartling, Shannon Scott, Rena Pandya, David Johnson, Ted Bishop, and Terry P Klassen. 2010. Storytelling as a communication tool for health consumers: development of an intervention for parents of children with croup. Stories to communicate health information. *BMC Pediatrics* 10, 1: 64. <https://doi.org/10.1186/1471-2431-10-64>
139. James C. Hayton, David G. Allen, and Vida Scarpello. 2004. Factor Retention Decisions in Exploratory Factor Analysis: a Tutorial on Parallel Analysis. *Organizational Research Methods* 7, 2: 191–205. <https://doi.org/10.1177/1094428104263675>
140. Marilyn M. Helms and Judy Nixon. 2010. Exploring SWOT analysis – where are we now?: A review of academic research from the last decade. *Journal of Strategy and Management* 3, 3: 215–251. <https://doi.org/10.1108/17554251011064837>

141. Jörg Henseler, Christian M. Ringle, and Marko Sarstedt. 2015. A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science* 43, 1: 115–135. <https://doi.org/10.1007/s11747-014-0403-8>
142. Erik Hermann. 2022. Artificial intelligence and mass personalization of communication content—An ethical and literacy perspective. *New Media & Society* 24, 5: 1258–1277. <https://doi.org/10.1177/14614448211022702>
143. Timothy R. Hinkin. 1998. A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires. *Organizational Research Methods* 1, 1: 104–121. <https://doi.org/10.1177/109442819800100106>
144. Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5: 1096257. <https://doi.org/10.3389/fcomp.2023.1096257>
145. Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. <https://doi.org/10.48550/arXiv.1805.03677>
146. Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghui Cheng. 2023. Toward Supporting Perceptual Complementarity in Human-AI Collaboration via Reflection on Unobservables. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1: 1–20. <https://doi.org/10.1145/3579628>
147. Marie Hornberger, Arne Bewersdorff, and Claudia Nerdel. 2023. What do university students know about Artificial Intelligence? Development and validation of an AI literacy test. *Computers and Education: Artificial Intelligence* 5: 100165. <https://doi.org/10.1016/j.caeai.2023.100165>
148. Marie Hornberger, Arne Bewersdorff, Daniel S. Schiff, and Claudia Nerdel. 2025. Development and validation of a short AI literacy test (AILIT-S) for university students. *Computers in Human Behavior: Artificial Humans* 5: 100176. <https://doi.org/10.1016/j.chbah.2025.100176>
149. Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research* 15, 9: 1277–1288. <https://doi.org/10.1177/1049732305276687>
150. Li-tze Hu and Peter M. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6, 1: 1–55. <https://doi.org/10.1080/10705519909540118>
151. Mo Hu, Guanglu Zhang, Leah Chong, Jonathan Cagan, and Kosa Goucher-Lambert. 2024. How Being Outvoted by AI Teammates Impacts Human-AI Collaboration. *International Journal of Human-Computer Interaction*: 1–18. <https://doi.org/10.1080/10447318.2024.2345980>
152. Cynthia Hunt, Spencer Yoder, Taylor Comment, Thomas Price, Bitu Akram, Lina Battestilli, Tiffany Barnes, and Susan Fisk. 2022. Gender, Self-Assessment, and Persistence in Computing: How gender differences in self-assessed ability reduce women’s persistence in computer science. In *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 1*, 73–83. <https://doi.org/10.1145/3501385.3543963>

153. Ben Hutchinson, Andrew Smart, Alex Hanna, Remi Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 560–575. <https://doi.org/10.1145/3442188.3445918>
154. Ha Sung Hwang, Liu Cun Zhu, and Qin Cui. 2023. Development and Validation of a Digital Literacy Scale in the Artificial Intelligence Era for College Students. *KSII Transactions on Internet and Information Systems* 17, 8: 2241–2258. <https://doi.org/10.3837/tiis.2023.08.016>
155. Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, Libuše Hannah Vepřek, and Gabrielle Quinn. 2023. Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making. *ACM Transactions on Computer-Human Interaction* 30, 5: 1–29. <https://doi.org/10.1145/3534561>
156. AI Now Institute. 2019. Discriminating Systems: Gender, Race, and Power in AI - Report. *AI Now Institute*. Retrieved July 7, 2025 from <https://ainowinstitute.org/publications/discriminating-systems-gender-race-and-power-in-ai-2>
157. Simon A. Jackson and Sabina Kleitman. 2014. Individual differences in decision-making and confidence: capturing decision tendencies in a fictitious medical test. *Metacognition and Learning* 9, 1: 25–49. <https://doi.org/10.1007/s11409-013-9110-y>
158. Daniel Kahneman. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, NY, US.
159. Henry F. Kaiser. 1960. The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement* 20, 1: 141–151. <https://doi.org/10.1177/001316446002000116>
160. Henry F. Kaiser. 1974. An Index of Factorial Simplicity. *Psychometrika* 39, 1: 31–36. <https://doi.org/10.1007/BF02291575>
161. Martin Kandlhofer and Gerald Steinbauer. 2018. A Driving License for Intelligent Systems. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1. <https://doi.org/10.1609/aaai.v32i1.11399>
162. Maurits Kaptein, Panos Markopoulos, Boris De Ruyter, and Emile Aarts. 2015. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies* 77: 38–51. <https://doi.org/10.1016/j.ijhcs.2015.01.004>
163. Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. 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 (CHI ’20)*, 1–14. <https://doi.org/10.1145/3313831.3376219>
164. Aisling Kelliher and Malcolm Slaney. 2012. Tell Me a Story. *IEEE Multimedia* 19, 1: 4–4. <https://doi.org/10.1109/MMUL.2012.13>
165. Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. “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*, 1–17. <https://doi.org/10.1145/3544548.3581001>
166. Scott King. 2023. The Union - How to Tell When a Product Is Truly Powered By AI. *Krista AI*. Retrieved June 26, 2025 from <https://www.krista.ai/how-to-tell-when-a-product-is-truly-powered-by-ai/>
167. René F. Kizilcec. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 2390–2395. <https://doi.org/10.1145/2858036.2858402>
168. Joshua Klayman. 1995. Varieties of Confirmation Bias. In *Psychology of Learning and Motivation*, Jerome Busemeyer, Reid Hastie and Douglas L. Medin (eds.). Academic Press, 385–418. [https://doi.org/10.1016/S0079-7421\(08\)60315-1](https://doi.org/10.1016/S0079-7421(08)60315-1)
169. Einat Kleider-Tesler, Mirit Barzillai, and Tami Katzir. 2022. Confidence and Calibration of Comprehension in Adolescence: Are They Domain-General or Domain-Specific? *Psychology* 13, 12: 1727–1747. <https://doi.org/10.4236/psych.2022.1312108>
170. Sabina Kleitman and Lazar Stankov. 2007. Self-confidence and metacognitive processes. *Learning and Individual Differences* 17, 2: 161–173. <https://doi.org/10.1016/j.lindif.2007.03.004>
171. Rex B. Kline. 2016. *Principles and practice of structural equation modeling*. The Guilford Press, New York London.
172. Martin J. Koch, Astrid Carolus, Carolin Wienrich, and Marc E. Latoschik. 2024. Meta AI literacy scale: Further validation and development of a short version. *Heliyon* 10, 21: e39686. <https://doi.org/10.1016/j.heliyon.2024.e39686>
173. Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. 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* (CHI '19), 1–14. <https://doi.org/10.1145/3290605.3300641>
174. Maximilian A. Kohl, Kevin Baum, Markus Langer, Daniel Oster, Timo Speith, and Dimitri Bohlender. 2019. Explainability as a Non-Functional Requirement. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 363–368. <https://doi.org/10.1109/RE.2019.00046>
175. Joseph A. Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction* 22, 1–2: 101–123. <https://doi.org/10.1007/s11257-011-9112-x>
176. Asher Koriati. 2012. The self-consistency model of subjective confidence. *Psychological Review* 119, 1: 80–113. <https://doi.org/10.1037/a0025648>
177. Joachim Krueger and Ross A. Mueller. 2002. Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology* 82, 2: 180–188. <https://doi.org/10.1037/0022-3514.82.2.180>
178. Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77, 6: 1121–1134. <https://doi.org/10.1037/0022-3514.77.6.1121>

179. Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 126–137. <https://doi.org/10.1145/2678025.2701399>
180. Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*, 1–10. <https://doi.org/10.1145/2207676.2207678>
181. Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, 3–10. <https://doi.org/10.1109/VLHCC.2013.6645235>
182. Nathan R. Kuncel, Deniz S. Ones, and David M. Klieger. 2014. In Hiring, Algorithms Beat Instinct. *Harvard Business Review*. Retrieved August 26, 2024 from <https://hbr.org/2014/05/in-hiring-algorithms-beat-instinct>
183. Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. <https://doi.org/10.1145/3290605.3300717>
184. Vivian Lai, Yiming Zhang, Chacha Chen, Q. Vera Liao, and Chenhao Tan. 2023. Selective Explanations: Leveraging Human Input to Align Explainable AI. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2: 357:1–357:35. <https://doi.org/10.1145/3610206>
185. Kathryn Ann Lambe, Gary O'Reilly, Brendan D. Kelly, and Sarah Curristan. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ Quality & Safety* 25, 10: 808–820. <https://doi.org/10.1136/bmjqs-2015-004417>
186. Colin Lankshear (ed.). 2008. *Digital literacies: concepts, policies and practices*. Lang, New York Bern Frankfurt am Main Berlin Vienna.
187. Matthias Carl Laupichler, Alexandra Aster, Nicolas Haverkamp, and Tobias Raupach. 2023. Development of the “Scale for the assessment of non-experts’ AI literacy” – An exploratory factor analysis. *Computers in Human Behavior Reports* 12: 100338. <https://doi.org/10.1016/j.chbr.2023.100338>
188. Matthias Carl Laupichler, Alexandra Aster, Jana Schirch, and Tobias Raupach. 2022. Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence* 3: 100101. <https://doi.org/10.1016/j.caeai.2022.100101>
189. Bongshin Lee, Nathalie Henry Riche, Petra Isenberg, and Sheelagh Carpendale. 2015. More Than Telling a Story: Transforming Data into Visually Shared Stories. *IEEE Computer Graphics and Applications* 35, 5: 84–90. <https://doi.org/10.1109/MCG.2015.99>
190. Min Hun Lee, Renee Bao Xuan Ng, Silvana Xinyi Choo, and Shamala Thilarajah. 2024. Interactive Example-Based Explanations to Improve Health Professionals’ Onboarding with AI for Human-AI Collaborative Decision Making. In *Frontiers in Artificial Intelligence and Applications*, Ulle Endriss, Francisco S. Melo, Kerstin Bach, Alberto Bugarín-Diz, José M. Alonso-Moral, Senén Barro and Fredrik Heintz (eds.). IOS Press. <https://doi.org/10.3233/FAIA241044>

191. Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1: 2053951718756684. <https://doi.org/10.1177/2053951718756684>
192. Seyoung Lee and Gain Park. 2024. Development and validation of ChatGPT literacy scale. *Current Psychology* 43, 21: 18992–19004. <https://doi.org/10.1007/s12144-024-05723-0>
193. Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9, 3. <https://doi.org/10.1214/15-AOAS848>
194. Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L. Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. 2022. Author Correction: Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence* 4, 10: 904–904. <https://doi.org/10.1038/s42256-022-00548-7>
195. Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), 1–15. <https://doi.org/10.1145/3313831.3376590>
196. Q. Vera Liao and Kush R. Varshney. 2022. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. Retrieved August 21, 2024 from <http://arxiv.org/abs/2110.10790>
197. Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 10: 147–159. <https://doi.org/10.1609/hcomp.v10i1.21995>
198. Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09), 2119–2128. <https://doi.org/10.1145/1518701.1519023>
199. Tomáš Lintner. 2024. A systematic review of AI literacy scales. *npj Science of Learning* 9, 1: 50. <https://doi.org/10.1038/s41539-024-00264-4>
200. Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3: 31–57. <https://doi.org/10.1145/3236386.3241340>
201. Han Liu, Vivian Lai, and Chenhao Tan. 2021. 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. <https://doi.org/10.1145/3479552>
202. Tania Lombrozo. 2006. The structure and function of explanations. *Trends in Cognitive Sciences* 10, 10: 464–470. <https://doi.org/10.1016/j.tics.2006.08.004>
203. Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive Psychology* 55, 3: 232–257. <https://doi.org/10.1016/j.cogpsych.2006.09.006>
204. Tania Lombrozo. 2016. Explanatory Preferences Shape Learning and Inference. *Trends in Cognitive Sciences* 20, 10: 748–759. <https://doi.org/10.1016/j.tics.2016.08.001>

205. Tania Lombrozo and Emily G. Liquin. 2023. Explanation Is Effective Because It Is Selective. *Current Directions in Psychological Science* 32, 3: 212–219. <https://doi.org/10.1177/09637214231156106>
206. Tania Lombrozo and Nadya Vasilyeva. 2017. *Causal Explanation*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199399550.013.22>
207. Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), 1–16. <https://doi.org/10.1145/3313831.3376727>
208. Joy Lu, Dokyun (Dk) Lee, Tae Wan Kim, and David Danks. 2019. Good Explanation for Algorithmic Transparency. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3503603>
209. Robert C. MacCallum, Shaobo Zhang, Kristopher J. Preacher, and Derek D. Rucker. 2002. On the practice of dichotomization of quantitative variables. *Psychological Methods* 7, 1: 19–40. <https://doi.org/10.1037/1082-989X.7.1.19>
210. Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3313831.3376445>
211. David Magis, Sébastien Béland, Francis Tuerlinckx, and Paul De Boeck. 2010. A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods* 42, 3: 847–862. <https://doi.org/10.3758/BRM.42.3.847>
212. Gideon Mann and Cathy O’Neil. 2016. Hiring algorithms are not neutral. *Harvard Business Review* 9: 2016.
213. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press, Cambridge.
214. Allan Martin and Jan Grudziecki. 2006. DigEuLit: Concepts and Tools for Digital Literacy Development. *Innovation in Teaching and Learning in Information and Computer Sciences* 5, 4: 249–267. <https://doi.org/10.11120/ital.2006.05040249>
215. Konstantina Martzoukou, Crystal Fulton, Petros Kostagiolas, and Charilaos Lavranos. 2020. A study of higher education students’ self-perceived digital competences for learning and everyday life online participation. *Journal of Documentation* 76, 6: 1413–1458. <https://doi.org/10.1108/JD-03-2020-0041>
216. John D. Mayer, A. T. Panter, and David R. Caruso. 2020. When people estimate their personal intelligence who is overconfident? Who is accurate? *Journal of Personality* 88, 6: 1129–1144. <https://doi.org/10.1111/jopy.12561>
217. Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW: 1–23. <https://doi.org/10.1145/3359174>
218. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys* 54, 6: 1–35. <https://doi.org/10.1145/3457607>

219. William Meredith. 1993. Measurement Invariance, Factor Analysis and Factorial Invariance. *Psychometrika* 58, 4: 525–543. <https://doi.org/10.1007/BF02294825>
220. Ines Mergel, Helen Dickinson, Jari Stenvall, and Mila Gasco. 2023. Implementing AI in the public sector. *Public Management Review*: 1–14. <https://doi.org/10.1080/14719037.2023.2231950>
221. Diane C. Mézière, Lili Yu, Erik D. Reichle, Titus von der Malsburg, and Genevieve McArthur. 2023. Using Eye-Tracking Measures to Predict Reading Comprehension. *Reading Research Quarterly* 58, 3: 425–449. <https://doi.org/10.1002/rrq.498>
222. Patrick Mikalef and Manjul Gupta. 2021. Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management* 58, 3: 103434. <https://doi.org/10.1016/j.im.2021.103434>
223. Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267: 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
224. Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. <https://doi.org/10.48550/arXiv.1712.00547>
225. Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3287560.3287596>
226. Lloyd H. Nakatani and John A. Rohrlich. 1983. Soft machines: A philosophy of user-computer interface design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '83)*, 19–23. <https://doi.org/10.1145/800045.801573>
227. Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys* 55, 13s: 1–42. <https://doi.org/10.1145/3583558>
228. Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence* 2: 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
229. Davy Tsz Kit Ng, Wenjie Wu, Jac Ka Lok Leung, Thomas Kin Fung Chiu, and Samuel Kai Wah Chu. 2024. Design and validation of the AI literacy questionnaire: The affective, behavioural, cognitive and ethical approach. *British Journal of Educational Technology* 55, 3: 1082–1104. <https://doi.org/10.1111/bjet.13411>
230. Raymond S. Nickerson. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2, 2: 175–220. <https://doi.org/10.1037/1089-2680.2.2.175>
231. Robert Nimmo, Marios Constantinides, Ke Zhou, Daniele Quercia, and Simone Stumpf. 2024. User Characteristics in Explainable AI: The Rabbit Hole of Personalization? In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3613904.3642352>
232. Donald A. Norman. 2002. *The Design of Everyday Things*. Basic Books, Inc., USA.

233. Geoff Norman. 2010. Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education* 15, 5: 625–632. <https://doi.org/10.1007/s10459-010-9222-y>
234. Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8: 112–121. <https://doi.org/10.1609/hcomp.v8i1.7469>
235. Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrira Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*, 340–350. <https://doi.org/10.1145/3397481.3450639>
236. Conor Nugent and Pádraig Cunningham. 2005. A Case-Based Explanation System for Black-Box Systems. *Artificial Intelligence Review* 24, 2: 163–178. <https://doi.org/10.1007/s10462-005-4609-5>
237. Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3–5: 393–444. <https://doi.org/10.1007/s11257-017-9195-0>
238. J.C. Nunnally. 1978. *Psychometric Theory*. McGraw-Hill, New York, NY, USA.
239. Adegboyega Ojo and Bahareh Heravi. 2018. Patterns in Award Winning Data Storytelling: Story Types, Enabling Tools and Competences. *Digital Journalism* 6, 6: 693–718. <https://doi.org/10.1080/21670811.2017.1403291>
240. Jennifer L. Osterhage, Ellen L. Usher, Trisha A. Douin, and William M. Bailey. 2019. Opportunities for Self-Evaluation Increase Student Calibration in an Introductory Biology Course. *CBE—Life Sciences Education* 18, 2: ar16. <https://doi.org/10.1187/cbe.18-10-0202>
241. Eoin D. O’Sullivan and Susie J. Schofield. 2019. A cognitive forcing tool to mitigate cognitive bias – a randomised control trial. *BMC Medical Education* 19, 1: 12. <https://doi.org/10.1186/s12909-018-1444-3>
242. Fred Paas, Alexander Renkl, and John Sweller. 2003. Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist* 38, 1: 1–4. https://doi.org/10.1207/S15326985EP3801_1
243. Stefan Palan and Christian Schitter. 2018. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17: 22–27. <https://doi.org/10.1016/j.jbef.2017.12.004>
244. Gerry Pallier. 2003. Gender Differences in the Self-Assessment of Accuracy on Cognitive Tasks. *Sex Roles* 48, 5/6: 265–276. <https://doi.org/10.1023/A:1022877405718>
245. Gerry Pallier, Rebecca Wilkinson, Vanessa Danthiir, Sabina Kleitman, Goran Knezevic, Lazar Stankov, and Richard D. Roberts. 2002. The Role of Individual Differences in the Accuracy of Confidence Judgments. *The Journal of General Psychology* 129, 3: 257–299. <https://doi.org/10.1080/00221300209602099>
246. Frank Pasquale. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press. <https://doi.org/10.4159/harvard.9780674736061>
247. Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on AI Literature Review.

