Advancing Pitfall-Aware Explanation Design for Human-Centered AI

Ariful Islam Anik

Department of Computer Science, University of Manitoba, anikmai@myumanitoba.ca

Andrea Bunt

Department of Computer Science, University of Manitoba, andrea.bunt@umanitoba.ca

Explanations are commonly proposed and deployed to improve the transparency and interpretability of AI systems, and are often assumed to increase trust, system understanding, and informed decision-making. However, human-centered evaluations show that explanations can also introduce unintended negative consequences, referred to as explainability pitfalls. In this position paper, we argue that explanations should be designed with likely pitfall pathways in mind, rather than waiting for pitfalls to surface through post-hoc evaluations. Drawing on prior work and empirical insights from our study with training dataset explanations, we illustrate that explanations can appear successful by conventional evaluation metrics while increasing cognitive effort and potentially creating conditions for inflated understanding and inappropriate reliance. We propose a pitfall-aware design agenda that makes explanation boundaries explicit through warning labels and stating non-goals, ensures explanations lead to meaningful user actions, and embeds pitfall awareness into organizational practices.

CCS CONCEPTS • Human-centered computing~Human computer interaction (HCI)~HCI design and evaluation

Additional Keywords and Phrases: Transparency, Training Dataset Explanations, Explainability Pitfalls, Pitfall-aware Explanations

ACM Reference Format:

Ariful Islam Anik and Andrea Bunt. 2026. Advancing Pitfall-Aware Explanation Design for Human-Centered AI. In *IUI MIRAGE 2026 workshop held at 31st International Conference on Intelligent User Interfaces, March 23, 2026, Paphos, Cyprus.* ACM, New York, NY, USA, 9 pages.

1 INTRODUCTION

To improve the interpretability and transparency of AI systems, the field of explainable AI (XAI) has developed a wide range of explanation techniques [1,2,6,11,16,17,22,23]. Explanations are commonly framed as mechanisms for increasing trust [36,42], supporting understanding [1,37], and enabling informed decision-making [11,53]. Yet, evidence from human-centered evaluations suggests that explanations do not always deliver these benefits. In some settings, explanations have little or no measurable impact [17,21,50], and in others they can prompt users to act against their own interests by fostering misplaced trust in systems and overreliance [18,26,34]. These unanticipated and unintended negative consequences are defined as explainability pitfalls [25].

Ehsan and Reidl distinguish explainability pitfalls from dark patterns of explainability [19] by intent [25]. While dark patterns are deliberately designed to manipulate users [19,29], explainability pitfalls emerge unintentionally [25], often as side effects of well-meaning transparency efforts. Prior literature has identified a range of such pitfalls, including uncalibrated trust and overreliance, misinterpretation of what explanations imply, increased cognitive load and frustration, and explanation designs that encourage passive consumption rather than action [25,41]. As AI explanations become more widely deployed, HCI and XAI researchers have increasingly called for deeper investigation of these pitfalls, including when they arise, how they manifest in practice, and what design strategies can mitigate them before

they reach users [24,25]. Despite growing calls to consider pitfalls at design time, they are still commonly approached as problems to be detected after deployment, typically surfaced through post hoc evaluation practices [27,43].

Our recent work on training dataset explanations, conditionally accepted at IUI 2026, helps illustrate these concerns. Training dataset explanations [6] are a data-centric approach that build upon the dataset documentation literature [33], communicating important information about datasets, such as the motivation, creation, composition, intended uses, distribution, and maintenance of a dataset. As part of a PhD dissertation (in progress) [5], we first conducted a scoping review (a method that 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 [8]) to synthesize how explainability pitfalls have been characterized in existing literature. Building on this review, we then empirically examined how information depth, an explanation design choice, can shape user outcomes by comparing summary and detailed training dataset explanations, alongside a Progressive Disclosure [57] manipulation (conditionally accepted at IUI 2026). Our results showed that detailed explanations increased perceived trust and understanding while also increasing cognitive load, and that participants consistently preferred the detailed explanations despite the added effort. Although we did not observe a clear pitfall in the form of excessive cognitive load, the findings point to subtler pitfall pathways. For example, under common evaluation metrics, the detailed explanations would appear unequivocally successful, yet they also create conditions in which users may overestimate how well they understand the system, consistent with patterns of illusion of explanatory depth [18].

