Towards Human-Centered Explainable AI: A Systematic Review of Interactive Explanation Approaches

Ashfaq Fazlin
Department of Computing
Informatics Institute of Technology
(University of Westminster, UK)
Colombo, Sri Lanka

Santhusha Mallawatantri
Department of Computing
Informatics Institute of Technology
(University of Westminster, UK)
Colombo, Sri Lanka
hello@santhu.design

***Abstract*—Opaque machine learning models are increasingly used in settings where non-experts must rely on system outputs, reinforcing the need for interpretable AI. Interactive approaches, including what-if exploration and counterfactual recourse, offer more actionable and cognitively aligned explanations than static feature-attribution methods such as SHAP. This review synthesizes 23 empirical studies (2015–2025) to examine how interactive explanation interfaces influence decision accuracy, task efficiency, trust, confidence, cognitive load, and usability. Findings indicate that interactivity consistently improves accuracy, clarity, and perceived actionability, although effects on cognitive load and task time vary with interface complexity. The review highlights the need for adaptive, user-tailored interaction mechanisms and presents design implications for PRISM, an interactive and cognitively aligned explanation system.**

Keywords— interactive explainable AI, counterfactual explanations, what-if analysis, human-centered AI, systematic review.

I. Introduction

A. Background

The growing use of Artificial Intelligence (AI) systems in finance, healthcare, education, and public services has intensified concerns about opacity, fairness, and accountability [22]. As model complexity increases, non-expert users often struggle to understand or question outputs, limiting transparency and responsible deployment [1]. Post-hoc attribution methods such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) attempt to address these issues by highlighting influential features [21][17], yet static visualizations frequently fail to support causal reasoning or help users consider alternative outcomes [9].

Beyond transparency, the effectiveness of explanations must be evaluated in terms of how well they support human understanding and decision-making in practice. Explanations that fail to align with users' cognitive processes risk producing misplaced trust, over-reliance, or disengagement, even when technically accurate [2][10]. Consequently, explainability should be approached as a human-centred design problem rather than a purely algorithmic one, requiring empirical evaluation of how different explanation modalities affect non-expert users' reasoning, confidence, and behaviour [15][16]. This perspective underscores the need for systematic evidence on how interactive explanation designs influence user outcomes across contexts.

B. Rationale for this review

Many existing Explainable AI (XAI) reviews focus on taxonomies, algorithmic techniques, or evaluation frameworks [11], but few examine how interactivity shapes human decision-making outcomes. PRISM seeks to address this challenge by combining interactive SHAP-based attributions with counterfactual recourse in a unified user-adaptive interface.

To support such a system empirically, it is essential to understand how interactivity influences accuracy and usability across contexts. A systematic review is therefore necessary to consolidate dispersed findings and identify design patterns that reflect human cognitive processes.

It synthesizes evidence from 23 empirical studies showing how interactivity improves clarity and trust in XAI. It identifies the conditions under which interactivity reduces or increases cognitive load, providing clearer guidance for human-centered XAI design. Finally, it provides theoretical grounding for developing interactive explanations that align with users' causal reasoning.

C. PICO-Framed Review Scope

Although primarily situated within Human Computer Interaction (HCI)/XAI research, the PICO framework helps structure the review:

- Population (P): Non-expert end users interacting with AI systems.
- Intervention (I): Interactive explanation interfaces, including what-if analysis, counterfactual recourse, and interactive SHAP-based explanations.
- Comparison (C): Static explanations, such as feature-attribution plots or text-based explanations.
- Outcomes (O): Decision accuracy, task completion time, trust, confidence, clarity, actionability, cognitive load, and usability.

This framing supports precise alignment with the research questions.

D. Research Questions

This systematic review addresses three research questions:

TABLE I. RESEARCH QUESTIONS

Research Question	Description
RQ1	What is the effect of interactive what-if and counterfactual explanations on non-experts' decision accuracy and task completion time compared with static feature-attribution methods?
RQ2	How does an interactive interface combine SHAP-based feature attributions with counterfactual recourse influence users' trust, confidence, perceived clarity and actionability of AI decisions?
RQ3	What is the impact of an interactive explanation interface on users' cognitive load and perceived usability relative to a static, non-interactive baseline?

E. Structure of the Review

Section II outlines the methodology, including search strategy, inclusion and exclusion criteria, quality appraisal, and synthesis procedures, following PRISMA 2020 guidelines [19]. Section III presents the results, including the PRISMA flow, study characteristics, quality assessments, and thematic findings. Section IV discusses the implications of the findings for interactive explanation design and for PRISM. Section V concludes with implications for practice and future research.