248. John W Payne, James R Bettman, and Eric J Johnson. 1988. Adaptive strategy selection in decision making. *Journal of experimental psychology: Learning, Memory, and Cognition* 14, 3: 534–552. <https://doi.org/10.1037/0278-7393.14.3.534>
249. Maria J. P. Peixoto, Akriti Pandey, Ahsan Zaman, and Peter R. Lewis. 2025. Who Benefits from AI Explanations? Towards Accessible and Interpretable Systems. <https://doi.org/10.48550/arXiv.2508.10806>
250. Maxine Pfannkuch, Matt Regan, Chris Wild, and Nicholas J. Horton. 2010. Telling Data Stories: Essential Dialogues for Comparative Reasoning. *Journal of Statistics Education* 18, 1: 9. <https://doi.org/10.1080/10691898.2010.11889479>
251. David Phipps. 2022. Switching from stakeholder. *Research Impact Canada*. Retrieved September 5, 2024 from <https://researchimpact.ca/featured/switching-from-stakeholder/>
252. Marc Pinski and Alexander Benlian. 2023. AI Literacy - Towards Measuring Human Competency in Artificial Intelligence. *Hawaii International Conference on System Sciences 2023 (HICSS-56)*. Retrieved from https://aisel.aisnet.org/hicss-56/cl/ai_and_future_work/3
253. Murat Polat. 2020. Analysis of multiple-choice versus open-ended questions in language tests according to different cognitive domain levels. *Novitas-ROYAL (Research on Youth and Language)* 14, 2: 76–96.
254. Denise F. Polit and Cheryl Tatano Beck. 2006. The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health* 29, 5: 489–497. <https://doi.org/10.1002/nur.20147>
255. Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, 1–52. <https://doi.org/10.1145/3411764.3445315>
256. Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in Explainable AI. <https://doi.org/10.48550/arXiv.1810.00184>
257. Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1776–1826. <https://doi.org/10.1145/3531146.3533231>
258. Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3173574.3173677>
259. Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. <https://doi.org/10.1145/3351095.3372873>
260. Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1: 1–22. <https://doi.org/10.1145/3512930>

261. McKenzie Raub. Bots, Bias and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices. 2018. *ARKANSAS LAW REVIEW* 71.
262. Steven P. Reise, Niels G. Waller, and Andrew L. Comrey. 2000. Factor analysis and scale revision. *Psychological Assessment* 12, 3: 287–297. <https://doi.org/10.1037/1040-3590.12.3.287>
263. Steven Paul Reise and Dennis A. Revicki (eds.). 2015. *Handbook of item response theory modeling: applications to typical performance assessment*. Routledge, New York London. <https://doi.org/10.4324/9781315736013>
264. Raymond Reiter. 1987. A theory of diagnosis from first principles. *Artificial Intelligence* 32, 1: 57–95. [https://doi.org/10.1016/0004-3702\(87\)90062-2](https://doi.org/10.1016/0004-3702(87)90062-2)
265. Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, GI Genius CADx Study Group, Giulio Antonelli, Halim Awadie, Sebastian Bernhofer, Sabela Carballal, Mário Dinis-Ribeiro, Agnès Fernández-Clotett, Glòria Fernández Esparrach, Ian Gralnek, Yuta Higasa, Taku Hirabayashi, Tatsuki Hirai, Mineo Iwatate, Miki Kawano, Markus Mader, Andreas Maieron, Sebastian Mattes, Tastuya Nakai, Ingrid Ordas, Raquel Ortigão, Oswaldo Ortiz Zúñiga, Maria Pellisé, Cláudia Pinto, Florian Riedl, Ariadna Sánchez, Emanuel Steiner, Yukari Tanaka, and Andrea Cherubini. 2022. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports* 12, 1: 14952. <https://doi.org/10.1038/s41598-022-18751-2>
266. Mijke Rhemtulla, Patricia É. Brosseau-Liard, and Victoria Savalei. 2012. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods* 17, 3: 354–373. <https://doi.org/10.1037/a0029315>
267. Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101. <https://doi.org/10.18653/v1/N16-3020>
268. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-Agnostic Interpretability of Machine Learning. <https://doi.org/10.48550/arXiv.1606.05386>
269. M. O. Riedl and R. M. Young. 2010. Narrative Planning: Balancing Plot and Character. *Journal of Artificial Intelligence Research* 39: 217–268. <https://doi.org/10.1613/jair.2989>
270. Vincent Robbemon, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, 223–233. <https://doi.org/10.1145/3503252.3531311>
271. Henry L. Roediger and Elizabeth J. Marsh. 2005. The Positive and Negative Consequences of Multiple-Choice Testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31, 5: 1155–1159. <https://doi.org/10.1037/0278-7393.31.5.1155>
272. Yves Rosseel. 2012. **lavaan**: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2. <https://doi.org/10.18637/jss.v048.i02>
273. Mary Beth Rosson and John M. Carroll. 2012. Scenario-Based Design. In *Human Computer Interaction Handbook* (3rd ed.). CRC Press.
274. Fumiko Samejima. 1969. Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika* 34, S1: 1–97. <https://doi.org/10.1007/BF03372160>

275. Wojciech Samek, Alexander Binder, Gregoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Muller. 2017. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems* 28, 11: 2660–2673. <https://doi.org/10.1109/TNNLS.2016.2599820>
276. James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*, 240–251. <https://doi.org/10.1145/3301275.3302308>
277. Thomas A. Schmitt. 2011. Current Methodological Considerations in Exploratory and Confirmatory Factor Analysis. *Journal of Psychoeducational Assessment* 29, 4: 304–321. <https://doi.org/10.1177/0734282911406653>
278. Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. “There Is Not Enough Information”: On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 1616–1628. <https://doi.org/10.1145/3531146.3533218>
279. Jane Secker and Emma Coonan (eds.). 2012. *Rethinking Information Literacy: A Practical Framework for Supporting Learning*. Facet. <https://doi.org/10.29085/9781856049528>
280. E Segel and J Heer. 2010. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics* 16, 6: 1139–1148. <https://doi.org/10.1109/TVCG.2010.179>
281. Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph “Jofish” Kaye. 2005. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*, 49–58. <https://doi.org/10.1145/1094562.1094569>
282. Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2: 1–22. <https://doi.org/10.1145/3415224>
283. Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users’ perceptions of fairness toward an algorithmic system. *Ethics and Information Technology* 24, 1: 2. <https://doi.org/10.1007/s10676-022-09623-4>
284. Ronal Singh, Tim Miller, Henrietta Lyons, Liz Sonenberg, Eduardo Velloso, Frank Vetere, Piers Howe, and Paul Dourish. 2023. Directive Explanations for Actionable Explainability in Machine Learning Applications. *ACM Transactions on Interactive Intelligent Systems* 13, 4: 1–26. <https://doi.org/10.1145/3579363>
285. Traci Sitzmann, Katherine Ely, Kenneth G. Brown, and Kristina N. Bauer. 2010. Self-Assessment of Knowledge: A Cognitive Learning or Affective Measure? *Academy of Management Learning & Education* 9, 2: 169–191. <https://doi.org/10.5465/amle.9.2.zqr169>
286. Paul Slovic, Melissa L. Finucane, Ellen Peters, and Donald G. MacGregor. 2007. The affect heuristic. *European Journal of Operational Research* 177, 3: 1333–1352. <https://doi.org/10.1016/j.ejor.2005.04.006>

287. Hannah R. Snyder, Naomi P. Friedman, and Benjamin L. Hankin. 2021. Associations Between Task Performance and Self-Report Measures of Cognitive Control: Shared Versus Distinct Abilities. *Assessment* 28, 4: 1080–1096. <https://doi.org/10.1177/1073191120965694>
288. Kacper Sokol and Peter Flach. 2020. One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency. *KI - Künstliche Intelligenz* 34, 2: 235–250. <https://doi.org/10.1007/s13218-020-00637-y>
289. Maarten W. van Someren, Yvonne F. Barnard, and Jacobijn A.C. Sandberg. 1994. *The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes*. Academic Press, Inc.
290. María T. Soto-Sanfiel, Ariadna Angulo-Brunet, and Christoph Lutz. 2024. The Scale of Artificial Intelligence Literacy for all (SAIL4ALL): A Tool for Assessing Knowledge on Artificial Intelligence in All Adult Populations and Settings. <https://doi.org/10.31235/osf.io/bvyku>
291. Frank Spillers. 2004. What is Progressive Disclosure? - ED. *Experience Dynamics*. Retrieved August 21, 2024 from <https://www.experiencedynamics.com/progressive-disclosure-the-best-interaction-design-technique/>
292. Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 107–120. <https://doi.org/10.1145/3301275.3302322>
293. Aaron Springer and Steve Whittaker. 2020. Progressive Disclosure: When, Why, and How Do Users Want Algorithmic Transparency Information? *ACM Transactions on Interactive Intelligent Systems* 10, 4: 1–32. <https://doi.org/10.1145/3374218>
294. Prachi Srivastava and Nick Hopwood. 2009. A Practical Iterative Framework for Qualitative Data Analysis. *International Journal of Qualitative Methods* 8, 1: 76–84. <https://doi.org/10.1177/160940690900800107>
295. Centre for Education Statistics and Evaluation. 2017. Cognitive load theory: Research that teachers really need to understand. *Sydney: Centre for Education Statistics and Evaluation*.
296. Teresa M. Stephens. 2020. Cognitive Debiasing: *Nurse Leader* 18, 4: 344–351. <https://doi.org/10.1016/j.mnl.2019.03.013>
297. Jiao Sun, Q. Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D. Weisz. 2022. Investigating Explainability of Generative AI for Code through Scenario-based Design. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)*, 212–228. <https://doi.org/10.1145/3490099.3511119>
298. Anneli Sundin, Karolin Andersson, and Robert Watt. 2018. Rethinking communication: integrating storytelling for increased stakeholder engagement in environmental evidence synthesis. *Environmental Evidence* 7, 1: 6. <https://doi.org/10.1186/s13750-018-0116-4>
299. John Sweller. 2010. Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educational Psychology Review* 22, 2: 123–138. <https://doi.org/10.1007/s10648-010-9128-5>
300. John Sweller. 2011. Cognitive Load Theory. In *Psychology of Learning and Motivation*. Elsevier, 37–76. <https://doi.org/10.1016/B978-0-12-387691-1.00002-8>

301. John Sweller, Jeroen J. G. Van Merriënboer, and Fred G. W. C. Paas. 1998. Cognitive Architecture and Instructional Design. *Educational Psychology Review* 10, 3: 251–296. <https://doi.org/10.1023/A:1022193728205>
302. Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*, 109–119. <https://doi.org/10.1145/3397481.3450662>
303. Jenny Tang, Eleanor Birrell, and Ada Lerner. 2022. Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys.
304. TemplarAdvisors. 2016. Why storytelling is essential to effective communication. *Templar Advisors*. Retrieved September 17, 2025 from <https://templaradvisors.com/2016/04/28/storytelling-effective-communication/>
305. Matthew S. Thiese, Brenden Ronna, and Ulrike Ott. 2016. P value interpretations and considerations. *Journal of Thoracic Disease* 8, 9: E928–E931. <https://doi.org/10.21037/jtd.2016.08.16>
306. David Thissen, Lynne Steinberg, and Howard Wainer. 1993. Detection of differential item functioning using the parameters of item response models. In *Differential item functioning*. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 67–113. <https://doi.org/10.1075/z.62.13kok>
307. Alice Thudt, Jagoda Walny, Theresia Gschwandtner, Jason Dykes, and John Stasko. 2018. Exploration and Explanation in Data-Driven Storytelling. In *Data-Driven Storytelling*. A K Peters/CRC Press.
308. Nava Tintarev and Judith Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop (ICDEW '07)*, 801–810. <https://doi.org/10.1109/ICDEW.2007.4401070>
309. Joseph K Torgesen. 2007. Research related to strengthening instruction in reading comprehension: Part 2. *Florida State University and National Center for Reading First Technical Assistance. Comprehension Conference*.
310. Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science* 185, 4157: 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
311. Michelle Vaccaro and Jim Waldo. 2019. The effects of mixing machine learning and human judgment. *Communications of the ACM* 62, 11: 104–110. <https://doi.org/10.1145/3359338>
312. Alexander Jam Van Deursen and Jan Agm Van Dijk. 2014. The digital divide shifts to differences in usage. *New Media & Society* 16, 3: 507–526. <https://doi.org/10.1177/1461444813487959>
313. Jeroen J. G. Van Merriënboer and John Sweller. 2005. Cognitive Load Theory and Complex Learning: Recent Developments and Future Directions. *Educational Psychology Review* 17, 2: 147–177. <https://doi.org/10.1007/s10648-005-3951-0>
314. Peter M. VanNostrand, Dennis M. Hofmann, Lei Ma, and Elke A. Rundensteiner. 2024. Actionable Recourse for Automated Decisions: Examining the Effects of Counterfactual Explanation Type and Presentation on Lay User Understanding. In *The 2024 ACM*

- Conference on Fairness, Accountability, and Transparency*, 1682–1700. <https://doi.org/10.1145/3630106.3658997>
315. Marcel V. J. Veenman. 2005. The assessment of metacognitive skills: what can be learned from multi-method designs? Retrieved from <https://api.semanticscholar.org/CorpusID:122725234>
 316. Tiffany C Veinot, Hannah Mitchell, and Jessica S Ancker. 2018. Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association* 25, 8: 1080–1088. <https://doi.org/10.1093/jamia/ocy052>
 317. Xaquín González Veira. 2017. I like data-driven storytelling (I like loaded words). *Xocas*. Retrieved September 17, 2025 from <https://medium.com/xocas/i-like-data-driven-storytelling-i-like-loaded-words-d066d83ee909>
 318. Wayne F. Velicer. 1976. Determining the Number of Components from the Matrix of Partial Correlations. *Psychometrika* 41, 3: 321–327. <https://doi.org/10.1007/BF02293557>
 319. Sahil Verma, Aditya Lahiri, John P. Dickerson, and Su-In Lee. 2021. Pitfalls of Explainable ML: An Industry Perspective. <https://doi.org/10.48550/arXiv.2106.07758>
 320. Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* 76: 89–106. <https://doi.org/10.1016/j.inffus.2021.05.009>
 321. Inc Visa. 2022. Data Experience Critique Framework. Retrieved from <https://developer.visa.com/images2/visa-chart-components/Data%20Experience%20Critique%20Framework.pdf>
 322. Michael Vössing, Niklas Kühl, Matteo Lind, and Gerhard Satzger. 2022. Designing Transparency for Effective Human-AI Collaboration. *Information Systems Frontiers* 24, 3: 877–895. <https://doi.org/10.1007/s10796-022-10284-3>
 323. Bingcheng Wang, Pei-Luen Patrick Rau, and Tianyi Yuan. 2023. Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology* 42, 9: 1324–1337. <https://doi.org/10.1080/0144929X.2022.2072768>
 324. Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 1–15. <https://doi.org/10.1145/3290605.3300831>
 325. Shan Wang, Fang Wang, Zhen Zhu, Jingxuan Wang, Tam Tran, and Zhao Du. 2024. Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications* 252: 124167. <https://doi.org/10.1016/j.eswa.2024.124167>
 326. Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*, 318–328. <https://doi.org/10.1145/3397481.3450650>
 327. Yuanyuan Wang, Stephen Chi-Fai Chan, and Grace Ngai. 2012. Applicability of Demographic Recommender System to Tourist Attractions: A Case Study on Trip Advisor. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 97–101. <https://doi.org/10.1109/WI-IAT.2012.133>

328. Zezhong Wang, Shunming Wang, Matteo Farinella, Dave Murray-Rust, Nathalie Henry Riche, and Benjamin Bach. 2019. Comparing Effectiveness and Engagement of Data Comics and Infographics. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. <https://doi.org/10.1145/3290605.3300483>
329. Patrick Weber, Marc Pinski, and Lorenz Baum. 2023. Toward an Objective Measurement of AI Literacy. *PACIS 2023 Proceedings*. Retrieved from <https://aisel.aisnet.org/pacis2023/60>
330. Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6: 70–79. <https://doi.org/10.1145/3282486>
331. Gordon Willis. 2005. *Cognitive Interviewing*. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America. <https://doi.org/10.4135/9781412983655>
332. Annika Wolff, Daniel Gooch, Jose J. Caverio Montaner, Umar Rashid, and Gerd Kortuem. 2016. Creating an Understanding of Data Literacy for a Data-driven Society. *The Journal of Community Informatics* 12, 3. <https://doi.org/10.15353/joci.v12i3.3275>
333. Jen-Her Wu and Yufei Yuan. 2003. Improving searching and reading performance: the effect of highlighting and text color coding. *Inf. Manage.* 40, 7: 617–637. [https://doi.org/10.1016/S0378-7206\(02\)00091-5](https://doi.org/10.1016/S0378-7206(02)00091-5)
334. Andre Wyatt. 2023. Advantages of SWOT Analysis (6 Benefits and 4 Limitations). *Status.net*. Retrieved September 17, 2025 from <https://status.net/articles/swot-analysis-process-benefits-limitations/>
335. Jr. Wyer (ed.). 2014. *Knowledge and Memory: the Real Story*. Psychology Press. <https://doi.org/10.4324/9781315806594>
336. L. Richard Ye and Paul E. Johnson. 1995. The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice. *MIS Quarterly* 19, 2: 157–172. <https://doi.org/10.2307/249686>
337. Hubert Dariusz Zajac, Natalia Rozalia Avlona, Finn Kensing, Tariq Osman Andersen, and Irina Shklovski. 2023. Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI. In *Proceedings of the 2023 AAI/ACM Conference on AI, Ethics, and Society (AIES '23)*, 351–362. <https://doi.org/10.1145/3600211.3604766>
338. Baobao Zhang and Allan Dafoe. 2019. Artificial Intelligence: American Attitudes and Trends. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3312874>
339. Helen Zhang, Anthony Perry, and Irene Lee. 2025. Developing and Validating the Artificial Intelligence Literacy Concept Inventory: an Instrument to Assess Artificial Intelligence Literacy among Middle School Students. *International Journal of Artificial Intelligence in Education* 35, 1: 398–438. <https://doi.org/10.1007/s40593-024-00398-x>
340. Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. 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*, 295–305. <https://doi.org/10.1145/3351095.3372852>
341. Lichun Zhou. 2020. Product advertising recommendation in e-commerce based on deep learning and distributed expression. *Electronic Commerce Research* 20, 2: 321–342. <https://doi.org/10.1007/s10660-020-09411-6>

- 342. Xinyan Zhou, Yan Li, Ching Sing Chai, and Thomas K. F. Chiu. 2025. Defining, enhancing, and assessing artificial intelligence literacy and competency in K-12 education from a systematic review. *Interactive Learning Environments* 33, 10: 5766–5788. <https://doi.org/10.1080/10494820.2025.2487538>
- 343. David Dunning. 2011. The Dunning–Kruger Effect. In *Advances in Experimental Social Psychology*. Elsevier, 247–296. <https://doi.org/10.1016/b978-0-12-385522-0.00005-6>
- 344. 5 Alternatives To SWOT Analysis Tackling its Weaknesses. 2023. *PESTLE Analysis*. Retrieved December 9, 2025 from <https://pestleanalysis.com/alternatives-to-swot-analysis/>
- 345. PAIR-code/knownyourdata. 2024. Retrieved August 26, 2024 from <https://github.com/PAIR-code/knownyourdata>
- 346. Userback: Your #1 User Feedback Software. Retrieved August 1, 2024 from <https://userback.io/>
- 347. Kaggle: Your Machine Learning and Data Science Community. Retrieved September 17, 2025 from <https://www.kaggle.com/>
- 348. Progressive Disclosure. *Nielsen Norman Group*. Retrieved December 9, 2024 from <https://www.nngroup.com/articles/progressive-disclosure/>

Appendices

Appendix A – Presentation Style Study Documents

A.1. Data Story Presentation for the Training Dataset Explanation

The explanation was presented as a single webpage where participants would scroll through it. The items would appear with animations, and participant could interact with the images and graphs.

Insights into the Data Used in the Hiring Systems

Following industry trends, Company “A” is planning to use an automated system based on machine-learning algorithms to inform their hiring decisions.

To train their system (i.e., to teach the system to properly interpret data and learn from it), they used an internal company dataset. We looked at the documentation provided for the dataset and using that knowledge, we can talk about the important properties of the dataset. We discuss the types of data the dataset contains, the individuals in the data, and how the data was collected, processed, and have been used till now.

The dataset contains data from the resumes of **17562** applicants. Now, what does each piece of data really tell us about the applicants?

For each applicant, the dataset contains information about 15 attributes. This includes attributes like -



Education



Experience



Skills

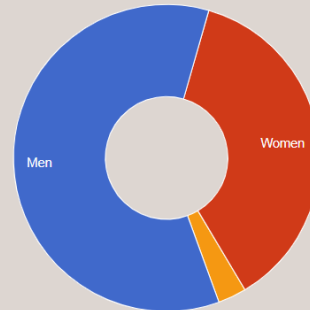


Age



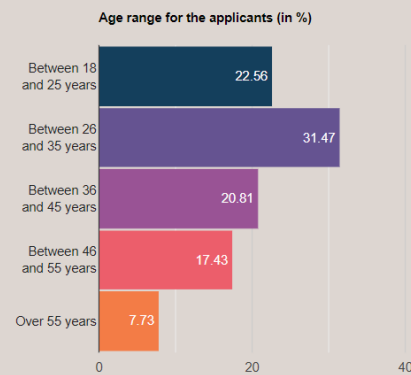
References

Among the applicants, **60% are men**.
The other 40% are women (37%)
and non-binary (3%) people.



Gender of the applicants

In terms of age, most individuals are **26 to 35**. The number of applicants *over 55* years old is the lowest. The other age brackets have similar numbers of applicants.



For the ethnicity of the applicants, the dataset dominantly contains applicants *who identify as White*. For all other ethnicities, there are a few to a moderate number of applicants present in the dataset.

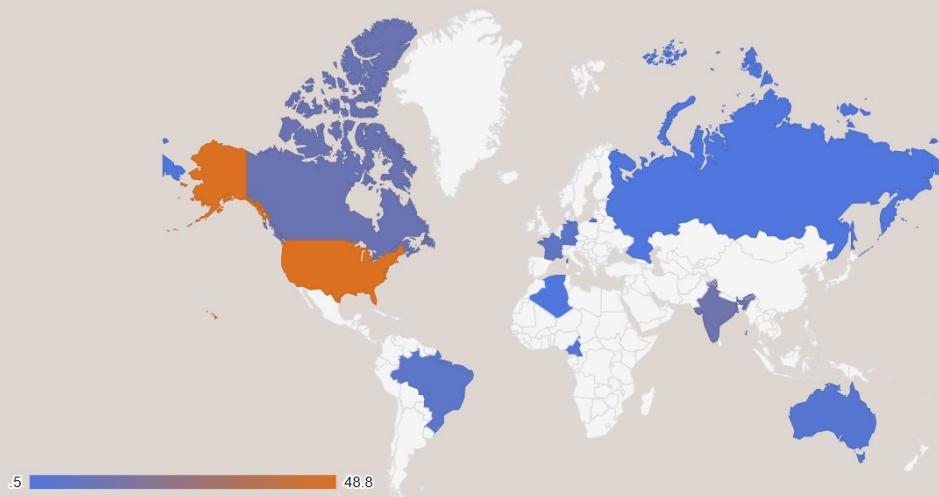
Ethnicity of the population in the dataset

👤 White 👤 Brown 👤 Hispanic 👤 Asian 👤 Black 👤 Others



• A Flourish data visualization

The dataset contains applicants from a range of countries; however, the majority are from the **USA**.