This observation echoes a broader concern that pitfalls can coexist with positive subjective reactions and may emerge through inflated understanding, excessive cognitive load, or misplaced confidence rather than through obvious observable failures. As a result, relying only on post-hoc evaluations risks missing important pathways through which explanations can introduce negative consequences.

In this position paper, we argue that explainability pitfalls should be treated as a forethought in designing explanations rather than an evaluation artifact to be identified post-hoc. We highlight several core directions toward pitfall-aware explanations, including anticipating common pitfall pathways during design, supporting meaningful user actions, and embedding pitfall awareness into organizational processes that extend beyond individual design decisions. Together, these directions motivate a broader design agenda for explanations that promote transparency while reducing the risk of unintended consequences.

2 BACKGROUND

Explainability pitfalls describe unintended negative downstream effects that can arise from AI explanations, even when those explanations are designed to promote transparency [25]. To understand what types of explainability pitfalls have been identified and how they appear, we conducted a scoping review [8] of using search terms such as "negative consequences/outcomes/impact of AI/ML explanations," "explainability pitfalls," and "harms of AI explanations." Our research included forward and backward citation tracking. We focused on literature that explicitly discussed negative consequences of explanations, as well as those that described how explanations could inadvertently result in such consequences. Our synthesis revealed two broad categories of explainability pitfalls.

The first category concerns the unintended impact of explanations on users, including uncalibrated trust [25,34,40], overreliance [10,13,40,50], high cognitive load [37,57,58], misinterpretation [7], and frustration [25,40,41]. For example, users may develop uncalibrated trust [34,40], accepting AI outputs without sufficient critical evaluation if the explanation appears clear or convincing, leading to overreliance [14,49]. Overly detailed or complex explanations can

contribute to high cognitive load [37,53], overwhelming users and leading to frustration [25], misinterpretation [7], and even overreliance [15].

The second category is directly related to explanation design, including distracting interfaces [41], designs that promote passive consumption of information [41], lack of actionability [25,40,41], excessive information volume and complexity [37,41], and ambiguity [41]. For example, if explanations are not designed with the users' information or visual literacy in mind [9,55], they can lead to misinterpretation and improper trust in the system. Similarly, a lack of actionability can leave users unsure of how to apply the information to improve outcomes [23,39], which can cause frustration [25].

Overall, this body of work highlights that explainability pitfalls arise both from how explanations shape users' perceptions and from how explanations are designed and presented. Importantly, many pitfalls stem from well-intentioned design choices rather than obvious failures, motivating closer attention to explanation design decisions that appear successful on the surface.

3 EMPIRICAL INSIGHTS FROM STUDYING TRAINING DATASET EXPLANATIONS

From our scoping review of explainability pitfalls, information depth (i.e., the level of detail and complexity included in the content of an explanation) emerged as a recurring design factor with the potential to introduce explainability pitfalls. Prior work has shown mixed effects of information depth on outcomes such as understanding, trust, and preference. Simpler explanations are often preferred [44,46,52] but may lead to incomplete understanding [53], while more detailed explanations can support richer mental models [17,37] at the cost of increased cognitive effort [37]. Despite this growing literature, little is known about how these trade-offs manifest in training dataset explanations [6], which differ from many other explanation types in both content and length.