So, we get a sense of the demographics of the applicants that was in the dataset. Let's now move our focus to how the data was collected.

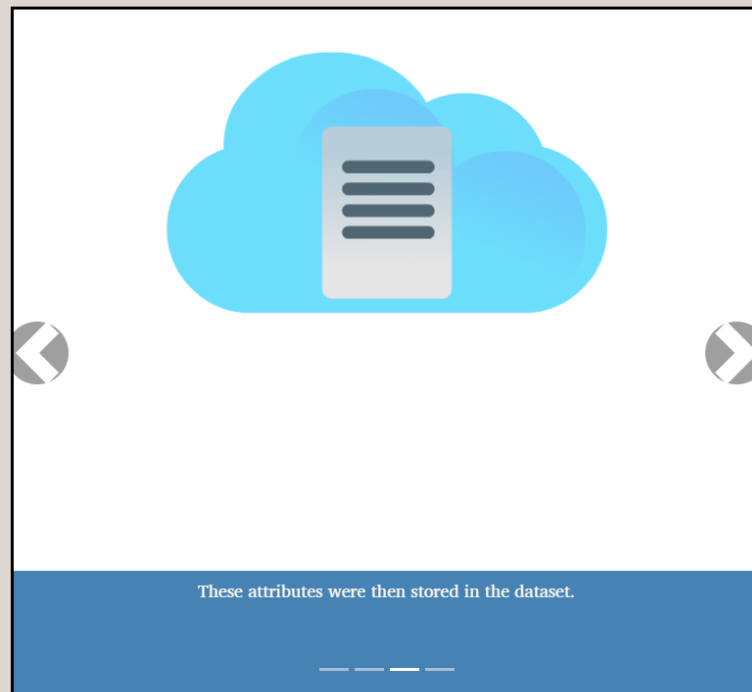
In 2020, Company “A” conducted the data collection using the resumes they received between 2016 and 2020.

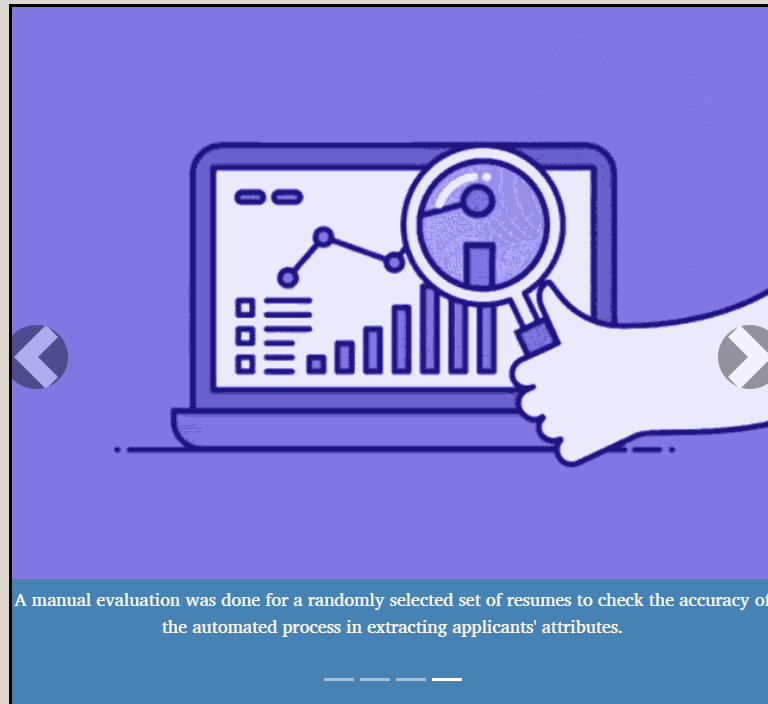
To gather relevant data from the resumes, they followed a sequence of steps. First, they extracted all the relevant attributes about the applicants from the resumes using an automated process (i.e., a computer program). These attributes were then stored in the dataset and all the other information about the applicants in the resumes was discarded.

After using the automated process, the company manually analyzed a randomly selected set of resumes to check the accuracy of the automated process in extracting applicants’ attributes from their resumes. The company was satisfied with the general accuracy of the automation and moved on to the next step of processing the data.

You can get a overview of the collection process in the following pictures.







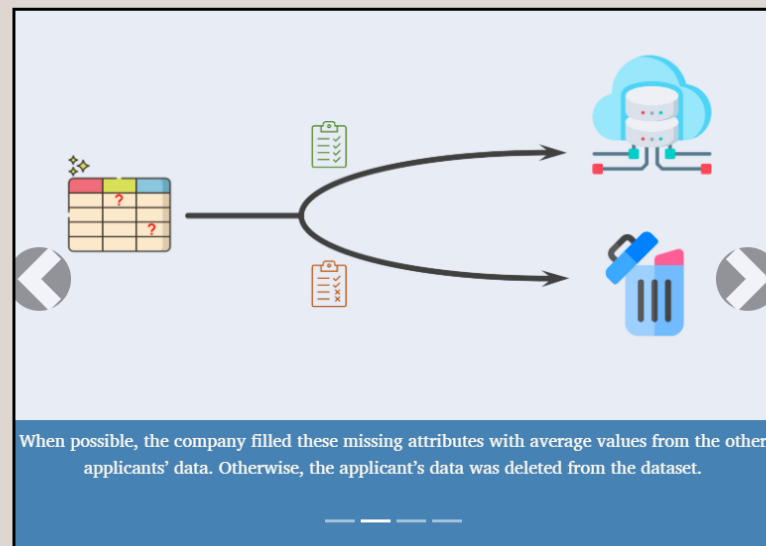
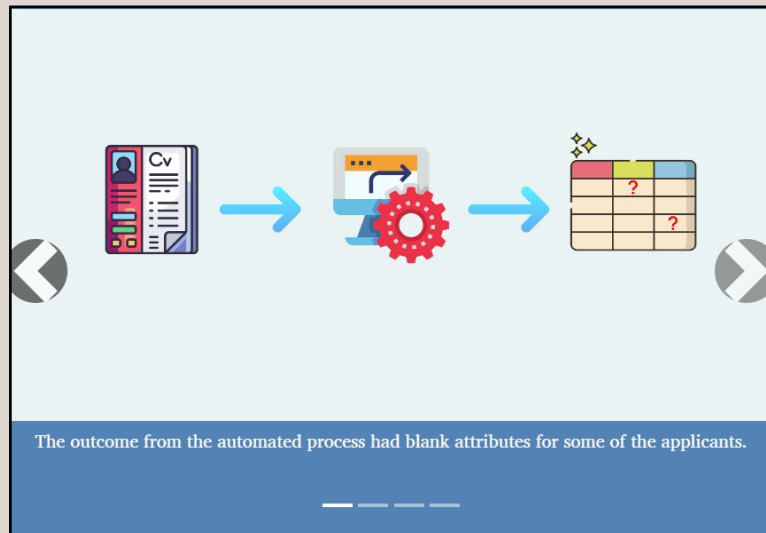
The company went through several processing steps to finalize the dataset.

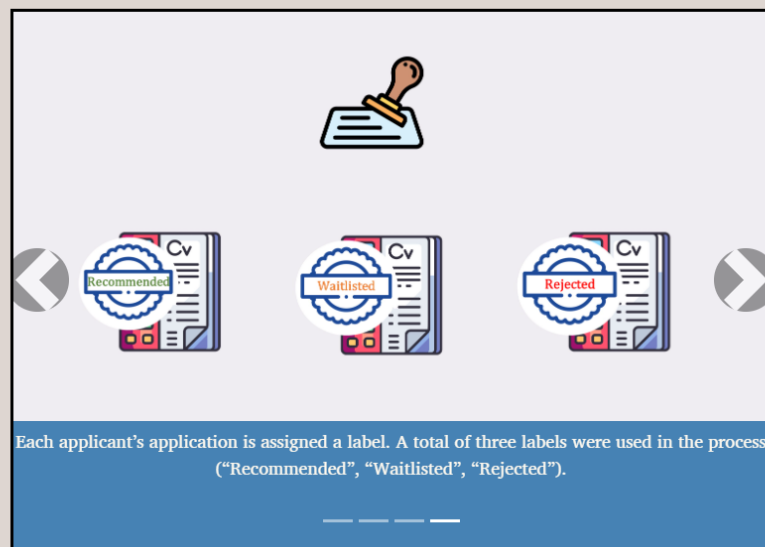
The automated process left some attributes blank for some of the applicants. When possible, the company filled these missing attributes with average values from the other applicants' data. Otherwise, the applicant's data was deleted from the dataset.

For educational qualifications and results, the people involved in the processing transformed the information into a value on a single scale to allow for better comparisons between applicants. They also assigned numerical values to prior experiences in terms of years spent at a position at a certain level.

Finally, each applicant's application was categorized and assigned a label by the people involved in the data processing. They used three different labels ("Recommended", "Waitlisted", and "Rejected") to categorize the applications. The labels were based on the actual decisions about the applicants when they applied to the company previously.

You can get a overview of the processings in the following pictures.





Finally, in 2021, the dataset was ready for use. After the above-mentioned process to prepare the dataset, it was finalized and released for internal use in 2021. The company states that other companies can also use this dataset. However, to get access to this dataset, they would need to contact Company “A”, since the dataset is not publicly released.

To date, company “A” is not aware of anybody else using their dataset.

This concludes the discussion on the attributes and the population of the dataset. The information seems a fair starting point to get insights into the fundamentals of the dataset. More information is available on-demand when requested to the dataset owners.

A.2. Q&A-based Presentation for the Training Dataset Explanation

One snippet of every category in the explanation is provided. The answers for the rest of the questions in each category (along with the questions) are also provided in this appendix.

The system's decisions/recommendations are influenced by its training dataset and how that is used to train the system. The training dataset for this hiring system has the following characteristics.

 Collection	Information about the amount of data, the source of the data, the collectors, and the labeling process	
How many instances are in the dataset?  The dataset contains data from resumes of 17562 applicants. For each applicant, the dataset contains information about 15 attributes. This includes attributes like - Education, Experience, Skills, Age, and References. Who collected the data?  What was the data collection process?  What tools were used in data collection?  Was any pre-processing done on the data?  How were the data labeled?  Any other comments? 		
 Demographics	Information on gender, ethnicity, age, and country distribution of the instances	
 Recommended Usage	List of recommended use cases for the dataset	
 General Information	Overview information about the dataset	

The pictures and visuals used in this webpage was taken from different sources. To see the full list of attributions, please click [here](#).

The system's decisions/recommendations are influenced by its training dataset and how that is used to train the system. The training dataset for this hiring system has the following characteristics.

	Collection	Information about the amount of data, the source of the data, the collectors, and the labeling process	>
	Demographics	Information on gender, ethnicity, age, and country distribution of the instances	✓
Gender distribution of the instances?		>	
Age distribution of the instances?		>	
Ethnicity distribution of the instances?		✓	
Ethnicity of the population in the dataset ♂ White ♂ Brown ♂ Hispanic ♂ Asian ♂ Black ♂ Others A Flourish data visualization			
Country distribution of the instances?		>	
	Recommended Usage	List of recommended use cases for the dataset	>
	General Information	Overview information about the dataset	>

The pictures and visuals used in this webpage was taken from different sources. To see the full list of attributions, please click [here](#).

The system's decisions/recommendations are influenced by its training dataset and how that is used to train the system. The training dataset for this hiring system has the following characteristics.

	Collection	Information about the amount of data, the source of the data, the collectors, and the labeling process	>
	Demographics	Information on gender, ethnicity, age, and country distribution of the instances	>
	Recommended Usage	List of recommended use cases for the dataset	✓
Suggested use cases for the dataset? > Where you should not use the dataset? ✓ The dataset should not be used for any other purpose (e.g., promotion, performance evaluation) other than determining the initial hiring decisions. Any other information to know before using the dataset? >			
	General Information	Overview information about the dataset	>

The pictures and visuals used in this webpage was taken from different sources. To see the full list of attributions, please click [here](#).

The system's decisions/recommendations are influenced by its training dataset and how that is used to train the system. The training dataset for this hiring system has the following characteristics.

	Collection	Information about the amount of data, the source of the data, the collectors, and the labeling process	>
	Demographics	Information on gender, ethnicity, age, and country distribution of the instances	>
	Recommended Usage	List of recommended use cases for the dataset	>
	General Information	Overview information about the dataset	▼
		When the dataset was released?	▼
		The dataset was released in 2021.	
		When was the data collected?	>
		Where the dataset was previously used?	>
		Was consent obtained from individuals related to the dataset?	>
		Have any updates been provided for the dataset?	>
		Is the dataset publicly available?	>
		Any other comments?	>

The pictures and visuals used in this webpage was taken from different sources. To see the full list of attributions, please click [here](#).

Below we provide all the questions and answers presented in the Q&A interface.

Collection: Information about the amount of data, the source of the data, the collectors, and the labeling process

Q. How many instances are in the dataset?

The dataset contains data from resumes of 17562 applicants. For each applicant, the dataset contains information about 15 attributes. This includes attributes like - Education, Experience, Skills, Age, and References.

Q. Who collected the data?

The data was collected by an internal committee within the company.

Q. What was the data collection process?

The company conducted the data collection based on the resumes they received between 2016 and 2020. They used an automated process to extract all the relevant attributes of the applicants. These attributes were then stored in the dataset and all the other information about the applicants in the resumes was discarded.

Q. What tools were used in data collection?

The resumes were first gathered manually from a company repository. Then an automated extraction tool was used to extract the data from the resumes.

Q. Was any pre-processing done on the data?

The use of the automated process in data collection resulted in missing attributes for some of the applicants. If applicable, these missing attributes were filled with the average values for the specific attribute. Otherwise, the applicant's data was deleted from the dataset. Further, applicants' educational qualifications and results to a single scale to allow better comparisons between applicants. Finally, numerical values were used to represent prior experiences in terms of years spent at a position at a certain level.

Q. How were the data labeled?

Each applicant was assigned a label based on the final decision. Three different labels ("Recommended", "Waitlisted", and "Rejected") were used. The labels were based on the actual decisions about the applicants when they applied to the company previously.

Q. Any other comments?

Each applicant was assigned a label based on the final decision. Three different labels (“Recommended”, “Waitlisted”, and “Rejected”) were used. The labels were based on the actual decisions about the applicants when they applied to the company previously.

Demographics: Information on gender, ethnicity, age, and country distribution of the instances

Q. Gender distribution of the instances?

60% of the applicants are men. The other 40% are women (37%) and non-binary (3%) people. (Accompanied with a pie graph, similar to the Data Story)

Q. Age distribution of the instances?

The age distribution of the applicants is provided in 5 groups. (Accompanied with a stacked bar graph, similar to the Data Story)

Q. Ethnicity distribution of the instances?

Presented as a pictograph, similar to the Data Story

Q. Country distribution of the instances?

Applicants from a range of countries are part of the dataset. (Accompanied with a map visualization, similar to the Data Story)

Recommended Usage: List of recommended use cases for the dataset

Q. Suggested use cases for the dataset?

The dataset is suggested to be used for hiring purposes only.

Q. Where should you not use the dataset?

The dataset should not be used for any other purpose (e.g., promotion, performance evaluation) other than determining the initial hiring decisions.

Q. Any other information to know before using the dataset?

N/A

General Information: Overview information about the dataset

Q. When was the dataset released?

The dataset was released in 2021.

Q. When was the data collected?

The dataset contains records from the resumes received between 2016 and 2020, which were gathered together in 2020.

Q. Where was the dataset previously used?

To date, the dataset has not been used anywhere.

Q. Was consent obtained from individuals related to the dataset?

No, individual consents were not taken. However, all the directly identifying information was discarded when the dataset was compiled.

Q. Have any updates been provided for the dataset?

No.

Q. Is the dataset publicly available?

No. However, one can get access to the dataset by contacting the owners.

Q. Any other comments?

No.

A.3. Recruitment Poster

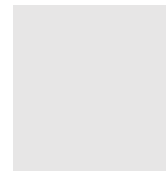
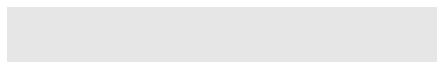


Interested in learning about how AI systems can explain themselves?



We are looking for people to participate in a study to evaluate and share their opinions on the different ways an AI system could explain how it was trained. The study will take approximately 60-90 minutes of your time and you will get \$25 cash or gift card (at your choosing).

If you are more than 18 years old and you are interested in participating in our study, please contact -



Research approved by University of Manitoba Joint Faculty Research Ethics Board. The Research Ethics Board can be reached by phone



A.4. Demographic and Background Questionnaire

About yourself

1. How old are you?
2. What is your gender?
3. In which country do you currently reside?
4. What is the highest level of education you have completed?
5. What is your educational background/specialization?

Questions on your knowledge about Artificial Intelligence (AI)

6. Have you ever taken or are currently taking any classes on Artificial Intelligence (AI)/ Machine Learning (ML)?
 - a. Yes
 - b. No
7. When it comes to AI/ML, I believe I have
 - a. No knowledge:
 - b. A little knowledge:
 - c. Some knowledge:
 - d. Moderate knowledge:
 - e. A lot of knowledge:
8. Which of the following is an unsupervised machine learning task?

- a. Distinguishing pictures containing cats from pictures not containing cats
 - b. Flagging text messages as appropriate or inappropriate
 - c. Divide data points into different clusters without any labels available
 - d. Predict the value of a house after training on a dataset with house features and values
 - e. I do not know the answer
9. What is the general goal of reinforcement learning?
- a. Maximize potential or expected punishment
 - b. Maximize potential or expected reward
 - c. Get to the goal as soon as possible
 - d. Avoid the most obstacles in any given state
 - e. I do not know the answer
10. Please provide your agreements with the following statements. (On a scale of 1-7) (taken from Wang et al. [323]. Reference number indicate the reference number in the main submission)
- a. I can distinguish between smart devices and non-smart devices.
 - b. I do not know how AI technology can help me.
 - c. I can identify the AI technology employed in the applications and products I use.

- d. I can skillfully use AI applications or products to help me with my daily work.
- e. It is usually hard for me to learn to use a new AI application or product.
- f. I can use AI applications or products to improve my work efficiency.
- g. I can evaluate the capabilities and limitations of an AI application or product after using it for a while.
- h. I can choose a proper solution from various solutions provided by a smart agent.
- i. I can choose the most appropriate AI application or product from a variety for a particular task.
- j. I always comply with ethical principles when using AI applications or products.
- k. I am never alert to privacy and information security issues when using AI applications or products.
- l. I am always alert to the abuse of AI technology.

A.5. Scenario Used in the Study – Automated Hiring System

A company is deciding on whether to purchase a computer system that is powered by machine learning algorithms to help informing the company's hiring decisions. This computer system will review and identify promising candidates that the company should consider from a pool of applications.

Before purchasing the system, the company would like people to scrutinize the system, focusing on how the system is trained and the data that it uses. To help the company scrutinize the system, the makers of the computer system have provided an explanation of the dataset it uses and descriptions of how the dataset was formulated. The company wants you to use the information in this explanation to conduct a SWOT analysis of the system. The primary objective of this analysis is to help the company identify the advantages of the system, the flaws that the system, how the system can potentially be improved further, and situations that could potentially introduce flaws into the system.

You will be given 25 minutes to interact with the explanation and conduct this analysis. During this time, we ask that you note at least 10 different points (strengths/weaknesses/opportunities/threats). If you want to provide more, you are welcome to do so. Please try to include a mix of strengths, weaknesses, opportunities to improve, and any threat to the system.

Please make sure you include at least one strength/weakness/opportunity/threat in your notes.

A.6. Informed Consent Form



Department of Computer Science
Faculty of Science
Discover the unknown + Invent the future

E2-445 EITC
University of Manitoba
Winnipeg, Manitoba, Canada
R3T 2N2
T: 204-474-8313
F: 204-474-7609

Research Project Title: Explanations in Support of Fairness Judgments of Data-Driven Decisions in Machine Learning Systems

Researchers:

Dr. Andrea Bunt, Professor, Department of Computer Science, University of Manitoba, [REDACTED]

Md Ariful Islam Anik, Graduate Research Assistant, Department of Computer Science, University of Manitoba, [REDACTED]

Research Sponsored by Natural Sciences and Engineering Research Council of Canada

Please take the time to read this carefully and to ensure you understand all the information.

This consent form, a copy of which will be left with you for your records and reference, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, you should feel free to ask. Please take the time to read this carefully and to understand any accompanying information.

You are invited to participate in a research study on the topic of evaluating the impact of the presentation styles of explanations that target training data on users' perceptions of them. The goal is to explore whether the presentation of information about the training data has any effect on the utility of the presented information. Your participation in this research study will involve interacting with our explanations in a given scenario presented in an online survey platform (Microsoft Forms) and providing written feedback on them. We will ask you to complete a background questionnaire at the beginning of the study and a post-scenario questionnaire after you have interacted with the explanation. Finally, we will ask you to participate in an interview. If you have any questions or concerns at this time or any time during the study, please feel free to ask the researcher for clarification.

The study will take place either remotely or in-person (based on your preference). If you choose to participate remotely, the session will take place over a video conferencing software (e.g., Zoom, Google Meet, Cisco Webex, Skype) of your choice. You need to have access to a computer if you want to participate remotely. Alternatively, if you choose to participate in-person, the session will take place in the HCI lab (E2-534 or E2-551) situated in the EITC building at the University of Manitoba (please take the time to read the appendix about the steps taken for ensuring your safety from COVID-19 for the in-person session). We will provide a computer for conducting the in-person session.

1970



2020

www.sci.umanitoba.ca/cs/

The benefits of participating in this research are that you may gain a greater understanding of how machine learning systems are trained and how training datasets contribute to the performance of machine learning systems. Further, the knowledge you will contribute to this study will be used to improve explanations from machine learning systems. The risks of this study are no greater than in everyday life.

Participation in this study is voluntary and will take approximately 60-90 minutes of your time. You will receive a \$25 (Canadian dollars) compensation in the form of cash or a gift card (at your choosing) after signing the consent form. You are free to withdraw from the study at any time up to the end of the study and/or refrain from answering any questions you prefer to omit. Even by withdrawing, you will keep your compensation.

We wish to record our discussions with you by using a digital audio recorder. The audio recording will assist our data analysis by allowing us to review the discussion and the study session in detail. Any information you choose to contribute to our discussion is completely confidential and will be used for anonymized research analysis. We may use anonymized quotes for purposes of dissemination. Your name will not be included or in any other way associated with the data presented in the result of this study.

We also wish to record your interactions with the explanations. This recording will contain all the actions (e.g., button clicked, mouse movements) you perform in the prototype and the feedback you provide in the interface. This recording will only be used for anonymized research analysis only. Your personal information will not be associated with this recording in any way.

By signing this consent form, you agree that you understand this and that we may use the recorded audio and the interaction video for data analysis purposes only. Unfortunately, if you do not wish to be recorded then you cannot participate in this study.

Data collected during this study will be retained for a maximum period of three years in a locked cabinet or in a password-protected computer in a locked office or laboratory in the EITC building at the University of Manitoba, to which only researchers associated with this project (Dr. Andrea Bunt, Md Ariful Islam Anik) have access. The data collected through the online survey platform (Microsoft Forms) and the interaction tool (www.userback.io) will be stored in their respective secured server, which is only accessible from a password-protected account of one of the researchers (Md Ariful Islam Anik). All data will be destroyed by December 2025. In addition, the University of Manitoba may look at research records to see that the research is being done in a safe and proper way. We intend to present the results as academic publications and a thesis that will be published in MSpace. Once published, the results of the study will be made available to the public for free at <http://hci.cs.umanitoba.ca/>. Again, no personal information about you will be included.

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research study. By doing this you also confirm that you are of the age of

majority in Canada (18 years or more). In no way does this waive your legal rights nor release the researchers, sponsors, or involved institutions from their legal and professional responsibilities.

This research has been approved by the University of Manitoba Joint Faculty Research Ethics Board. If you have any concerns or complaints about this project, please contact Dr. Andrea Bunt at [REDACTED] or [REDACTED] or the Human Ethics Coordinator at [REDACTED] or [REDACTED]. A copy of this consent form has been given to you to keep for your records and reference.

☐ I wish to receive a summary of the findings.

☐ I wish to receive a copy of the interview transcript of the audio recording to confirm its accuracy.

Please write your email address if you checked the box above:

Participant's email address: _____

Participant's Signature: _____ Date _____

Researcher's Signature: _____ Date _____

A.7. Post Study Questionnaire

Based on your interaction with the hiring system, please provide your agreement to the following statements (on a scale of 1-7)

1. I am confident in the system
2. The system is reliable
3. The system is safe
4. The system is consistent
5. The system is usable
6. The system is accountable
7. The system is predictable
8. I can trust the system
9. The system is free of bias
10. The system upholds ethical and moral standards
11. The system's explanation is reasonable
12. I would agree with the hiring recommendations made by the system
13. It was clear to me which data the system uses to make recommendations
14. I understood which data was used by the system when recommending applicants
15. I understood which characteristics were considered to recommend applicants
16. The explanation was easy to follow

17. I liked how the information was presented to me
18. I understood the explanation
19. The explanation was too detailed
20. The explanation contained sufficient information for me to do the SWOT analysis.
21. I am satisfied with the breadth of information provided in the explanation
22. I am satisfied with the depth of information provided in the explanation
23. My understanding of how machine learning systems use data has improved after seeing the explanation
24. My vocabulary of machine learning and AI has increased after interacting with the explanation
25. I am confident that I can provide a clear summary of the data to another person based on what I saw in the explanation

A.8. Ethics Certificate



**UNIVERSITY
OF MANITOBA**

**Research Ethics
and Compliance**

Human Ethics
 208-194 Dafoe Road
 Winnipeg, MB
 Canada R3T 2N2
 Phone +204-474-7122
 Email: humanethics@umanitoba.ca

PROTOCOL APPROVAL

TO: **Andrea Bunt**
Principal Investigator

FROM: **Julia Witt, Chair**
Joint-Faculty Research Ethics Board (JFREB)

Re: **Protocol J2019:083**
"Explanations in Support of Fairness Judgments of Data-Driven Decisions in Machine Learning Systems"

Effective: November 13, 2019

Expiry: November 13, 2020

Joint-Faculty Research Ethics Board (JFREB) has reviewed and approved the above research. JFREB is constituted and operates in accordance with the current *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*.

This approval is subject to the following conditions:

1. Approval is granted for the research and purposes described in the application only.
2. Any modification to the research or research materials must be submitted to JFREB for approval before implementation.
3. Any deviations to the research or adverse events must be submitted to JFREB as soon as possible.
4. This approval is valid for one year only and a Renewal Request must be submitted and approved by the above expiry date.
5. A Study Closure form must be submitted to JFREB when the research is complete or terminated.
6. The University of Manitoba may request to review research documentation from this project to demonstrate compliance with this approved protocol and the University of Manitoba *Ethics of Research Involving Humans*.

Funded Protocols:

- Please e-mail a copy of this Approval, identifying the related UM Project Number, to the Research Grants Officer at

A.9. TCPS 2: CORE Certificate



Appendix B – Information Depth Study Documents

B.1. Recruitment Poster



**University
of Manitoba**



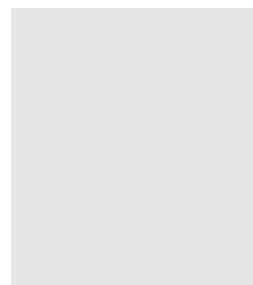
Interested in learning about how AI systems can explain themselves?



We are looking for people to participate in a study to evaluate and share their opinions on the different ways an AI system could explain how it was trained.

The study will take approximately 90-120 minutes of your time and you will get \$30 in e-transfer or digital gift card (at your choosing).

If you are more than 18 years old and you are interested in participating in our study, please contact -



*This research has been approved by the Research Ethics Board at the University of Manitoba, Fort Garry campus.
The Research Ethics Board can be reached by phone*



B.2. Pre-screening Questionnaire

1. Please provide your agreement with the following statements. (On a scale of 1-7)
(taken from Wang et al. [323], reference number indicate the reference number in the main submission)
 - a. I can distinguish between smart devices and non-smart devices.
 - b. I do not know how AI technology can help me.
 - c. I can identify the AI technology employed in the applications and products I use.
 - d. I can skillfully use AI applications or products to help me with my daily work.
 - e. It is usually hard for me to learn to use a new AI application or product.
 - f. I can use AI applications or products to improve my work efficiency.
 - g. I can evaluate the capabilities and limitations of an AI application or product after using it for a while.
 - h. I can choose a proper solution from various solutions provided by a smart agent.
 - i. I can choose the most appropriate AI application or product from a variety for a particular task.
 - j. I always comply with ethical principles when using AI applications or products.

- k. I am never alert to privacy and information security issues when using AI applications or products.
 - l. I am always alert to the abuse of AI technology.
- 2. What is your educational background/specialization?
 - a. Computer Science
 - b. Engineering
 - c. Non-engineering
- 3. Have you ever taken or are currently taking any classes on Artificial Intelligence (AI)/ Machine Learning (ML)?
 - a. Yes
 - b. No

B.3. Demographic Questions Asked to Participants

1. How old are you?
2. What is your gender?
 - a. Woman
 - b. Man
 - c. Non-binary
 - d. Prefer not to say
 - e. Other (please type)
3. In which country do you currently reside?
4. What is the highest level of education you have completed?
 - a. Some high school or less
 - b. High school diploma or GED
 - c. Some college, but no degree
 - d. Associates, or technical degree
 - e. Bachelor's degree
 - f. Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, etc.)
 - g. Prefer not to say

B.4. Explanation Prototypes

a. *Prototype 1 – Admission System (Detailed – No Progressive Disclosure)*

The system's decisions/recommendations are influenced by its training dataset and how that is used to train the system. The training dataset for this admission recommendation system has the following characteristics.



Collection

Information about the amount of data, the source of the data, the collectors, and the labeling process

How many instances are in the dataset?

The dataset contains application data for 11837 students who applied to the University of X for graduate programs in prior years. For each applicant, the dataset contains primary information like - Prior Education, GPA, Institutions, Experience, Extra-curricular skills, References.

Who collected the data?

The data was gathered by internal committee at the University of X, with the permission provided by the Graduate Admissions Committee of the University of the X. The internal committee consisted 5 members who brought different expertise to oversee the collection process.

What was the data collection process?

The committee conducted the data collection based on the applications that were submitted to the admission portal in the last 8 years. They used an automated process to extract all the relevant attributes of the applicants. These attributes were then stored in the dataset and all the personally identifying information was discarded. Some additional unspecified automatic filtering was done. The accuracy of the provided information in the resumes was not verified.

What tools were used in data collection?

An automated text extraction tool was used to extract the data from the applications.

Was any pre-processing done on the data?

All the reference letter scores were transformed into a numerical grade from 1 to 5 where a score of 1 indicates a poor reference letter and a score of 5 indicates an excellent reference letter. This was done manually by three different people. The use of the automated process in data collection resulted in missing attributes for some of the applicants. If applicable, these missing attributes were filled with the average values for the specific attribute.

How were the data labeled?

Each of the applicants data was labeled as "recommended", "not recommended", or "waitlisted" based on the manual decision made for the students.

Any other comments?

As an additional method of verification, The company ran a manual analysis for some randomly selected applicants to analyze the accuracy of the automated process. They were satisfied with the general accuracy of the automation.

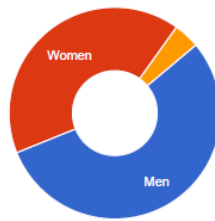


Demographics

Information on gender, ethnicity, age, and country distribution of the instances

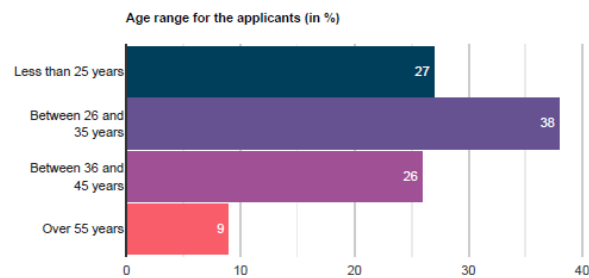
Gender distribution of the instances?

55% of the applicants are men. The other % are women (41%) and non-binary (4%) people.



Age distribution of the instances?

The age distribution of the applicants are provided in 4 groups.

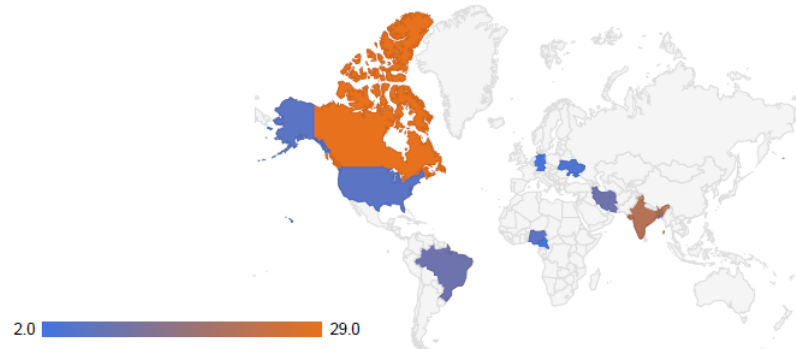


Ethnicity distribution of the instances?

Ethnicity distribution wasn't mentioned explicitly in the dataset. However, the dataset contains student data from more than 87 countries that cover a range of ethnicities.

Country distribution of the instances?

Applicants from a range of countries are part of the dataset. Countries with a percentage of less than 2% are not shown.



Recommended Usage

List of recommended use cases for the dataset

Suggested use cases for the dataset?

The dataset was suggested to be used for recommending outcomes for student applications in graduate programs.

Where you should not use the dataset?

The dataset should not be used to profile students for noneducational purposes (e.g., loan application, housing application).

Any other information to know before using the dataset?

N/A



General Information

Overview information about the dataset

When the dataset was released?

The dataset was released in 2021.

When was the data collected?

The dataset contains records from the applications received between 2016 and 2020, which were gathered together in 2020.

Where the dataset was previously used?

To date, the dataset has not been used anywhere.

Was consent obtained from individuals related to the dataset?

No, individual consents were not taken. However, individuals already consented for their information to be used for admission decisions and all the directly identifying information was discarded when the dataset was compiled.

Have any updates been provided for the dataset?

Yes. Updated in January 2023 to include the admission data for 2021 and 2022.

Is the dataset publicly available?

No. However, one can get access to the dataset by contacting the owners.

Any other comments?

If any errors are found, please contact the owners for potential fixes.

b. Prototype 2 – Admission System (Detailed – Progressive Disclosure)

The system's decisions/recommendations are influenced by its training dataset and how that is used to train the system. The training dataset for this admission recommendation system has the following characteristics.

 Collection	<u>Information about the amount of data, the source of the data, the collectors, and the labeling process</u>	
<u>How many instances are in the dataset?</u>  The dataset contains application data for 11837 students who applied to the University of X for graduate programs in prior years. For each applicant, the dataset contains primary information like - Prior Education, GPA, Institutions, Experience, Extra-curricular skills, References. <u>Who collected the data?</u>  <u>What was the data collection process?</u>  <u>What tools were used in data collection?</u>  <u>Was any pre-processing done on the data?</u>  <u>How were the data labeled?</u>  <u>Any other comments?</u> 		
 Demographics	<u>Information on gender, ethnicity, age, and country distribution of the instances</u>	
 Recommended Usage	<u>List of recommended use cases for the dataset</u>	
 General Information	<u>Overview information about the dataset</u>	

c. Prototype 3 – Admission System (Summary – No Progressive Disclosure)

The system's decisions/recommendations are influenced by its training dataset and how that is used to train the system. The training dataset for this admission recommendation system has the following characteristics.

 Collection	The dataset contains data from over 11 thousands applicants' resumes, covering important attributes like prior education, GPA, experiences, skills, and references. It was collected by an internal committee using a combination of manual and automated processes. The data was labeled based on previous decisions.
 Demographics	The dataset provides the gender, age, and country distribution of the applicants. Gender distribution is provided in 3 categories, age is grouped into four categories, and country distributions is also provided.
 Recommended Usage	The dataset is recommended for graduate admission decision recommendation purposes only, and caution is advised against using it for other purposes like housing applications or loan applications.
 General Information	The dataset was released in 2021 and contains data from resumes received between 2015 and 2020. It was further updated in 2023 to include data from 2021 and 2022. It has not been used elsewhere, and individual consents were not obtained. The dataset is not publicly available, but access can be requested from the owners.

d. Prototype 4 - Admission System (Summary – Progressive Disclosure)

The system's decisions/recommendations are influenced by its training dataset and how that is used to train the system. The training dataset for this admission recommendation system has the following characteristics.

	Collection	The dataset contains data from over 11 thousands applicants' resumes, covering important attributes like prior education, GPA, experiences, skills, and references. It was collected by an internal committee using a combination of manual and automated processes. The data was labeled based on previous decisions.
	Demographics	click here to reveal information
	Recommended Usage	click here to reveal information
	General Information	click here to reveal information

e. Prototype 5 – Hiring System (Detailed – No Progressive Disclosure)

The system's decisions/recommendations are influenced by its training dataset and how that is used to train the system. The training dataset for this hiring system has the following characteristics.



Collection

Information about the amount of data, the source of the data, the collectors, and the labeling process

How many instances are in the dataset?

The dataset contains data from resumes of 17562 applicants. For each applicant, the dataset contains information about 15 attributes. This includes attributes like - Education, Experience, Skills, Age, and References.

Who collected the data?

The data was collected by an internal committee within the company. The committee consisted 6 executive members who were from different background and were assigned on the project with the goal of creating a diverse dataset.

What was the data collection process?

The company conducted the data collection based on the resumes they received between 2016 and 2020. They used an automated process to extract all the relevant attributes of the applicants. These attributes were then stored in the dataset and all the other information about the applicants in the resumes was discarded. Some additional unspecified automatic filtering was done. The accuracy of the provided information in the resumes was not verified.

What tools were used in data collection?

The resumes were first gathered manually from a company repository. Then an automated extraction tool was used to extract the data from the resumes.

Was any pre-processing done on the data?

The use of the automated process in data collection resulted in missing attributes for some of the applicants. If applicable, these missing attributes were filled with the average values for the specific attribute. Otherwise, the applicant's data was deleted from the dataset. Further, applicants' educational qualifications and results to a single scale to allow better comparisons between applicants. Finally, numerical values were used to represent prior experiences in terms of years spent at a position at a certain level.

How were the data labeled?

Each applicant was assigned a label based on the final decision. Three different labels ("Recommended", "Waitlisted", and "Rejected") were used. The labels were based on the actual decisions about the applicants

when they applied to the company previously. For the resumes that were “Waitlisted”, the label represents the initial decision for the applicant.

Any other comments?

The company ran a manual analysis for some randomly selected resumes to analyze the accuracy of the automated process in extracting applicants' attributes from the resumes. They were satisfied with the general accuracy of the automation.

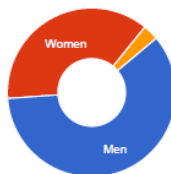


Demographics

Information on gender, ethnicity, age, and country distribution of the instances

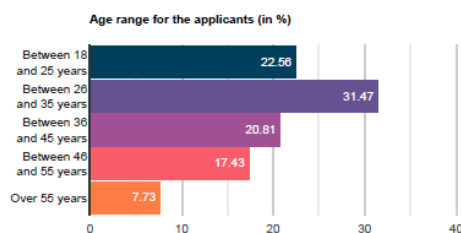
Gender distribution of the instances?

60% of the applicants are men. The other 40% are women (37%) and non-binary (3%) people.



Age distribution of the instances?

The age distribution of the applicants are provided in 5 groups.



Ethnicity distribution of the instances?

Ethnicity of the population in the dataset

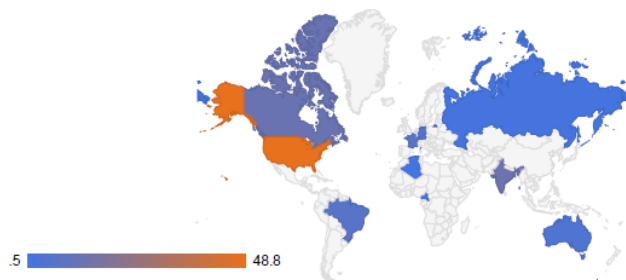
White Brown Hispanic Asian Black Others



* A Flourish data visualization

Country distribution of the instances?

Applicants from a range of countries are part of the dataset.



Recommended Usage

List of recommended use cases for the dataset

Suggested use cases for the dataset?

The dataset is suggested to be used for hiring purposes only.

Where you should not use the dataset?

The dataset should not be used for any other purpose (e.g., promotion, performance evaluation) other than determining the initial hiring decisions.

Any other information to know before using the dataset?

N/A



General Information

Overview information about the dataset

When the dataset was released?

The dataset was released in 2021.

When was the data collected?

The dataset contains records from the resumes received between 2016 and 2020, which were gathered together in 2020.

Where the dataset was previously used?

To date, the dataset has not been used anywhere.

Was consent obtained from individuals related to the dataset?

No, individual consents were not taken. However, all the directly identifying information was discarded when the dataset was compiled.

Have any updates been provided for the dataset?

No.

Is the dataset publicly available?

No. However, one can get access to the dataset by contacting the owners.

Any other comments?

If any errors are found, please contact the owners for potential fixes.

f. Prototype 6 – Hiring System (Detailed – Progressive Disclosure)

The system's decisions/recommendations are influenced by its training dataset and how that is used to train the system. The training dataset for this hiring system has the following characteristics.

	<u>Collection</u>	<u>Information about the amount of data, the source of the data, the collectors, and the labeling process</u>	➤
<u>How many instances are in the dataset?</u> ▼ The dataset contains data from resumes of 17562 applicants. For each applicant, the dataset contains information about 15 attributes. This includes attributes like - Education, Experience, Skills, Age, and References. <u>Who collected the data?</u> ➤ <u>What was the data collection process?</u> ➤ <u>What tools were used in data collection?</u> ➤ <u>Was any pre-processing done on the data?</u> ➤ <u>How were the data labeled?</u> ➤ <u>Any other comments?</u> ➤			
	<u>Demographics</u>	<u>Information on gender, ethnicity, age, and country distribution of the instances</u>	➤
	<u>Recommended Usage</u>	<u>List of recommended use cases for the dataset</u>	➤
	<u>General Information</u>	<u>Overview information about the dataset</u>	➤

g. Prototype 7 – Hiring System (Summary – No Progressive Disclosure)

The system's decisions/recommendations are influenced by its training dataset and how that is used to train the system. The training dataset for this hiring system has the following characteristics.