In our study, conditionally accepted at IUI 2026, we treated information depth as an explanation design factor (which can also serve as a potential pitfall pathway) and examined Progressive Disclosure (i.e., a UI strategy that reveals explanations on demand [57]) as a candidate strategy for managing the risks associated with increased depth. To study this, we compared summary and detailed training dataset explanations that communicated the same core content at different levels of depth, and examined whether presenting information with or without Progressive disclosure can manage the cognitive load. We also evaluate how information depth and Progressive Disclosure shaped users' understanding, perceptions, and cognitive load. Participants used these explanations to assess AI systems and then completed questionnaires assessing perceived trust, fairness, understanding, learning, and cognitive load, followed by semi-structured interviews. Figure 1 illustrates the combinations of explanation depth and Progressive Disclosure examined in the study.

Our findings from the study with 32 participants revealed a clear trade-off associated with information depth. Detailed explanations led to higher perceived trust, fairness, understanding, and learning than summary explanations, but also resulted in significantly higher cognitive load. Importantly, participants consistently preferred the detailed explanations despite the added effort. Progressive Disclosure did not reliably mitigate the cognitive costs of increased depth, nor did it meaningfully affect trust or perceived understanding, although it did support perceived learning. In the next section, we interpret these patterns through the lens of explainability pitfalls and discuss what they imply for pitfall-aware explanation design.