	Collection	The dataset contains data from over 17 thousands applicants' resumes, covering 15 attributes. It was collected by an internal committee using a combination of manual and automated processes. The data was labeled based on previous decisions.
	Demographics	The dataset provides the gender, age, ethnicity, and country distribution of the applicants. Gender distribution is provided in 3 categories, age is grouped into five categories, and ethnicity and country distributions are also provided.
	Recommended Usage	The dataset is recommended for hiring purposes only, and caution is advised against using it for other purposes like promotion or performance evaluation.
	General Information	The dataset was released in 2021 and contains data from resumes received between 2016 and 2020. It has not been used elsewhere, and individual consents were not obtained. The dataset is not publicly available, but access can be requested from the owners.

h. Prototype 8 – Hiring System (Summary – Progressive Disclosure)

The system's decisions/recommendations are influenced by its training dataset and how that is used to train the system. The training dataset for this hiring system has the following characteristics.

 Collection	click here to reveal information
 Demographics	The dataset provides the gender, age, ethnicity, and country distribution of the applicants. Gender distribution is provided in 3 categories, age is grouped into five categories, and ethnicity and country distributions are also provided.
 Recommended Usage	click here to reveal information
 General Information	click here to reveal information

B.5. Scenarios Used in the Study

The following are the two scenarios used in the study. They were selected for their similarity in significance and familiarity to a diverse audience.

a. Automated Hiring System Scenario

A company is deciding on whether to purchase a computer system that is powered by machine learning algorithms to help inform the company's hiring decisions. This computer system will review and identify promising candidates that the company should consider from a pool of applications.

Before purchasing the system, the company would like people to scrutinize the system, focusing on how the system is trained and the data that it uses. To help the company scrutinize the system, the makers of the computer system have provided an explanation of the dataset it uses and descriptions of how the dataset was formulated. The company wants you to use the information in this explanation to conduct an analysis of the system. The primary objective of this analysis is to help the company identify the advantages of the system and its flaws and shortcomings.

You will be given 20 minutes to interact with the explanation and conduct this analysis. During this time, we ask that you note 8 different points (strengths/weaknesses). If you want to provide more, you are welcome to do so. Please try to include a mix of strengths and weaknesses of the system.

Please make sure you include at least one strength and one weakness in your feedback.

b. Automated Decision Recommendation System Scenario

The University of X is deciding on whether to purchase a computer system that can help inform the student recruiter to give admission decisions to new graduate applicants. This computer system will assess applicants based on a machine learning model that profiles students based on the application materials and identify the promising applicants.

Before purchasing the system, the university would like people to scrutinize the system, focusing on how the system is trained and the data that it uses. To help scrutinize the system, the makers of the computer system have provided an explanation of the dataset it uses and descriptions of how the dataset was formulated. The university wants you to use the information in this explanation to conduct an analysis of the system. The primary objective of this analysis is to help the university to identify the advantages of the system and its flaws and shortcomings.

You will be given 20 minutes to interact with the explanation and conduct this analysis. During this time, we ask that you note 8 different points (strengths/weaknesses). If you want to provide more, you are welcome to do so. Please try to include a mix of strengths and weaknesses of the system.

Please make sure you include at least one strength and one weakness in your feedback.

B.6. Informed Consent Form



**University
of Manitoba**

**Department of Computer Science
Faculty of Science**
Discover the unknown + Invent the future

E2-445 EITC
University of Manitoba
Winnipeg, Manitoba, Canada
R3T 2N2
T: 204-474-8313
F: 204-474-7609

Research Project Title: Explanations in Support of Fairness Judgments of Data-Driven Decisions in Machine Learning Systems

Researchers:

Dr. Andrea Bunt, Professor, Department of Computer Science, University of Manitoba, [REDACTED]

Md Ariful Islam Anik, Graduate Research Assistant, Department of Computer Science, University of Manitoba, [REDACTED]

Research Sponsored by Natural Sciences and Engineering Research Council of Canada

Please take the time to read this carefully and to ensure you understand all the information.

This consent form, a copy of which will be left with you for your records and reference, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, you should feel free to ask. Please take the time to read this carefully and to understand any accompanying information.

You are invited to participate in a research study on the topic of evaluating the impact of volume and presentation of the contents for explanations that target training data. The goal is to explore how does the volume and presentation of information about the training data effect on the utility of the presented information. Your participation in this research study will involve interacting with our explanations in two given scenarios presented in an online survey platform (Microsoft Forms) and providing written feedback on them. We will ask you to complete a background questionnaire at the beginning of the study and two post-scenario questionnaires after you have interacted with the explanations. Finally, we will ask you to participate in an interview. If you have any questions or concerns at this time or any time during the study, please feel free to ask the researcher for clarification.

The study will take place either remotely or in-person (based on your preference). If you choose to participate remotely, the session will take place over a video conferencing software (e.g., Zoom, Google Meet, Cisco Webex, Skype) of your choice. You need to have access to a computer if you want to participate remotely. Alternatively, if you choose to participate in-person, the session will take place in the HCI lab (E2-534 or E2-551) situated in the EITC building at the University of Manitoba. We will provide a computer for conducting the in-person session.

The benefits of participating in this research are that you may gain a greater understanding of how machine learning systems are trained and how training datasets contribute to the performance of machine learning

1970



2020

www.sci.umanitoba.ca/cs/

systems. Further, the knowledge you will contribute to this study will be used to improve explanations from machine learning systems. The risks of this study are no greater than in everyday life.

Participation in this study is voluntary and will take approximately 90-120 minutes of your time. You will receive a \$30 (Canadian dollars) compensation in the form of e-transfer or a digital gift card (at your choosing) after signing the consent form. You are free to withdraw from the study at any time up to the end of the study and/or refrain from answering any questions you prefer to omit. Even by withdrawing, you will keep your compensation.

We wish to record our discussions with you by using a digital audio recorder. The audio recording will assist our data analysis by allowing us to review the discussion and the study session in detail. Any information you choose to contribute to our discussion is completely confidential and will be used for anonymized research analysis. We may use anonymized quotes for purposes of dissemination. Your name will not be included or in any other way associated with the data presented in the result of this study.

We also wish to record your interactions with the explanations. This recording will contain all the actions (e.g., button clicked, mouse movements) you perform in the prototype and the feedback you provide in the interface. This recording will only be used for anonymized research analysis only. Your personal information will not be associated with this recording in any way.

By signing this consent form, you agree that you understand this and that we may use the recorded audio and the interaction video for data analysis purposes only. Unfortunately, if you do not wish to be recorded then you cannot participate in this study.

Data collected during this study will be retained for a maximum period of three years in a locked cabinet or in a password-protected computer in a locked office or laboratory in the EITC building at the University of Manitoba, to which only researchers associated with this project (Dr. Andrea Bunt, Md Ariful Islam Anik) have access. The data collected through the online survey platform (Microsoft Forms) and the interaction tool (www.userback.io) will be stored in their respective secured server, which is only accessible from a password-protected account of one of the researchers (Md Ariful Islam Anik). All data will be destroyed by December 2026. We intend to present the results as academic publications and a thesis that will be published in MSpace. Once published, the results of the study will be made available to the public for free at <http://hci.cs.umanitoba.ca/>. Again, no personal information about you will be included.

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to participate as a subject. In no way does this waive your legal rights nor release the researchers, sponsors, or involved institutions from their legal and professional responsibilities. You are free to withdraw from the study at any time, and /or refrain from answering any questions you prefer to omit, without prejudice or consequence. Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation.

The University of Manitoba may look at your research records to see that the research is being done in a safe and proper way.

This research has been approved by the Research Ethics Board at the University of Manitoba, Fort Garry campus. If you have any concerns or complaints about this project, you may contact any of the above-named persons or the Human Ethics Officer at [REDACTED] A copy of this consent form has been given to you to keep for your records and reference.

☐ I wish to receive a summary of the findings.

Please write your email address if you checked the box above:

Participant's email address: _____

Participant's Signature: _____ Date _____

Researcher's Signature: _____ Date _____

B.7. Post-Scenario Questionnaire

Please provide your agreement with the following statements (on a scale of 1-7).

1. I am confident in the system
2. The system is reliable
3. The system is safe
4. The system is consistent
5. The system is usable
6. The system is accountable
7. The system is predictable
8. I can trust the system
9. The system is free of bias
10. The system upholds ethical and moral standards
11. The system's explanation is reasonable
12. I would agree with the hiring recommendations made by the system
13. It was clear to me which data the system uses to make recommendations
14. I understood which data was used by the system when recommending applicants
15. I understood which characteristics were considered to recommend applicants
16. The explanation was easy to follow

17. I liked how the information was presented to me
18. I understood the explanation
19. The explanation was too detailed
20. The explanation contained sufficient information for me to identify strengths/weaknesses.
21. I am satisfied with the breadth of information provided in the explanation
22. I am satisfied with the depth of information provided in the explanation
23. My understanding of how machine learning systems use data has improved after seeing the explanation
24. My vocabulary of machine learning and AI has increased after interacting with the explanation
25. I am confident that I can provide a clear summary of the data to another person based on what I saw in the explanation

Please try to answer the following questions without seeking additional help from the explanation (i.e., looking at it again).

26. What kind of information was in the demographic section? (Just one sentence is fine)
27. Was there any updates provided to the dataset? (Yes/No/Maybe)

- 28. What kind of processes were used in data collection?
- 29. Were there any steps to label the data? (Yes/No/Maybe)
- 30. Where the data should not be used?

Please answer the following questions about the explanation? (on a scale of 1-7) (NASA TLX [137])

- 31. How mentally demanding was the explanation?
- 32. How physically demanding was the explanation?
- 33. How hurried or rushed was the pace of the task?
- 34. How successful were you in accomplishing what you were asked to do?
- 35. How hard did you have to work to accomplish your level of performance?
- 36. How insecure, discouraged, irritated, stressed, and annoyed were you?

Finally,

- 37. I would recommend purchasing the admission system based on the explanation. (yes/no/maybe)
- 38. How confident are you with your recommendation? (out of 1-10)

B.8. Ethics Certificate



**UNIVERSITY
OF MANITOBA**

**Research Ethics
and Compliance**

Human Ethics
 208-194 Dafoe Road
 Winnipeg, MB
 Canada R3T 2N2
 Phone +204-474-7122
 Email: humanethics@umanitoba.ca

PROTOCOL APPROVAL

TO: **Andrea Bunt**
Principal Investigator

FROM: **Julia Witt, Chair**
Joint-Faculty Research Ethics Board (JFREB)

Re: **Protocol J2019:083**
"Explanations in Support of Fairness Judgments of Data-Driven Decisions in Machine Learning Systems"

Effective: November 13, 2019

Expiry: November 13, 2020

Joint-Faculty Research Ethics Board (JFREB) has reviewed and approved the above research. JFREB is constituted and operates in accordance with the current *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*.

This approval is subject to the following conditions:

1. Approval is granted for the research and purposes described in the application only.
2. Any modification to the research or research materials must be submitted to JFREB for approval before implementation.
3. Any deviations to the research or adverse events must be submitted to JFREB as soon as possible.
4. This approval is valid for one year only and a Renewal Request must be submitted and approved by the above expiry date.
5. A Study Closure form must be submitted to JFREB when the research is complete or terminated.
6. The University of Manitoba may request to review research documentation from this project to demonstrate compliance with this approved protocol and the University of Manitoba *Ethics of Research Involving Humans*.

Funded Protocols:

- Please e-mail a copy of this Approval, identifying the related UM Project Number, to the Research Grants Officer at

B.9. Comprehension Questionnaire Data Analysis

Here we provide sample graded answers with justifications.

Question	Sample Answers from Participants	Graded Score	Justification
What kind of information was in the demographic section?	facts about applicants (age, gender, country, and ethnicity distribution)	1	Demonstrates the recollection of all the demographic distributions present
	ranges of ages and countries of applicants	0.5	Partial recollection of the demographic distribution present in the explanation
	17 thousand applicants who submitted resumes	0	Irrelevant answer
What kind of processes were used in data collection?	An internal committee using a mix of manual and automated processes	1	Accurate information based on the explanation
	automated	0.5	Partially correct answer
	Students graded the resumes and them AI analyzed it	0	Irrelevant answer
Where the data should not be used?	to profile students for noneducational purposes (e.g., loan application, housing application).	1	Accurate demonstration of information in the explanation
	In public	0	Incorrect answer based on the explanation

Appendix C – AI Literacy Study Documents

C.1. Initial Set of Likert-Scale and Multiple-Choice Items

a. Likert Items

Item	Source
1. I can recognize AI-driven systems in various applications.	Wang et al. [323]
2. I can identify intelligent systems from non-intelligent ones.	Wang et al. [323]
3. I can differentiate between narrow AI and general AI.	Original
4. I understand how AI is programmed and trained using data.	Original
5. I can program a basic AI application.	Ng et al. [229]
6. I can describe the strengths and weaknesses of current AI systems.	Wang et al. [323]
7. I am aware that AI systems can be unreliable in creative tasks.	Chan and Zhou [68]
8. I understand the concept of neural networks as a representation in AI.	Original
9. I feel confident in explaining how AI systems make decisions using learned patterns.	Lee and Park [192]
10. I am aware that AI decisions can reflect the biases present in its training data.	Chan and Zhou [68]
11. I am aware that AI systems capability and limitations are mediated by what it learns from the data.	Wang et al. [323]
12. I can describe the role humans play in guiding AI systems.	Original
13. I understand how AI systems adjust based on human feedback and input.	Original
14. I believe AI systems are effective at responding to human actions in real-time.	Original
15. I understand how different disciplines contribute to the development of AI.	Original

16. I believe collaboration between different fields is necessary to improve AI-human interaction.	Original
17. I am confident in my ability to interpret the data used in AI training.	Original
18. I understand how AI systems learn from data over time.	Original
19. I can explain the machine learning process in detail.	Original
20. I am able to critically evaluate data used to train AI for biases and inaccuracies.	Hwang et al. [154]
21. I can identify biases in data used for training AI systems.	Lee and Park [192]
22. I understand the role of sensors in AI systems and data collection.	Original
23. I am aware of the ethical challenges and responsibilities in AI development.	Lee and Park [192]
24. I believe that ethical concerns in AI development are adequately addressed by existing regulations.	Original
25. I believe that AI systems need more regulation to prevent unethical usage.	Original

b. Multiple-Choice Items

Item	Source
1. Which of these systems is powered by AI? a) Google Maps b) Calculator c) A regular thermostat d) An analog clock	Original
2. In which of these areas is AI typically is applied? a) Detecting Credit Card Fraud b) Cryptocurrency mining c) Web tracking d) Encryption for instant messaging services	Hornberger et al. [147]

3. Imagine you are chatting with an assistant on the Internet. How could you proceed to find out whether you are interacting with a human or with an AI? a) you can save yourself the effort, since you can no longer distinguish between a human and an AI in written communication b) ask a difficult factual question, since only a human can answer it c) make a few typing errors in your text, then the AI can no longer understand you, but a human can d) make an ironic remark, because this is better understood by humans	Hornberger et al. [147]
4. What is a primary feature of intelligence? a) Following preset commands b) Ability to learn and adapt c) Memorizing large sets of information d) Moving objects with precision	Original
5. What makes AI intelligent? a) AI can walk and talk b) AI has an artificial brain c) AI is at least as intelligent as humans d) AI acts rationally to achieve a particular goal as well as possible	Hornberger et al. [147]
6. Why do AI systems behave intelligently? a) they have no feelings that could distract them from their task b) they think autonomously and pursue their own goals c) they have been programmed to try to achieve a given goal as well as possible d) they are built similar to the human brain and therefore have a similar intelligence	Hornberger et al. [147]
7. Which of the following best describes narrow AI? a) A system designed to solve many types of problems across domains b) A system focused on solving one specific problem c) A hypothetical future AI that mimics human cognition d) An AI that can handle any task equally well	Original
8. In AI, a distinction can be made between "narrow" and "general" AI. "Narrow AI" refers to AI systems that have capabilities in a limited area. "General AI", on the other hand, is said to be capable of a very broad range of tasks, similar to humans. Which of these examples could be considered strong AI? a) intelligent virtual assistant (e.g. Alexa)	Hornberger et al. [147]

b) fully self-driving car c) powerful search engine (e.g. Google) d) strong AI does not exist at the moment	
9. What can weak AI NOT do? a) make decisions under uncertainty b) solve a wide range of tasks c) solve a task better than a human d) learn from unstructured data	Hornberger et al. [147]
10. Which of the following is critical to the successful programming of AI? a) Manual input of decision paths b) Large datasets for training c) Human trainers guiding every decision d) Fixed rule-based logic	Original
11. One limitation of AI is: a) Its inability to store vast amounts of data b) Its lack of creativity and abstract thinking c) Its poor speed in data processing d) Its inability to follow simple rules	Original
12. In which task is AI already superior to humans? a) Detecting tumors b) Programming software c) Translating novels d) Designing cancer therapies	Hornberger et al. [147]
13. In which areas are humans still superior to AI? a) Predict extreme weather events from weather data b) Find a proof for a mathematical theorem c) Answer quiz questions d) Play poker	Hornberger et al. [147]
14. Which of the following data representations is commonly used in AI? a) Flowcharts b) Neural networks c) Pie charts	Original

d) Tables	
15. What are knowledge representations in the field of AI? a) particular representation of knowledge in order to communicate it to humans through AI b) sensors that capture information from the environment c) information about the world that can be processed by a computer d) an algorithm that generates knowledge from data	Hornberger et al. [147]
16. An AI is supposed to divide images into meaningful parts. Input is always an image. Which input-output pair is NOT meaningful: a) Image -> number b) Image -> vector c) Image -> image d) Image -> video	Hornberger et al. [147]
17. AI makes decisions primarily based on: a) Human intuition b) Predefined logical rules c) Patterns learned from data d) Random guessing	Original
18. AI decision-making can be biased due to: a) Poor hardware b) Biased training data c) The programming language used d) Inconsistent network speed	Original
19. AI decision-making differs from human decision-making in that AI: a) Is purely logic-based b) Relies on learned experience without emotional influence c) Can be creative and intuitive d) Makes decisions faster than humans	Original
20. How do AI systems make decisions? a) based on mathematical-logical principles b) based on links defined by programmers c) based on quantum entanglement d) based on artificial intuition	Hornberger et al. [147]

21. What is a key criterion for the quality of a model in machine learning? a) it can predict the output values of the test data as well as possible b) it contains as few variables as possible c) it is as well adapted as possible to the training data d) the predictions are as unambiguous as possible	Hornberger et al. [147]
22. How does supervised learning differ from unsupervised learning? a) in supervised learning, the output values of the training data are known b) in supervised learning, humans must supervise the AI during learning and intervene if necessary c) in supervised learning all computational steps are documented d) in supervised learning, stricter legal regulations apply	Hornberger et al. [147]
23. In a human–AI interaction, the human typically: a) Passively observes AI behavior b) Provides feedback to guide AI c) Replaces AI in decision-making d) Does not interact with AI once it starts functioning	Original
24. How can humans influence the outcome of machine learning? a) calculation of the accuracy of the prediction b) randomized division into test and training data c) selection of the model d) abstraction of the model	Hornberger et al. [147]
25. To what extent can humans influence the outcome of machine learning? a) Humans can hardly influence the result of machine learning, since it runs automatically b) The results can only be influenced when selecting the data the model learns with c) Humans can influence the result during development at several process steps d) Humans can only influence the interpretation of the results	Hornberger et al. [147]
26. A reactive AI system is one that: a) Acts independently without adjusting to external input b) Adjusts its actions based on the user’s input or environment c) Always follows a predefined set of rules d) Operates in a fixed pattern	Original

27. Which discipline provides critical insights into human behavior for AI interaction design? a) Psychology b) Chemistry c) Geography d) Economics	Original
28. Which of the following systems often use AI? a) Flight surveillance systems b) Geo-positioning systems c) 3D printing systems d) Inventory management systems	Hornberger et al. [147]
29. Which of the following interdisciplinary research fields is also a subfield of AI? a) Blockchain b) Natural Language Processing c) Psychology of Learning d) Bioinformatics	Hornberger et al. [147]
30. What is crucial for training a machine learning model? a) Large amounts of diverse data b) High-speed internet c) Simple pre-written rules d) Human supervision at all times	Original
31. What is the benefit of data visualizations? a) Maintaining transparency b) Preparation of training for image recognition c) Communication of results d) Conducting statistical tests	Hornberger et al. [147]
32. Machine learning refers to: a) Feeding pre-programmed instructions to an AI b) AI recognizing patterns in data and improving with experience c) AI making random decisions d) AI memorizing data for future use	Original