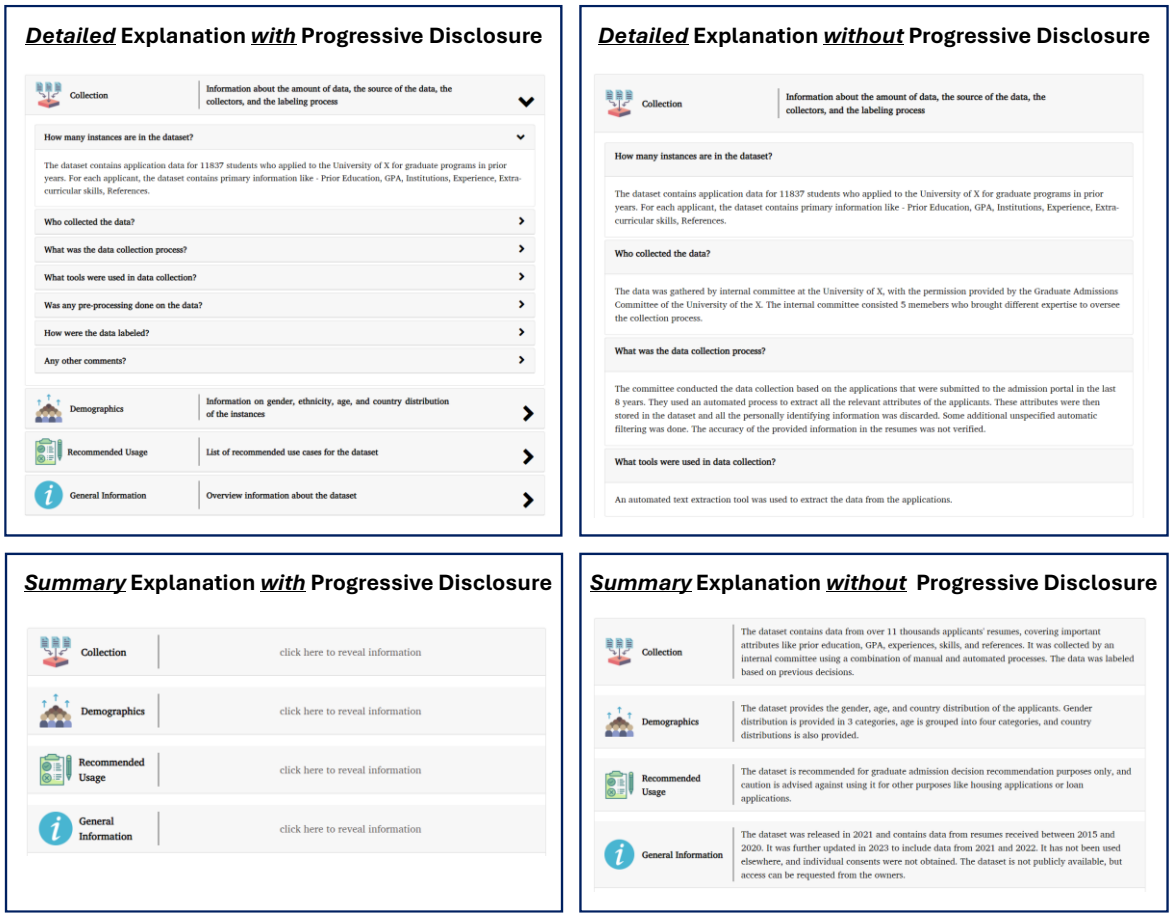

Figure 1: Illustrative snippets from the four training dataset explanation variants.

4 REFLECTION THROUGH THE LENS OF EXPLAINABILITY PITFALLS

Although this position paper does not present any new empirical findings beyond those reported in our IUI 2026 paper (conditionally accepted), it offers a new interpretation of those results through the lens of explainability pitfalls. On the surface, participants' higher perceived trust, understanding, and a clear preference for detailed explanations despite the increased cognitive load suggest that users value comprehensiveness over brevity. However, this pattern also raises less visible concerns about how explanations that feel successful can still create pitfall pathways.

While we did not observe a clear pitfall in the form of excessive cognitive load, which remained moderate even for detailed explanations (mean = 3.5 on a 7-point Likert scale), the findings still surface risks about perceived understanding. We attempted to assess objective understanding using a critique task (i.e., asking participants users to assess a system's advantages and disadvantages [7]) and by evaluating the accuracy of their critiques. However, critique accuracy is a limited proxy for understanding, as it reflects whether participants can articulate plausible strengths and weaknesses, which may not capture deeper comprehension. It is also challenging to evaluate critique accuracy reliably, since participants drew different, yet defensible, conclusion from the same information. In contrast, perceived understanding primarily reflects confidence in understanding, which can increase even when objective understanding does not. This

misalignment can foster an illusion of explanatory depth [18], in which users feel they understand the system more deeply than they do. This inflated sense of understanding can, in turn, create conditions for overreliance, particularly in contexts where users must make repeated or high-stakes decisions.

Additionally, we observed a pattern in the study that participants selectively engaged with parts of the detailed explanations, while overlooking others, suggesting another potential pitfall pathway. Such selective engagement is consistent with confirmation bias [47], where participants attend to explanation elements that support their initial impressions. From the outside, the accompanying increases in perceived understanding and trust can make the explanation appear successful. At the same time, this pattern can reflect shallow engagement and contribute to the formation of an incomplete mental model. Conversely, summary explanations sometimes elicited uncertainty and doubt, pointing to a different risk where insufficient detail undermines users' ability to reason critically about the system.

As our study was scenario-based and did not require participants to make any consequential decisions, the practical effects of inflated understanding or misplaced confidence may only emerge in sustained or higher-stakes contexts. Still, these observations highlight the value of considering potential pitfall pathways during design and evaluation, motivating the directions we outline next: designing explanations with forethought, supporting meaningful user actions, operationalizing pitfall-aware practices, and evaluating explanations to anticipate pitfall pathways.

5 PATH FORWARD: DESIGNING PITFALL-AWARE EXPLANATIONS

5.1 Pitfalls as a Forethought Rather Than an Afterthought

A promising step toward pitfall-aware explanation design is to treat explainability pitfalls as a design input rather than an afterthought. Designers should explicitly anticipate how explanations could be misunderstood, inappropriately trusted, or cognitively burdensome, even when they are technically sound. Doing so, however, requires practical support for identifying likely failure modes before explanations are built (i.e., seamful design [24]). One way to operationalize this is to provide pitfall taxonomies that map common explanation goals [20,48,59] (e.g., increasing trust or improving understanding) to known pitfalls (e.g., overreliance [49] or the illusion of explanatory depth [18]). Such tools can prompt designers to ask, "what could go wrong if this explanation works too well?" alongside standard design questions about usefulness and clarity. A systematic review (i.e., a means of evaluating and interpreting available research relevant to a particular research question, topic area, or phenomenon of interest [35]) of explainability pitfalls and how they manifest in practice can serve as a starting point for developing such resources.

Another proactive approach could be using warning labels [62] that communicate the scope, limits, and potential dangers of an explanation. Similar to hazard labels in other domains [12,28,63], these labels can clarify what an explanation is intended to support and what it should not be used for. For example, a training dataset explanation could include a short label stating that the information is intended to support high-level assessment of data coverage and limitations, but not to verify the correctness or fairness of individual predictions. Further, it could warn against the possible risks such as miscalibrated trust. More broadly, explicitly stating non-goals [54] of explanations (e.g., that they do not guarantee correctness, ensure fairness, or replace human judgment) can make explanation boundaries visible and reduce the chance that users treating explanations as broader justifications than intended.

5.2 Explanations Need to End in Action

Many explainability pitfalls stem from explanations that inform without enabling actions [25,40,41]. 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 or inappropriate reliance [49]. 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 interface.

In our own study, we did not implement these action pathways because our goal was to isolate how information depth and Progressive Disclosure shape users' understanding and perceptions in a controlled setting. We instead treat actionability as a key next design problem for data-centric explanations, one that can be addressed by pairing explanatory insights with plausible next steps at the point of use [4,33,56]. 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. A practical heuristic for assessing actionability could be to ask whether an explanation makes at least one appropriate next step visible [31].

5.3 Operationalizing Pitfall-Aware Practices

Pitfall-aware explanation design requires organizational support, since many pitfalls emerge over time and across contexts rather than during initial interface design. Teams can conduct contextual pitfall audits to examine how explanations behave under conditions such as high-stakes systems or time pressure [30]. These audits do not guarantee correctness, but surface plausible risk pathways and guide mitigation choices [45,51]. Teams can also conduct periodical pitfall reviews, similar to privacy [64] or accessibility reviews [32,38], where explanation designs are checked for known risk patterns and assumptions about users and contexts. Insights from audits and reviews can be documented in a shared pitfall registry, so that lessons learned can inform future explanation work.

5.4 Evaluation to Anticipate Pitfall Pathways

Finally, evaluation practices should increasingly aim to anticipate pitfall pathways rather than identifying them only after deployment. Beyond perceived trust and understanding, one promising direction is to examine calibration between users' confidence and their objective understanding [3]. However, designing objective checks that are valid and resistant to superficial strategies remains an open challenge. In our study, the critique task could only serve as a proxy for understanding. More commonly used comprehension probes can also inadvertently capture reading effort rather than understanding [60], and counterfactual interpretation tasks [61] can be difficult to standardize across explanation types and domains. Despite these challenges, lightweight objective checks can still be useful if they are tailored to the explanation's intended use and designed to discourage superficial strategies. For example, comprehension probes can test key limitations rather than recall, and counterfactual tasks can be standardized around a small set of comparable changes. Used iteratively during design, these methods can surface potential pitfalls while mitigation is still feasible.

6 CONCLUSION

This position paper argues for treating explainability pitfalls as a forethought in explanation design rather than an afterthought in evaluation. Drawing on prior work and our empirical insights from a study with training dataset explanations, we show that explanations can appear successful by common metrics while still creating conditions for inflated understanding and inappropriate reliance. To address this gap, we propose a pitfall-aware design agenda that makes explanation boundaries explicit, ensures explanations support meaningful user actions, and embeds pitfall awareness into organizational practices. Together, these directions aim to move human-centered XAI toward explanation systems that support transparency while remaining robust to unintended consequences.

REFERENCES

[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] 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

[5] Ariful Islam Anik. 2026. Investigating Explanations that Target Training Data. PhD Dissertation, University of Manitoba. (In progress)

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] Anders Boman, Mécia Miguel, Ida Andersson, and Daniel Slunge. 2024. The effect of information about hazardous chemicals in consumer products on behaviour – A systematic review. *Science of The Total Environment* 947: 174774. https://doi.org/10.1016/j.scitotenv.2024.174774

[13] Zana Buçinca, 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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*.