33. Why can systems based on machine learning obtain good results? a) their work is often observed by humans and corrected if necessary b) they think similarly to humans, but faster c) they can draw conclusions from large amounts of data and thus improve their model d) they are derived from expert systems in which expert knowledge is stored	Hornberger et al. [147]
34. What data do AI-based recommender systems of streaming services use? a) the data of every user when using the service b) all data that a user of the service leaves on the Internet c) the data of other users, but not one's own d) only one's own data when using the service	Hornberger et al. [147]
35. What is a common issue with training data that leads to biased AI decisions? a) Lack of sufficient data b) Biases present in the dataset c) Using unsupervised learning models d) Slow processing speeds	Original
36. You are testing a machine learning model that is supposed to classify photos of animals. You notice that the model is better at recognizing cats than dogs. What could be the reason for this? a) dogs are more difficult to recognize than cats, since there are fewer images of dogs on the internet b) small objects (cats) are better recognized than large ones (dogs) c) most models are generally better at recognizing cats than dogs d) the training data of the dogs were not representative for all dog breeds	Hornberger et al. [147]
37. What is the first step in developing a machine learning model? a) Selecting an algorithm b) Preprocessing the data c) Training the model d) Evaluating the results	Original
38. Which statement about the steps in the machine learning process is correct? a) the steps of the process are based on behaviorist learning theories b) the development of machine learning is an iterative process	Hornberger et al. [147]

c) the steps of the process of supervised and unsupervised learning are basically the same d) the steps of the process can be run backwards to generate an (artificial) data set	
39. What should be considered in machine learning when dividing the data into training and test data? a) the data should be divided into parts of as equal size as possible b) the data should be randomly divided into training and test data sets c) the test data should be of higher quality than the training data d) the training and test data should be as different from each other as possible	Hornberger et al. [147]
40. AI systems use sensors primarily to: a) Make decisions independently b) Collect real-time data from the environment c) Improve their ethical decision-making d) Communicate with other systems	Original
41. Which of the following is an ethical concern in AI? a) AI speed b) AI bias and fairness c) AI's inability to perform calculations d) AI programming languages	Original
42. AI systems often struggle with ethical decisions because: a) They rely solely on data patterns b) They lack moral reasoning abilities c) They prioritize efficiency over fairness d) All of the above	Original
43. Which ethical principles should be considered when developing AI? a) Holism, fairness, transparency b) Prevention of harm, transparency, fairness c) Fairness, holism, prevention of harm d) Transparency, prevention of harm, holism	Hornberger et al. [147]
44. What is the "black-box" problem? a) AI entails a residual risk that is hard to calculate b) it is often difficult to determine how AI makes decisions	Hornberger et al. [147]

c) users are often not informed about the application of AI d) many users have little knowledge about AI	
45. Which societal challenge is frequently mentioned in the context of AI? a) lack of investment incentives in the educational system b) chip shortage in industry due to the high computational cost of AI c) high error rate in AI-enabled manufacturing d) replacement of human workforce by AI	Hornberger et al. [147]
46. What are central risks in using AI for predictive policing? a) Vulnerability to hacking b) Discrimination against suspects based on origin and status c) Lack of legal certainty in the event of AI breakdown d) Undermining the authority of police officers	Hornberger et al. [147]
47. Which legal challenges do AI applications entail? a) Users of AI have no option for legal protection b) Protecting the rights of AI itself c) Lawyers do not understand the importance of AI d) Limited control of AI because of its autonomy	Hornberger et al. [147]

C.2. Item Sets after Expert Review and Pilot Studies

a. Likert-Scale Items

Likert-scale Item	AI literacy framework dimension
1. I can recognize AI-driven systems across different applications.	Recognize AI
2. I can distinguish between intelligent systems and non-intelligent systems.	Understanding Intelligence
3. I can differentiate between narrow AI and general AI.	General vs. Narrow
4. I understand the process of programming AI applications.	Programmability
5. I understand the process of training AI applications using data.	ML Steps
6. I am capable of programming a basic AI application.	Programmability
7. I can clearly articulate the strengths of current AI systems.	AI's Strength and Weakness
8. I can clearly articulate the limitations of current AI systems.	AI's Strength and Weakness
9. I am aware that AI systems can be unreliable.	AI's Strength and Weakness
10. I have a clear understanding of neural networks as a model in AI.	Representations
11. I can explain how AI systems make decisions based on learned patterns.	Decision-making
12. I am aware that AI decisions can reflect biases present in their training data.	Critically Interpreting Data
13. I recognize that the capabilities and limitations of AI systems are shaped by what they learn from data.	Learning from Data
14. I can describe the role of humans in influencing AI systems.	Human Role in AI
15. I understand how AI systems adapt based on inputs from human.	Action and Reaction
16. I am aware that AI systems can effectively respond to human inputs in real-time.	Action and Reaction

17. I understand how different disciplines contribute to the development of AI as a field.	Interdisciplinarity
18. I recognize the importance of collaboration across disciplines is necessary for improving Human-AI interaction.	Interdisciplinarity
19. I am confident in my ability to interpret data used in AI training.	Data Literacy
20. I am confident in my ability to analyze data used in AI training.	Data Literacy
21. I understand how AI systems evolve by learning from data over time.	Learning from Data
22. I can thoroughly explain the machine learning process and its components.	ML Steps
23. I am capable of critically evaluating training data for inaccuracies.	Critically Interpreting Data
24. I can effectively identify biases in data used to train AI systems.	Critically Interpreting Data
25. I understand the role of sensors in AI systems.	Sensors
26. I recognize the importance of sensors in collecting data for AI systems.	Sensors
27. I am aware of the ethical challenges associated with AI development.	Ethics
28. I am aware of the ethical responsibilities associated with AI development.	Ethics
29. I recognize that existing regulations play a role in addressing ethical concerns in AI development.	Ethics
30. I recognize that AI systems may require stricter regulation to prevent unethical usage.	Ethics

b. Multiple-Choice Items

The associated competency is provided in parentheses after the questions. The correct answer is indicated in italics.

Multiple-Choice Item	AI literacy framework dimension
1. Which of the following systems is powered by AI? a) Google Maps b) Calculator c) A traditional thermostat d) An analog clock	Recognize AI
2. Imagine you are chatting with a virtual assistant. How can you determine whether you are interacting with a human or an AI? a) Assume it's impossible to distinguish between a human and an AI in written communication. b) Ask a challenging factual question, as only a human can answer it. c) Intentionally make a few typing errors; an AI won't understand, but a human will. d) Make an ironic remark, as humans are better at interpreting irony.	Understanding Intelligence
3. What is a primary characteristic of intelligence? a) Following preset commands b) The ability to learn and adapt c) Memorizing large amounts of information d) Moving objects with precision	Understanding Intelligence
4. What convinces us for an AI to be intelligent? a) The ability to walk and talk b) The presence of an artificial brain c) Being at least as intelligent as humans d) Acting rationally to achieve specific goals effectively	Understanding Intelligence
5. Which statement best describes narrow AI? a) A system capable of solving diverse problems across multiple domains b) A system designed to address a single, specific problem c) A speculative future AI that mimics human thought processes d) An AI capable of performing any task as effectively as humans	General vs. Narrow
6. What is something narrow AI is unable to do? a) Make decisions under uncertain conditions	General vs. Narrow

b) Perform a wide range of tasks across different domains c) Solve a task with a greater efficiency than a human d) Learn from unstructured data	
7. What has been the most significant factor in the advancement of AI as a field? a) Manually coding decision pathways b) The availability of large datasets for training AI models c) Continuous human oversight for every AI decision d) The reliance on fixed rule-based algorithms	Learning from Data
8. What is a major limitation of current AI systems? a) An inability to store large volumes of data b) A lack of creativity and abstract reasoning skills c) Slow data processing capabilities d) Difficulty following simple instructions	AI's Strength and Weakness
9. In which task is AI already outperforming humans? a) Accurately detecting tumors in medical imaging b) Programming complex software c) Translating lengthy novels d) Designing innovative cancer treatments	AI's Strength and Weakness
10. In which domain do humans still outperform AI? a) Predicting extreme weather events from data b) Proving mathematical theorems c) Answering factual quiz questions d) Playing poker	AI's Strength and Weakness
11. Which data structure is most commonly used in AI systems? a) Flowcharts that map out decision-making processes b) Neural networks that simulate interconnected brain-like structures c) Pie charts for visualizing datasets d) Tables that organize raw data in rows and columns	Representations
12. How are knowledge representations defined in the context of AI? a) Ways to present knowledge so it is understandable to humans through AI b) Devices that collect data from the environment for AI processing	Representations

c) Structured information about the world that computers can process and use d) Algorithms that derive knowledge directly from datasets	
13. What is the primary foundation for AI decision-making? a) Human intuition b) Predefined logical rules c) Patterns learned from data d) Random guessing	Decision-making
14. What is a major source of bias in AI decision-making? a) Hardware performance b) Training data c) The programming language used to build the AI d) Network speed	Critically Interpreting Data
15. How does AI decision-making differ from human decision-making? a) AI is entirely logic-based b) AI relies on learned patterns and lacks emotional influence in decision-making c) AI is capable of creativity and intuition d) AI always makes faster decisions than humans	Decision-making
16. On what principles do AI systems base their decisions? a) Mathematics and logic b) Predefined relationships set by programmers c) Quantum entanglement d) Artificial intuition	Decision-making
17. What distinguishes supervised learning from unsupervised learning? a) In supervised learning, the output values of the training data are known b) In supervised learning, humans continuously supervise the training process c) In supervised learning, all steps are fully documented d) Supervised learning is subject to stricter legal regulations	Decision-making
18. At which stage can humans influence the outcomes of machine learning? a) By calculating prediction accuracy	Human-Role in AI

b) By randomly splitting data into training and testing sets c) By selecting the model used in the learning process d) By abstracting the structure of the model	
19. To what extent can humans influence the results of machine learning? a) Humans have minimal influence, as the process is fully automated b) Influence is limited to selecting the training data c) Humans can affect results at several stages of the development process d) Humans can only influence the interpretation of results	Human-Role in AI
20. What best describes an AI system? a) Acts independently without adapting to external input b) Responds and adjusts its actions based on user input or environmental changes c) Operates strictly according to predefined rules d) Follows a fixed, unchanging pattern	Action and Reaction
21. Which discipline contributes critical insights into human behavior for designing AI interactions? a) Psychology b) Chemistry c) Geography d) Mathematics	Interdisciplinarity
22. Which interdisciplinary research area is also a subfield of AI? a) Blockchain b) Natural Language Processing c) Psychology of Learning d) Bioinformatics	Interdisciplinarity
23. What is essential for training a machine learning model? a) Large amounts of diverse data b) High-speed internet c) Simple handcrafted rules d) Constant human supervision	ML Steps
24. What is the primary benefit of using data visualizations? a) Storing data effectively	Data Literacy

b) Preparing images for recognition tasks c) Conveying insights and clearly communicating results d) Performing statistical tests	
25. Why do machine learning systems produce strong results? a) They are extremely powerful in computation b) They think like humans, but faster c) They analyze large datasets to identify patterns and improve models d) They are based on expert systems with stored knowledge	Learning from Data
26. You observe that a machine learning model is better at recognizing cats than dogs. What could explain this difference? a) There are fewer online images of dogs compared to cats, making them harder to identify b) Smaller animals like cats are inherently easier for models to recognize than larger ones like dogs c) Most machine learning models are naturally better at recognizing cats d) The training data for dogs was not diverse or representative of all dog breeds	Critically Interpreting Data
27. What is the first step in training a machine learning model? a) Selecting an algorithm b) Preprocessing and cleaning the data c) Training the model d) Evaluating the results	ML Steps
28. Which statement accurately describes the machine learning process? a) It is grounded in behaviorist learning theories b) It is an iterative process that involves repeated refinement c) Supervised and unsupervised learning follow identical steps d) The process can be reversed to generate artificial datasets	ML Steps
29. What should be considered when splitting data into training and test sets? a) Ensure the datasets are equal in size b) Randomly divide the data into training and test sets c) Use higher-quality data for the test set than the training set d) Ensure training and test data are as distinct as possible	ML Steps

30. Why do AI systems use sensors? a) To make decisions independently b) To collect real-time data from the environment c) To enhance their ethical decision-making capabilities d) To communicate with other systems	Sensors
31. Which is a common ethical concern related to AI? a) The speed of AI systems b) Bias and fairness in AI decision-making c) AI's limitations in performing calculations d) The programming languages used to develop AI	Ethics
32. Why do AI systems often struggle with fair and ethical decision-making? a) They rely solely on data patterns b) They lack moral reasoning abilities c) They prioritize efficiency over fairness d) All of the above	Ethics
33. Which ethical principles should guide the development of AI systems? a) Sustainability, justice, and respect for privacy b) Transparency, fairness, and prevention of harm c) Inclusivity, safety, and adaptability d) Responsibility, creativity, and neutrality	Ethics
34. What is the "black-box" problem in AI? a) AI poses unpredictable risks that are difficult to calculate b) It is often unclear how AI systems arrive at their output c) Users are not always informed about the use of AI systems d) Many people lack basic knowledge about AI technologies	Ethics
35. Which societal challenge is frequently associated with AI? a) Lack of investment in education systems b) Shortages of computer chips due to AI's computational demands c) High error rates in AI-powered manufacturing processes d) Displacement of human workers by AI systems	Ethics

C.3. Background and Demographic Questions

1. How old are you?
 - a. 18-24
 - b. 25-34
 - c. 35-44
 - d. 45-54
 - e. 55-64
 - f. 65+
2. What is your gender?
 - a. Woman
 - b. Man
 - c. Non-binary
 - d. Prefer not to say
 - e. Other (please type)
3. In which country do you currently reside?
4. What is the highest level of education you have completed?
 - a. No formal education
 - b. High school diploma or equivalent

- c. Some college, but no degree
 - d. Associates, or technical degree
 - e. Bachelor's degree
 - f. Master's degree
 - g. Doctoral degree
 - h. Other
5. What is the primary field of your educational background?
- a. Arts / Humanities
 - b. Business / Management
 - c. Computer Science / Information Technology
 - d. Education
 - e. Engineering
 - f. Law
 - g. Mathematics
 - h. Medical / Healthcare
 - i. Natural Sciences
 - j. Social Sciences
 - k. Statistics / Data Science

l. Other

6. Have you received any formal education or training specifically related to Artificial Intelligence (AI) or Data Science?
 - a. Yes, I hold a degree specializing in AI or Data Science.
 - b. Yes, I have taken three or more courses in AI or Data Science as part of my degree program.
 - c. Yes, I have taken one or two formal courses related to AI or Data Science as part of my degree program.
 - d. No formal education, but I have gained knowledge in AI or Data Science through on-the-job experience.
 - e. No formal education, but I have self-taught knowledge in AI or Data Science.
 - f. No formal education, but I have attended workshops or seminars on AI or data science.
 - g. No formal education or training related to AI or Data Science.

C.4. Demographic Details of the Development and Validation Sample

The development sample consisted of 288 participants and the validation sample consisted of 188 participants. We summarize the demographic distribution of both samples across age, gender identity, education level, educational background, and prior AI training in Table 16 - 20.

Table 16: Age distribution of participants in the development and validation samples.

Age group	Development Sample (n, %)	Validation Sample (n, %)
<i>18-24</i>	59 (20.5%)	41 (21.8%)
<i>25-34</i>	132 (45.8%)	67 (35.6%)
<i>35-44</i>	51 (17.7%)	37 (19.7%)
<i>45-54</i>	28 (9.7%)	19 (10.1%)
<i>55-64</i>	15 (5.2%)	18 (9.6%)
<i>65+</i>	3 (1.0%)	6 (3.2%)

Table 17: Gender identity distribution of participants in the development and validation samples.

Identity	Development Sample (n, %)	Validation Sample (n, %)
<i>Woman</i>	143 (49.7%)	93 (49.5%)
<i>Man</i>	144 (50%)	95 (50.5%)
<i>Non-binary</i>	1 (0.3%)	-

Table 18: Highest level of education completed by participants in the development and validation samples.

Education Level	Development Sample (n, %)	Validation Sample (n, %)
No formal education	-	1 (0.5%)
High school	40 (13.9%)	17 (9.0%)
Some college, no degree	33 (11.5%)	14 (7.4%)
Associates or technical degree	14 (4.9%)	9 (4.8%)
Bachelors	129 (44.8%)	97 (51.6%)
Masters	63 (21.9%)	41 (21.8%)
Doctorate	9 (3.1%)	9 (4.8%)

Table 19: Primary field of educational background among participants in the development and validation samples.

Field	Development Sample (n, %)	Validation Sample (n, %)
<i>Arts / Humanities</i>	44 (15.3%)	11 (5.9%)
<i>Business / Management</i>	55 (19.1%)	41 (21.8%)
<i>Computer Science / IT</i>	67 (23.3%)	58 (30.9%)
<i>Education</i>	18 (6.2%)	15 (8.0%)
<i>Engineering</i>	20 (6.9%)	15 (8.0%)
<i>Law</i>	14 (4.9%)	5 (2.7%)
<i>Mathematics</i>	5 (1.7%)	3 (1.6%)
<i>Medical / Healthcare</i>	16 (5.6%)	20 (10.6%)
<i>Natural Sciences</i>	18 (6.2%)	6 (3.2%)
<i>Other</i>	9 (3.1%)	2 (1.1%)
<i>Social Sciences</i>	17 (5.9%)	9 (4.8%)
<i>Statistics / Data Science</i>	5 (1.7%)	3 (1.6%)

Table 20: Prior AI training reported by participants in the development and validation samples.

We grouped responses to the item “Have you received any formal education or training specifically related to AI or Data Science?” as follows: *Formal training* = degree in AI/Data Science; three or more AI/Data Science courses; one or two AI/Data Science courses. Informal training = on-the-job experience; self-taught; workshops or seminars. No training = no formal education or training in AI/Data Science.

Training	Development Sample (n, %)	Validation Sample (n, %)
<i>No Training</i>	110 (38.2%)	44 (23.4%)
<i>Informal Training</i>	119 (41.3%)	80 (42.6%)
<i>Formal training</i>	58 (20.1%)	64 (34.0%)

Overall, the development and validation samples were broadly comparable in terms of age and gender distributions, with participants spanning a wide range of educational levels and backgrounds. The validation sample contained a slightly higher proportion of participants with formal AI training and with degrees in technical fields such as computer science and healthcare.

C.5. Ethics Certificate



University of Manitoba | Research Ethics and Compliance

Human Research Ethics - Fort Garry
66 Chancellors Circle
Winnipeg, MB R3T 2N2
humanethics@umanitoba.ca

PROTOCOL APPROVAL

Effective: February 3, 2025

Expiry: February 2, 2026

Principal Investigator: Andrea Bunt
Protocol Number: HE2024-0362
Protocol Title: *AI Knowledge Assessment Study*

Martha Koch, Acting Chair, REB2

Research Ethics Board 2 has reviewed and approved the above research. The Office of Human Research Ethics (OHRE) is constituted and operates in accordance with the current *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans- TCPS 2 (2022)*.

Please note the following important information about your protocol approval:

- i. Approval is granted for the research and purposes described in the protocol only.
- ii. Any changes to the protocol or research materials must be approved by the OHRE **before implementation**.
- iii. Any **deviations** to the research or **adverse events** must be reported to the OHRE immediately through an **REB Event**.
- iv. This approval is valid **for one year only**. A Renewal Request must be submitted and approved prior to the above expiry date.
- v. A **Protocol Closure** must be submitted to the OHRE when the research is complete or if the research is terminated.
- vi. The University of Manitoba may request to audit your research documentation to confirm compliance with this approved protocol, and with the UM [Ethics of Research Involving Humans](#) policies and procedures.

C.6. Study Materials

a. Recruitment Notice

We are researchers at the Human-computer Interaction Lab at University of Manitoba, and we are currently conducting a research study to explore ways to measure people's knowledge of and experiences with Artificial Intelligence (AI) technologies. We welcome users with all levels of AI knowledge and experiences to participate.

As part of the study, you would be asked to complete a questionnaire on your AI knowledge and experiences. The goal is to see how different people respond to questions about AI proficiency. Your participation in the study will take approximately 20 minutes of your time. You will receive a \$6 (Canadian dollars) compensation for your time and effort.

If you would like more information on the study, please contact Ariful Islam Anik, a PhD student at the University of Manitoba. You can also contact the Principal Investigator, Dr. Andrea Bunt.

This research has been approved by the Research Ethics Board at the University of Manitoba, Fort Garry campus.

b. Informed Consent Form



**University
of Manitoba**

Department of Computer Science
Faculty of Science

Discover the unknown + Invent the future

E2-445 EITC

University of Manitoba

Winnipeg, Manitoba, Canada

R3T 2N2

T: 204-474-8313

F: 204-474-7609

Study Title: AI Knowledge Assessment Study

Principal Investigator: Dr. Andrea Bunt, Professor, Department of Computer Science, University of Manitoba, [REDACTED]

Co-Investigators: Md Ariful Islam Anik, Graduate Research Assistant, Department of Computer Science, University of Manitoba, [REDACTED]

Sponsor: Natural Sciences and Engineering Research Council of Canada

Conflicts of Interest and Undue Influence: There are no conflicts of interest associated with this study.

You are being invited to participate in a research study. This consent form, a copy of which you can download or print, is only part of the process of informed consent. If you want more details about something mentioned here, or information not included here, feel free to ask any of the people named above. Please take the time to read this document and any accompanying information carefully. It is very important that you understand:

- what is being asked of you,
- what the risks and benefits of participation are, and
- how the information you provide will be used and stored.

Purpose of the study:

You are invited to participate in a research study on the topic of exploring effective methods for measuring individuals' AI knowledge and experiences and to assess whether these measurements can be used to categorize users based on their AI proficiency. Your participation in this research study will involve filling out a questionnaire we designed for assessing AI knowledge and experiences. If you have any questions or concerns at this time or any time during the study, please feel free to ask the researchers for clarification.

Study Procedures:

The study will take place remotely, with the content presented via Microsoft Forms. After providing consent, you will proceed to the questionnaire. Participation in this study is voluntary and will take approximately 20 minutes of your time. The correctness of your responses will not impact you in any way, but we encourage you to answer thoughtfully and carefully to help us gather meaningful insights. Additionally, you are free to refrain from answering any questions you prefer to omit.

Study Risks:

This survey is minimal risk, meaning that the risks are no greater than what you might encounter in your daily life.

Study Benefits:

The benefits of participating in this research are that you contribute to the development of a measurement tool that can help estimate an individual's AI knowledge and experiences. The questions about AI might also spark your interest in AI and how it impacts our life.

Compensation:

You will receive a \$6 (Canadian dollars) compensation on your prolific account for your time and effort.

Use and Storage of Information:

All the information you provide as a participant in this study is confidential which means we as researchers must keep it safe. We will do our best; however, it is not possible to guarantee absolute confidentiality. We will only share your personal information if required through court order or law (e.g., suspected child or elder abuse).

Only Prolific will see your personal information to provide your honorarium. We will not be able to link the personal information you share with them with the research information you share with us. We will collect your anonymous responses to our questionnaire. Additionally, we will collect some indirectly identifying information, such as age, gender, and education level. The data will be collected through Microsoft Forms and will be securely stored in a privately shared MS Teams on the University of Manitoba servers. Access to the data will be restricted to the research team, who will use university-provided, password-protected email accounts to access it. In addition, the University of Manitoba may look at research records to see that the research is being done in a safe and proper way.

Data Destruction:

All the data will be retained for a maximum period of three years and will be destroyed by December 2027.

Dissemination:

This study is part of co-investigator's PhD thesis, which will be disseminated on MSpace, and we intend to disseminate the results through academic journals and/or conferences. No personal information about you will be included.

Withdrawing:

Your participation in this research study is voluntary. You can choose to do only the activities and/or answer only the questions that you are comfortable with. You may withdraw from the study for any reason. You do not have to explain why. You will not be penalized in any way. Should you desire to withdraw partway through the study, you can do so by closing this tab or window on your browser. If you quit without clicking the submit button at the end of the survey, none of your answers will be saved, so your data will not be used.

in the research. However, to receive your honorarium you must click through to the end of the study and submit it. You do not have to respond to any subsequent questions, except for the final question, which provides an option to withdraw your data. If you decide to withdraw, your data will be destroyed. However, if you do not select the option and submit your responses, due to the nature of the study, we will not be able to remove your information as it cannot be linked back to you.

Summary of Results:

A summary of the results will be made available to the public at the HCI lab website (<http://hci.cs.umanitoba.ca/>). We anticipate providing the summary of results by December 2025. Given that we are not collecting any identifying information, we will not have the ability to send participants the summary directly.

Questions or Concerns:

Designated University of Manitoba personnel (and/or Natural Sciences and Engineering Research Council of Canada) may check that this study is being done safely and properly. To do this, they may review the research records.

This study has been reviewed and approved by a Research Ethics Board at the University of Manitoba. However, this does not mean that participation is risk-free. If you have questions about your rights as a research participant, you may contact the Office of Human Research Ethics [REDACTED]

If you have any questions, concerns, or complaints about this study, you may contact any members of the research team listed on the first page or the Office of Human Research Ethics.

Consent:

By clicking continue below, you indicate that you have understood to your satisfaction the information regarding participation in the research study. By doing this you also confirm that you are at least 18 years of age. In no way does this waive your legal rights nor release the researchers, sponsors, or involved institutions from their legal and professional responsibilities. You are free to withdraw from the study at any time, without prejudice or consequence by simply closing the questionnaire without pressing the final “Submit” button. Once you have submitted your responses you will no longer be able to withdraw your data since your responses are anonymous.

By participating in this study, I agree that:

- I have read the above information.
- I have had the opportunity to ask and have answered all of my questions.
- I understand what is being asked of me.

- I will be taking part in a research study.
- I may freely stop or leave the research study activities at any time.
- My information may be shared outside the University of Manitoba.
- I do not waive my legal rights by participating in the study.

Notice Regarding Collection, Use, and Disclosure of Personal Information

Your personal information is being collected under the authority of The University of Manitoba Act. The University of Manitoba is committed to preserving your right to privacy. The information you provide will be used by the University to support our research. Your personal information will not be used or disclosed for other purposes, unless permitted by *The Freedom of Information and Protection of Privacy Act* or *The Personal Health Information Act*. If you have any questions about the collection of personal information: Ph: 204-474-9462 or Email: fippa@umanitoba.ca

C.7. Likert-Scale Items Analysis

a. *Exploratory Factor Analysis*

We first assessed factorability of the Likert-scale data and determined the number of factors to extract. Next, we examined factor loadings for the full item pool, refined the structure by dropping weak or cross-loading items, and then interpreted the emergent factors. The following subsections (4.1.1–4.1.4) present these stages in detail.

1. *Initial Model Specifications*

- Extraction method: Principal Axis Factoring (PAF)
- Rotation: *Oblimin* (allowing correlated factors)
- Number of factors: 2 (based on parallel analysis, scree plot, MAP test)
- Sample: Development sample (N = 288)

2. *Factor Loadings for all Likert-Scale Items*

Table 21: Factor loadings, communalities (h^2), uniqueness (u^2), and complexity (com) of Likert-scale items in the development sample. The highlighted items were considered weak based on predetermined criteria.

Item	Factor 1	Factor 2	h^2	u^2	com
I can recognize AI-driven systems across different applications.	.49	.36	.56	.44	1.8
I can distinguish between intelligent systems and non-intelligent systems.	.48	.31	.48	.52	1.7
I can differentiate between narrow AI and general AI	.71		.54	.46	1.0
I understand the process of programming AI applications.	.89		.69	.31	1.0

I understand the process of training AI applications using data.	.62		.60	.40	1.3
I am capable of programming a basic AI application.	.88		.63	.37	1.1
I can clearly articulate the strengths of current AI systems.	.63		.59	.41	1.3
I can clearly articulate the limitations of current AI systems.	.52	.34	.57	.43	1.7
I am aware that AI systems can be unreliable.		.63	.30	.70	1.4
I have a clear understanding of neural networks as a model in AI.	.82		.65	.35	1.0
I can explain how AI systems make decisions based on learned patterns.	.64		.65	.35	1.3
I am aware that AI decisions can reflect biases present in their training data.		.74	.48	.52	1.1
I recognize that the capabilities and limitations of AI systems are shaped by what they learn from data.		.78	.54	.46	1.0
I can describe the role of humans in influencing AI systems.		.60	.63	.37	1.5
I understand how AI systems adapt based on inputs from human.		.59	.59	.41	1.4
I am aware that AI systems can effectively respond to human inputs in real-time.		.66	.52	.48	1.0
I understand how different disciplines contribute to the development of AI as a field.	.43	.39	.51	.49	2.0
I recognize the importance of collaboration across disciplines is necessary for improving Human-AI interaction.		.62	.58	.42	1.2
I am confident in my ability to interpret data used in AI training.	.71		.68	.32	1.1
I am confident in my ability to analyze data used in AI training.	.73		.66	.34	1.1

I understand how AI systems evolve by learning from data over time.	.38	.49	.58	.42	1.9
I can thoroughly explain the machine learning process and its components.	.93		.77	.23	1.0
I am capable of critically evaluating training data for inaccuracies.	.74		.61	.39	1.0
I can effectively identify biases in data used to train AI systems.	.58	.32	.62	.38	1.6
I understand the role of sensors in AI systems.	.85		.70	.30	1.0
I recognize the importance of sensors in collecting data for AI systems.	.81		.66	.34	1.0
I am aware of the ethical challenges associated with AI development.		.69	.54	.46	1.0
I am aware of the ethical responsibilities associated with AI development.		.67	.58	.42	1.1
I recognize that existing regulations play a role in addressing ethical concerns in AI development.		.60	.50	.50	1.2
I recognize that AI systems may require stricter regulation to prevent unethical usage.		.74	.45	.55	1.1

Factor Statistics

Factor	SS Loadings	Proportion Variance	Cumulative Variance	Proportion Explained
Factor 1	10.47	0.35	0.35	0.60
Factor 2	6.97	0.23	0.58	0.40

Factor correlation: $r = .51$

Factors score adequacy: Correlation with true factors = .98 (Factor 1), .96 (Factor 2)

3. Factor Loadings after Dropping Weak Items

Table 22: Factor loadings, communalities (h^2), uniqueness (u^2), and complexity (com) of Likert-scale items after dropping weak items.

Item	Factor 1	Factor 2	h^2	u^2	com
I can differentiate between narrow AI and general AI	.70		.53	.47	1.0
I understand the process of programming AI applications.	.88		.69	.31	1.0
I understand the process of training AI applications using data.	.62		.60	.40	1.3
I am capable of programming a basic AI application.	.87		.63	.37	1.1
I can clearly articulate the strengths of current AI systems.	.61		.57	.43	1.3
I am aware that AI systems can be unreliable.		.63	.30	.70	1.3
I have a clear understanding of neural networks as a model in AI.	.81		.65	.35	1.0
I can explain how AI systems make decisions based on learned patterns.	.63		.65	.35	1.4
I am aware that AI decisions can reflect biases present in their training data.		.74	.48	.52	1.0
I recognize that the capabilities and limitations of AI systems are shaped by what they learn from data.		.77	.54	.46	1.0
I can describe the role of humans in influencing AI systems.		.59	.61	.39	1.5
I understand how AI systems adapt based on inputs from human.		.58	.58	.42	1.5
I am aware that AI systems can effectively respond to human inputs in real-time.		.66	.51	.49	1.1

I recognize the importance of collaboration across disciplines is necessary for improving Human-AI interaction.		.61	.56	.44	1.3
I am confident in my ability to interpret data used in AI training.	.71		.68	.32	1.2
I am confident in my ability to analyze data used in AI training.	.72		.66	.34	1.1
I can thoroughly explain the machine learning process and its components.	.92		.78	.22	1.0
I am capable of critically evaluating training data for inaccuracies.	.72		.60	.40	1.0
I understand the role of sensors in AI systems.	.85		.71	.29	1.0
I recognize the importance of sensors in collecting data for AI systems.	.80		.67	.33	1.0
I am aware of the ethical challenges associated with AI development.		.70	.56	.44	1.0
I am aware of the ethical responsibilities associated with AI development.		.68	.59	.41	1.1
I recognize that existing regulations play a role in addressing ethical concerns in AI development.		.60	.50	.50	1.2
I recognize that AI systems may require stricter regulation to prevent unethical usage.		.74	.47	.53	1.1

Factor Statistics

Factor	SS Loadings	Proportion Variance	Cumulative Variance	Proportion Explained
Factor 1	8.45	0.35	0.35	0.60
Factor 2	5.67	0.24	0.59	0.40

Factor correlation: $r = .48$

Factors score adequacy: correlation of scores with factors = .98 (PA1), .96 (PA2)

4. Interpretation of Factors

The final EFA solution suggested a two-factor structure. **Factor 1: Functional Knowledge of AI Systems**, captured items related to programming, training processes, understanding of machine learning, sensors, and the evaluation of training data. **Factor 2: AI in Society**, included items focused on ethics, regulations, human influence, bias, and interdisciplinary collaboration. The naming of the two factors was based on qualitative evaluation of item content and conceptual alignment with prior work on AI literacy. Together, the two factors accounted for 59% of the total variance, with strong factor score adequacy values, supporting the robustness and interpretability of the factor structure.

b. Item Response Theory (IRT) Results for Likert-Scale Factors

To further evaluate item-level functioning, we applied the Graded Response Model (GRM) to the finalized Likert-scale factors. The GRM estimates both discrimination parameters (a), which indicate how well items differentiate between participants with varying levels of competence, and threshold parameters (b_1 – b_6), which reflect the points along the latent trait continuum at which respondents transition between response categories. Results for Factor 1 (Functional Knowledge of AI Systems) and Factor 2 (AI in Society) are presented in Tables 23 and 24, respectively, along with illustrative item characteristic curves in Figure 29 and 30.

1. Factor 1 (Functional Knowledge of AI Systems)

Table 23: Item Response Theory (IRT) parameter estimates for Factor 1 (Functional Knowledge of AI Systems) using the Graded Response Model (GRM).

Factor 1: Functional Knowledge of AI Systems	Discrimination (a)	Thresholds (b1-b6)
I can differentiate between narrow AI and general AI.	2.20	2.30, 0.69, -0.03, -1.47, -3.00, -4.45
I understand the process of programming AI applications.	2.95	2.47, 0.67, -0.74, -2.26, -3.94, -5.90
I understand the process of training AI applications using data.	2.21	3.46, 1.86, 0.73, -0.31, -1.93, -3.96
I am capable of programming a basic AI application.	2.86	0.50, -0.94, -1.76, -2.89, -4.06, -5.56
I can clearly articulate the strengths of current AI systems.	2.06	3.00, 1.55, 0.48, -0.72, -2.54, -4.58
I have a clear understanding of neural networks as a model in AI.	2.85	2.61, 0.75, -0.47, -1.90, -3.50, -5.72
I can explain how AI systems make decisions based on learned patterns.	2.37	3.55, 1.77, 0.66, -0.82, -2.56, -4.28
I am confident in my ability to analyze data used in AI training.	2.49	3.36, 1.88, 0.28, -0.97, -3.14, -5.43
I can thoroughly explain the machine learning process and its components.	3.92	2.15, 0.20, -1.58, -3.24, -5.34, -7.74
I am capable of critically evaluating training data for inaccuracies.	2.43	2.77, 1.17, -0.20, -1.42, -3.35, -5.39
I understand the role of sensors in AI systems.	2.80	1.52, 0.45, -0.68, -1.93, -3.54, -5.18

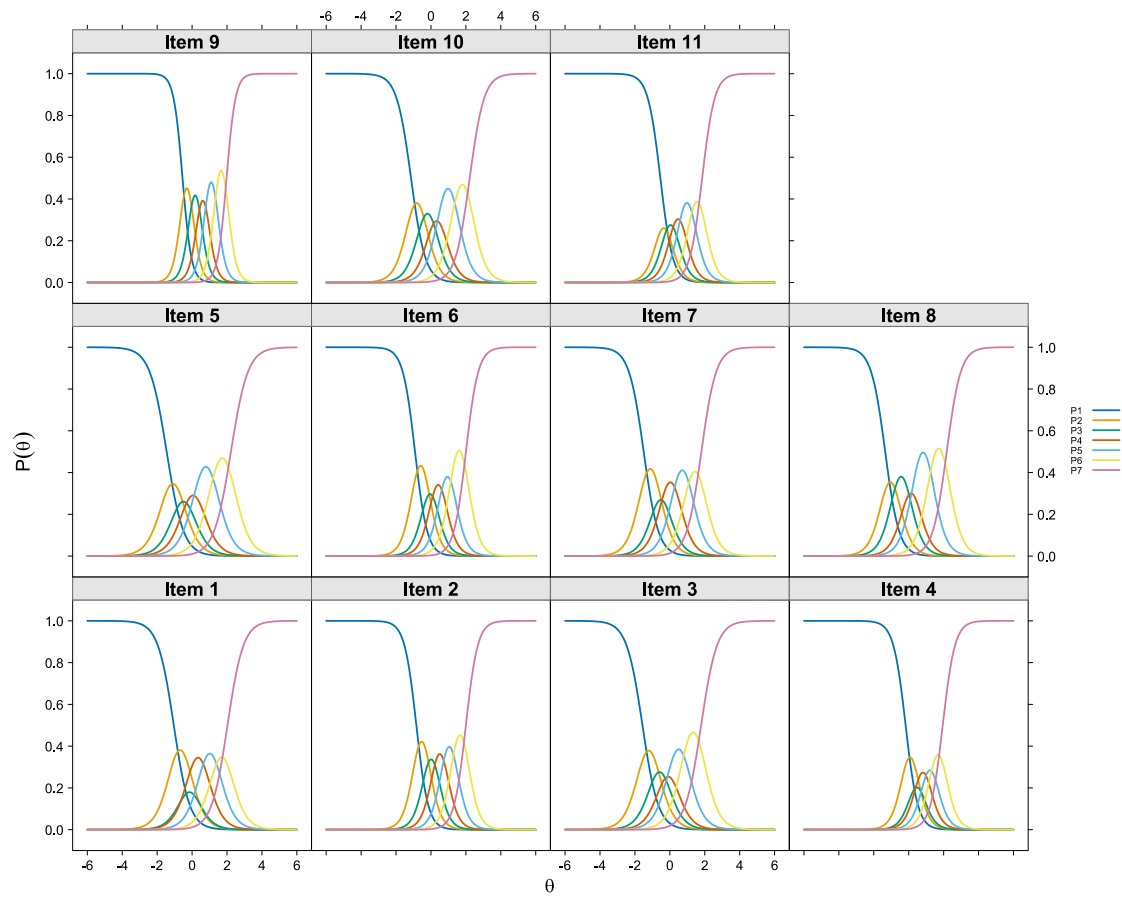


Figure 29: Item characteristic curves for Factor 1 (Functional Knowledge of AI Systems) based on the Graded Response Model.

2. Factor 2 (AI in Society)

Table 24: Item Response Theory (IRT) parameter estimates for Factor 2 (AI in Society) using the Graded Response Model (GRM).

Factor 2: AI in Society	Discrimination (a)	Thresholds (b1-b6)
I am aware that AI decisions can reflect biases present in their training data.	1.76	5.06, 3.41, 2.88, 1.71, 0.28, -1.84
I recognize that the capabilities and limitations of AI systems are shaped by what they learn from data.	2.15	7.05, 4.83, 3.78, 2.19, 0.58, -1.77
I am aware that AI systems can effectively respond to human inputs in real-time.	2.33	6.74, 4.68, 3.36, 1.94, -0.15, -2.29
I recognize the importance of collaboration across disciplines is necessary for improving Human-AI interaction.	2.19	5.90, 4.22, 2.96, 1.66, -0.28, -2.62
I am aware of the ethical challenges associated with AI development.	1.79	5.50, 3.87, 2.70, 1.55, -0.06, -2.16
I recognize that existing regulations play a role in addressing ethical concerns in AI development.	1.82	5.22, 3.58, 2.44, 0.91, -0.76, -2.73
I understand how AI systems adapt based on inputs from human.	2.59	5.55, 4.05, 2.66, 1.40, -0.78, -3.13
I am aware that AI systems can be unreliable. *	1.01	6.16, 3.99, 3.55, 2.13, 1.01, -0.23
I can describe the role of humans in influencing AI systems.	2.56	6.15, 4.11, 2.71, 0.96, -0.95, -3.21

Note – The starred item was removed after IRT analysis due to low discrimination power.

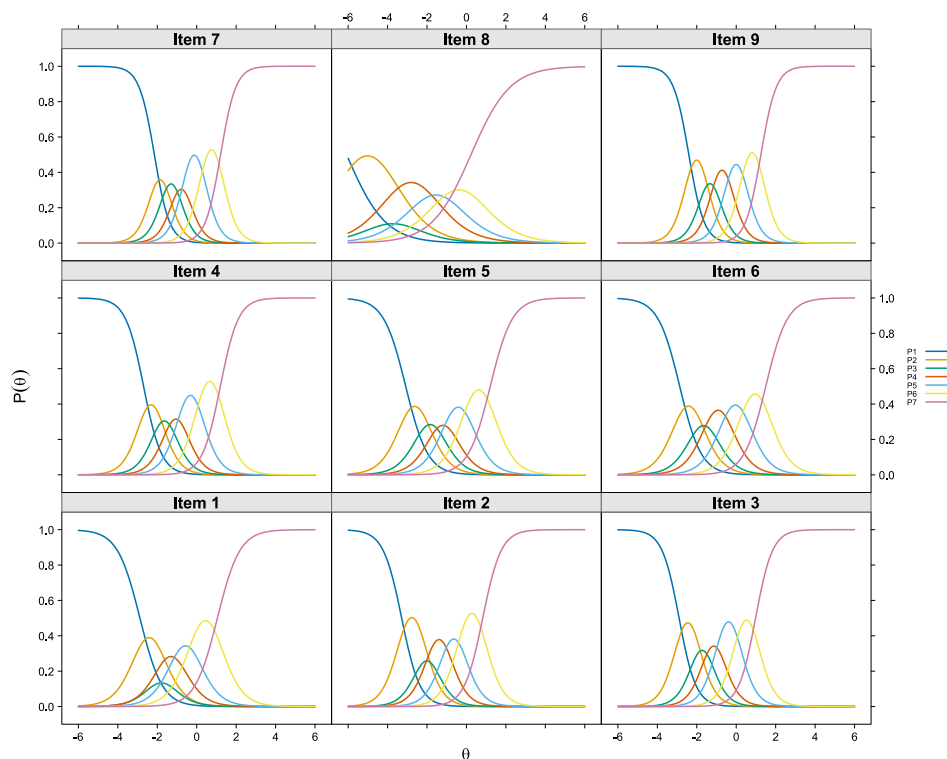


Figure 30: Item characteristic curves for Factor 2 (AI in Society) based on the Graded Response Model.