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

[26] 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

[27] Alessandro Gambetti, Qiwei Han, Hong Shen, and Claudia Soares. 2025. A Survey on Human-Centered Evaluation of Explainable AI Methods in Clinical Decision Support Systems. https://doi.org/10.48550/arXiv.2502.09849

[28] Gerald J. Gorn, Anne M. Lavack, Craig R. Pollack, and Charles B. Weinberg. 1997. An Experiment in Designing Effective Warning Labels. *Health Marketing Quarterly* 14, 2: 43–61. https://doi.org/10.1300/J026v14n02_05

[29] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3174108

[30] 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

[31] Rex Hartson. 2003. Cognitive, physical, sensory, and functional affordances in interaction design. *Behaviour & Information Technology* 22, 5: 315–338. https://doi.org/10.1080/01449290310001592587

[32] W3C Web Accessibility Initiative (WAI). WCAG-EM Overview: Website Accessibility Conformance Evaluation Methodology. *Web Accessibility Initiative (WAI)*. Retrieved December 17, 2025 from https://www.w3.org/WAI/test-evaluate/conformance/wcag-em/

[33] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2023. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Computing Surveys* 55, 5: 1–29. https://doi.org/10.1145/3527848

[34] 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

[35] Barbara Kitchenham and others. 2004. Procedures for performing systematic reviews. *Keele, UK, Keele University* 33, 2004: 1–26.

[36] 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

[37] 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

[38] Jonathan Lazar, Daniel Goldstein, Anne Taylor, Todd Green, Lindsay Lawrence, and Greg Harris. 2015. *Ensuring digital accessibility through process and policy*. Morgan Kaufmann, Amsterdam, Netherlands.

[39] 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

[40] 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

[41] 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

[42] 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

[43] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. 2022. The Conflict Between Explainable and Accountable Decision-Making Algorithms. In *2022 ACM Conference on Fairness Accountability and Transparency*, 2103–2113. https://doi.org/10.1145/3531146.3534628

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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

[49] Samir Passi and Mihaela Vorvoreanu. Overreliance on AI Literature Review.

[50] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman 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

[51] 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

[52] 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

[53] 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

[54] H. Schulzrinne, H. Tschofenig, J. Morris, J. Cuellar, J. Polk, and J. Rosenberg. 2007. *Common Policy: A Document Format for Expressing Privacy Preferences*. RFC Editor. https://doi.org/10.17487/rfc4745

[55] Jane Secker and Emma Coonan (eds.). 2012. *Rethinking Information Literacy: A Practical Framework for Supporting Learning*. Facet. https://doi.org/10.29085/9781856049528

[56] 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

[57] 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

[58] 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

[59] 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

[60] Joseph K Torgesen. Research related to strengthening instruction in reading comprehension: Part 2.

[61] 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

[62] Michael S. Wogalter, Stephen W. Jarrard, and S. Noel Simpson. 1992. Effects of Warning Signal Words on Consumer-Product Hazard Perceptions. *Proceedings of the Human Factors Society Annual Meeting* 36, 13: 935–939. https://doi.org/10.1177/154193129203601303

[63] Michael S. Wogalter, Christopher B. Mayhorn, and Kenneth R. Laughery. 2021. WARNINGS AND HAZARD COMMUNICATIONS. In *HANDBOOK OF HUMAN FACTORS AND ERGONOMICS* (1st ed.), Gavriel Salvendy and Waldemar Karwowski (eds.). Wiley, 644–667. https://doi.org/10.1002/9781119636113.ch24

[64] 2023. The Digital Privacy Playbook - Privacy Impact Assessments. Retrieved December 17, 2025 from https://www.canada.ca/en/government/system/digital-government/digital-privacy-playbook/privacy-impact-assessments.html