C.8. Multiple-Choice Items Analysis

We applied Item Response Theory (IRT) models to both the development and validation samples. These analyses allowed us to estimate item parameters, assess model fit, and examine the extent to which individual items reliably captured variation in participants' factual AI understanding.

a. Item Response Theory (IRT) Results for Multiple-Choice Items on Development Sample

We first examined the development sample using the 3-parameter logistic (3PL) model, which estimates item discrimination (a), difficulty (b), and guessing (g) parameters. This model is well suited to multiple-choice items because it accounts for the probability of

answering correctly by chance. Table 25 reports the parameter estimates for all items, Figure 31 presents the corresponding item characteristic curves, and Table 26 summarizes item-level model fit.

Table 25: Item Response Theory (IRT) parameter estimates for multiple-choice items in the development sample (discrimination, difficulty, and guessing parameters).

Item	Discrimination (a)	Difficulty (b)	Guessing (g)
Which of the following systems is powered by AI?	4.07	-0.15	0.85
Imagine you are chatting with a virtual assistant. How can you determine whether you are interacting with a human or an AI?	0.56	-0.13	0.00
What is a primary characteristic of intelligence?	1.61	-1.10	0.39
What convinces us for an AI to be intelligent?	1.09	-0.85	0.00
Which statement best describes narrow AI?	1.57	-1.20	0.00
What is something narrow AI is unable to do?	1.68	-0.32	0.16
What has been the most significant factor in the advancement of AI as a field?	4.41	-0.37	0.45
What is a major limitation of current AI systems?	0.66	-1.47	0.00
In which task is AI already outperforming humans?	3.25	0.70	0.16
In which domain do humans still outperform AI?	1.21	1.56	0.10
Which data structure is most commonly used in AI systems?	1.94	0.34	0.34
How are knowledge representations defined in the context of AI?	18.44	1.05	0.29
What is the primary foundation for AI decision-making?	0.77	-3.20	0.01
What is a major source of bias in AI decision-making?	3.96	-0.17	0.43

How does AI decision-making differ from human decision-making?	1.46	-1.55	0.00
On what principles do AI systems base their decisions?	-0.16	-1.32	0.00
What distinguishes supervised learning from unsupervised learning?	4.17	1.35	0.28
At which stage can humans influence the outcomes of machine learning?	3.34	0.89	0.50
To what extent can humans influence the results of machine learning?	1.27	-0.48	0.07
What best describes an AI system?	0.92	-1.20	0.00
Which discipline contributes critical insights into human behavior for designing AI interactions?	1.58	0.10	0.46
Which interdisciplinary research area is also a subfield of AI?	11.82	0.78	0.47
What is essential for training a machine learning model?	3.02	-0.78	0.34
What is the primary benefit of using data visualizations?	2.36	0.08	0.33
Why do machine learning systems produce strong results?	1.38	-1.40	0.00
You observe that a machine learning model is better at recognizing cats than dogs. What could explain this difference?	1.30	-1.34	0.00
What is the first step in training a machine learning model?	1.26	-0.80	0.01
Which statement accurately describes the machine learning process?	1.55	0.30	0.14
What should be considered when splitting data into training and test sets?	1.49	1.25	0.10
Why do AI systems use sensors?	2.04	0.47	0.46
Which is a common ethical concern related to AI?	1.92	-1.63	0.00

Why do AI systems often struggle with fair and ethical decision-making?	1.36	-0.16	0.34
Which ethical principles should guide the development of AI systems?	9.27	0.76	0.48
What is the “black-box” problem in AI?	5.39	0.70	0.36
Which societal challenge is frequently associated with AI?	1.83	-1.03	0.00

Note - Discrimination (a) = how well the item differentiates between participants of different ability levels. Difficulty (b) = location on the ability scale (θ) where the item is most informative. Guessing (g) = indicates the probability of answering correctly by chance.

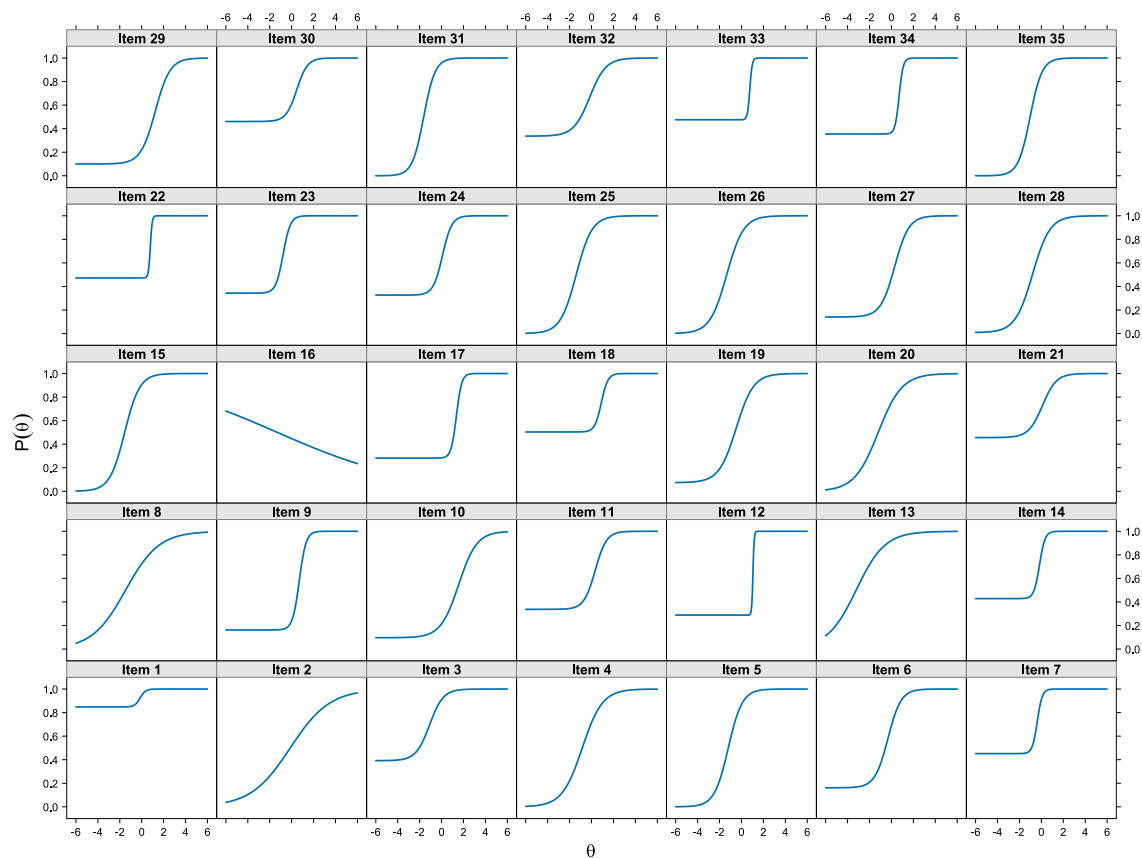


Figure 31: Item characteristic curves (ICCs) for the 35 multiple-choice items in the development sample based on the 3-parameter logistic model. Each curve shows the probability of a correct response across levels of latent ability (θ), illustrating item discrimination, difficulty, and guessing behavior.

Table 26: Item-level model fit statistics for multiple-choice items in the development sample (S - X^2 , df , RMSEA, and p -values).

Item	S - X^2	df	RMSEA	p -value
Which of the following systems is powered by AI?	11.79	8	0.04	.16
Imagine you are chatting with a virtual assistant. How can you determine whether you are interacting with a human or an AI?	15.93	20	0.00	.72
What is a primary characteristic of intelligence?	23.07	16	0.04	.11
What convinces us for an AI to be intelligent?	12.67	18	0.00	.81
Which statement best describes narrow AI?	19.03	15	0.03	.21

What is something narrow AI is unable to do?	21.27	17	0.03	.22
What has been the most significant factor in the advancement of AI as a field?	19.41	13	0.04	.11
What is a major limitation of current AI systems?	18.57	20	0.00	.55
In which task is AI already outperforming humans?	13.72	16	0.00	.62
In which domain do humans still outperform AI?	17.80	17	0.01	.40
Which data structure is most commonly used in AI systems?	33.53	19	0.05	.02*
How are knowledge representations defined in the context of AI?	34.23	19	0.05	.01*
What is the primary foundation for AI decision-making?	17.56	17	0.01	.42
What is a major source of bias in AI decision-making?	26.64	14	0.06	.02*
How does AI decision-making differ from human decision-making?	19.53	15	0.03	.19
On what principles do AI systems base their decisions?	29.36	23	0.03	.17
What distinguishes supervised learning from unsupervised learning?	21.66	20	0.02	.36
At which stage can humans influence the outcomes of machine learning?	18.01	20	0.00	.59
To what extent can humans influence the results of machine learning?	19.77	17	0.02	.29
What best describes an AI system?	33.35	19	0.05	.02*
Which discipline contributes critical insights into human behavior for designing AI interactions?	19.64	19	0.01	.42
Which interdisciplinary research area is also a subfield of AI?	11.61	19	0.00	.90
What is essential for training a machine learning model?	10.69	12	0.00	.56
What is the primary benefit of using data visualizations?	25.51	17	0.04	.08
Why do machine learning systems produce strong results?	13.00	15	0.00	.60
You observe that a machine learning model is better at recognizing cats than dogs. What could explain this difference?	25.04	17	0.04	.09

What is the first step in training a machine learning model?	31.90	18	0.05	.02*
Which statement accurately describes the machine learning process?	15.34	16	0.00	.50
What should be considered when splitting data into training and test sets?	18.02	17	0.01	.38
Why do AI systems use sensors?	17.43	19	0.00	.56
Which is a common ethical concern related to AI?	10.03	11	0.00	.53
Why do AI systems often struggle with fair and ethical decision-making?	39.89	18	0.07	.002**
Which ethical principles should guide the development of AI systems?	27.76	19	0.04	.09
What is the “black-box” problem in AI?	19.27	18	0.02	.38
Which societal challenge is frequently associated with AI?	21.73	13	0.05	.06

*Note. $S-X^2$ = item-level chi-square fit statistic, df = degrees of freedom, RMSEA = root mean square error of approximation, p -value = significance of $S-X^2$ test. Items with significant misfit ($p < .05$) are flagged with **

b. Item Response Theory (IRT) Results for Multiple-Choice Items on Validation Sample

We then applied the same 3PL model to the validation sample to confirm the stability of the multiple-choice items’ psychometric properties. Table 27 provides parameter estimates, Figure 32 presents the item characteristic curves, and Table 28 reports model fit statistics. Together, these results highlight which items maintained strong performance across samples and which items may require further refinement.

Table 27: Item Response Theory (IRT) parameter estimates for multiple-choice items in the validation sample (discrimination, difficulty, and guessing parameters).

Item	Discrimination (a)	Difficulty (b)	Guessing (g)
What is one of the primary characteristics of intelligence?	0.94	-2.54	0.00
What convinces us for an AI to be intelligent?	0.92	-1.25	0.00
Which statement best describes narrow AI?	2.34	-0.91	0.00
What is something narrow AI is unable to do?	2.61	-0.24	0.20
In which task is AI already outperforming humans?	8.51	0.49	0.10
In which domain do humans still outperform AI?	4.17	0.73	0.20
Which data structure is most commonly used in AI systems?	6.10	0.33	0.59
How does AI decision-making differ from human decision-making?	2.21	-1.11	0.02
What distinguishes supervised learning from unsupervised learning?	6.66	0.54	0.25
At which stage can humans influence the outcomes of machine learning?	15.85	0.39	0.46
What is essential for training a machine learning model?	5.89	-0.37	0.57
What is the primary benefit of using data visualizations?	3.99	0.02	0.29
Why do machine learning systems produce strong results?	1.06	-1.76	0.01
You observe that a machine learning model is better at recognizing cats than dogs. What could explain this difference?	1.96	-0.91	0.00
Which statement accurately describes the machine learning process?	2.16	0.03	0.24
What should be considered when splitting data into training and test sets?	2.85	0.55	0.04

Which is a common ethical concern related to AI?	1.90	-0.98	0.25
Which societal challenge is frequently associated with AI?	1.52	-1.14	0.00

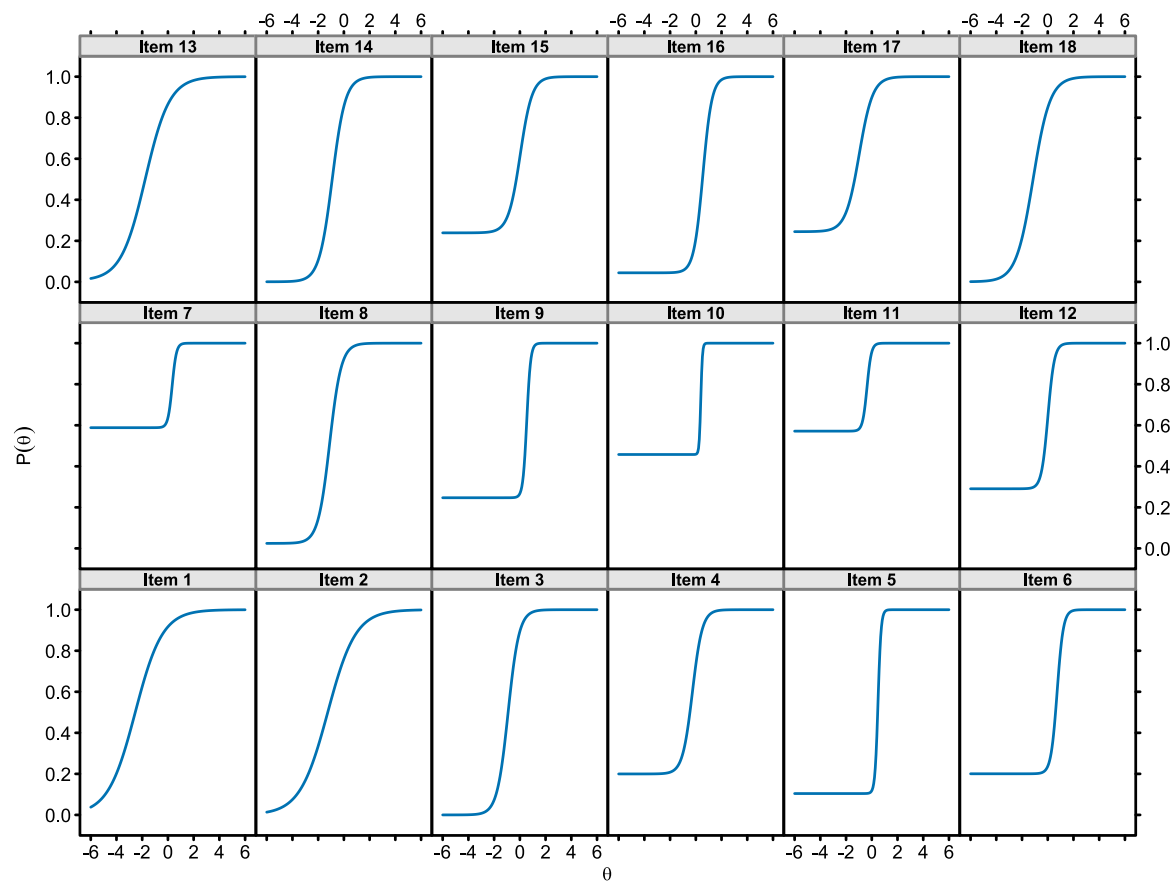


Figure 32: Item characteristic curves (ICCs) for multiple-choice items in the validation sample based on the 3-parameter logistic model.

Table 28: Item-level model fit statistics for multiple-choice items in the validation sample ($S-X^2$, df, RMSEA, and p-values).

Item	$S-X^2$	df	RMSEA	p-value
What is one of the primary characteristics of intelligence?	7.83	7	0.025	.35
What convinces us for an AI to be intelligent?	9.56	11	0.000	.57
Which statement best describes narrow AI?	9.49	7	0.044	.22
What is something narrow AI is unable to do?	6.18	8	0.000	.63
In which task is AI already outperforming humans?	6.81	6	0.027	.34
In which domain do humans still outperform AI?	11.61	9	0.039	.24
Which data structure is most commonly used in AI systems?	15.77	9	0.063	.07
How does AI decision-making differ from human decision-making?	4.60	6	0.000	.60
What distinguishes supervised learning from unsupervised learning?	4.33	9	0.000	.89
At which stage can humans influence the outcomes of machine learning?	9.55	9	0.018	.39
What is essential for training a machine learning model?	6.06	6	0.007	.42
What is the primary benefit of using data visualizations?	17.46	9	0.071	.04
Why do machine learning systems produce strong results?	14.11	9	0.055	.12
You observe that a machine learning model is better at recognizing cats than dogs. What could explain this difference?	8.99	7	0.039	.25
Which statement accurately describes the machine learning process?	19.70	11	0.065	.05
What should be considered when splitting data into training and test sets?	6.11	7	0.000	.53
Which is a common ethical concern related to AI?	12.76	8	0.056	.12
Which societal challenge is frequently associated with AI?	6.27	9	0.000	.71

C.9. Measurement Invariance Testing for Likert-Scale Factors

We tested measurement invariance of the two-factor Likert structure across gender and prior AI training groups. Establishing invariance is critical for ensuring that observed differences reflect true differences in different groups. Following standard practice, we sequentially evaluated configural, metric, and scalar models and compared changes in fit indices (ΔCFI , RMSEA, SRMR) as well as chi-square difference tests.

Table 29: Multi-group measurement invariance testing across gender.

Model	χ^2	df	CFI	ΔCFI	TLI	RMSEA	SRMR
Config	670.35	298	.879	–	.861	.115	.076
Metric	692.40	315	.877	.002	.866	.113	.095
Scalar	705.53	332	.878	–.001	.875	.109	.095

Interpretation: Configural, metric, and scalar invariance were all supported across gender ($\Delta\text{CFI} < .01$, nonsignificant χ^2 difference tests), indicating that the factor structure, loadings, and intercepts were equivalent for men and women.

Table 30: Multi-group measurement invariance testing across prior AI training groups

Model	χ^2	df	CFI	ΔCFI	TLI	RMSEA	SRMR
Config	909.87	447	.808	–	.779	.129	.095
Metric	955.90	481	.803	.005	.790	.126	.115
Scalar	1018.33	515	.791	.012	.792	.125	.129
Partial Scalar*	1005.60	513	.795	.008	.795	.124	.128

Note – One intercept (“I am aware that AI systems can effectively respond to human inputs in real-time.”) was freed to restore fit.

Interpretation: Across prior AI training groups, configural and metric invariance were supported. Scalar invariance initially failed ($\Delta\text{CFI} = .012$, $p = .011$). Allowing one intercept to vary improved model fit ($\Delta\text{CFI} = .008$, nonsignificant χ^2 difference), supporting *partial scalar invariance*. This indicates valid latent mean comparisons across training groups with only minor item-level non-invariance.

C.10. Differential Item Functioning for Multiple-Choice Items

We conducted Differential Item Functioning (DIF) analyses of the multiple-choice items. DIF testing evaluates whether individual items operate differently across groups with the same underlying ability, which could indicate potential bias. Using likelihood-ratio tests within the 3PL IRT framework and controlling false discovery rate (FDR), we compared item functioning across gender and training groups.

Table 31: Differential item functioning (DIF) analyses of multiple-choice items across gender and training groups using 3PL IRT likelihood-ratio tests with FDR correction.

Grouping	Method	Items Flagged	Notes
Gender	3PL-IRT (LR, FDR)	0	No evidence of DIF across gender.
Training	3PL-IRT (LR, FDR)	0	All items functioned equivalently across training groups.

Interpretation: DIF analyses showed that the multiple-choice items functioned consistently across gender and training groups.

C.11. Regression Analysis

To further examine the relationship between perceived and actual competence, we conducted regression analyses predicting calibration mismatch from self-assessed competence while controlling gender and prior AI training. We estimated robust regression models (HC3 standard errors) using both raw and standardized self-assessment scores as predictors, allowing us to assess the stability of effects across different scaling approaches.

Table 32: Regression of calibration mismatch on self-assessed competence (raw score model) with robust standard errors (HC3)

Predictor	B	SE	95% CI	t	<i>p</i>
(Intercept)	−2.97	0.17	–	−17.12	< .001
Self-assess	0.68	0.05	[0.58, 0.77]	13.61	< .001
Gender (men = 1)	−0.19	0.09	–	−2.07	.039
Training (informal)	0.11	0.11	–	0.96	.339
Training (formal)	0.14	0.17	–	0.83	.406

Interpretation: The effect of self-assessed competence was positive and highly significant ($b = 0.68$, 95% CI [0.58, 0.77], $p < .001$), indicating that participants with lower self-assessments underestimated their competence, while those with higher self-assessments overestimated. Gender showed a small negative effect, suggesting men tended slightly more toward underestimation when holding self-assessment and training constant.

Table 33: Regression of calibration mismatch on standardized self-assessed competence (z-score model) with robust standard errors (HC3).

Predictor	B	SE	95% CI	t	<i>p</i>
(Intercept)	0.02	0.10	–	0.16	.873
Self-assess (z)	0.84	0.06	[0.72, 0.96]	13.61	< .001
Gender (men = 1)	–0.19	0.09	–	–2.07	.039
Training (informal)	0.11	0.11	–	0.96	.339
Training (formal)	0.14	0.17	–	0.83	.406

Interpretation: Self-assessed competence (standardized) was a strong positive predictor of calibration mismatch ($b = 0.84$, 95% CI [0.72, 0.96], $p < .001$). Gender showed a small but significant negative effect ($b = -0.19$, $p = .039$), indicating that men were slightly more likely to underestimate their competence when self-assessment and training were held constant. Training level effects were not statistically significant.