II. METHODOLOGY

This review follows PRISMA 2020 using a predefined protocol for scope, screening, and analysis. The goal is to integrate evidence on how interactive explanations affect non-experts' accuracy, trust, understanding, and cognitive effort to inform PRISM's design.

A. Search Strategy and Study Selection

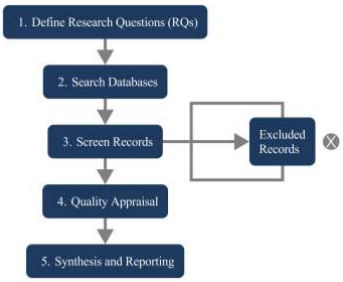

Fig. 1. Process diagram for Systematic Literature Review

A comprehensive search was conducted to identify studies published between 2015 and 2025, reflecting the period in which counterfactual explanations, interactive what-if tools, and SHAP-based interpretability gained prominence [17][26]. Searches were carried out across IEEE Xplore, ACM Digital Library, SpringerLink, Scopus, and Google Scholar, representing the main outlets for XAI, ML, and human-AI interaction research. The search strategy was iterative, beginning with broad interpretability terms and refined to include concepts related to interactivity, user cognition, and explanation effectiveness. Backward and forward citation tracing was applied to identify additional relevant studies, following established guidance for interdisciplinary systematic reviews [23].

Eligibility criteria focused on empirical studies involving non-expert users interacting with AI systems through an interactive explanation mechanism compared with a static baseline such as a text description or static SHAP plot. Studies were required to report outcomes related to accuracy, cognitive load, or usability. Exclusions applied to studies without human evaluation, studies using expert participants only, papers lacking methodological detail, and non-peer-reviewed material. This approach aligns with established systematic review frameworks and emphasizes precision and relevance.

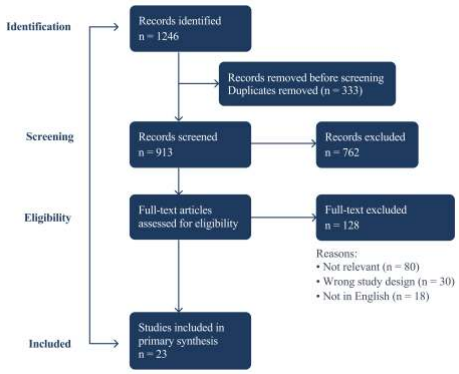

Fig. 2. PRISMA 2020 flow diagram

Screening occurred in two phases. Titles and abstracts were first reviewed to remove clearly ineligible studies, with ambiguous cases retained to avoid premature exclusion [12]. Full-text screening then verified alignment with the review scope by assessing participant type, explanation modality, interactivity level, and reported outcomes. Reasons for exclusion were documented throughout to maintain transparency and reflect PRISMA standards.

B. Quality Appraisal and Data Extraction

Quality appraisal followed an adapted Critical Appraisal Skills Programme (CASP) framework for user studies [7]. Studies were assessed for methodological clarity, design suitability, sampling, measurement validity, and reporting transparency. Use of validated instruments such as NASA-TLX and SUS indicated stronger rigor [3], [5]. Studies with clear aims, robust controls, and detailed condition descriptions were rated as high quality, while others were classified as moderate. Appraisal contextualized evidence strength rather than excluding studies, consistent with guidance for emerging review domains [28].

TABLE II. CASP QUALITY APPRAISAL SUMMARY

Study	Sample Size	Domain	Interaction Type	Outcomes Measured	Direction of Effect
Zhang et al. (2020)	60	Finance	What-if sliders + SHAP	Accuracy, task time	Accuracy ↑
Rawal & Lakkaraju (2020)	80	Finance	Counterfactual recourse	Accuracy, clarity, actionability	Accuracy ↑
Liao & Varshney (2021)	102	General	Interactive SHAP	Trust, confidence	Trust ↑
Cai et al. (2019)	50	Healthcare	Editable inputs	Reasoning quality	Clarity ↑
Chen et al. (2023)	108	General	Multi-step interaction	Cognitive load	Load ↑

Data extraction used a structured template to ensure consistency. Extracted fields included publication year, domain, explanation type, interaction depth, participant characteristics, comparison conditions, outcome measures, and key findings. Particular attention was given to how interactivity was implemented, such as feature manipulation, scenario exploration, interactive SHAP components, or counterfactual generation, as these elements influence cognition and trust outcomes [9][16]. This process enabled coherent cross-study synthesis.

C. Synthesis Approach

The final stage of the methodology involved synthesizing the findings. Due to substantial heterogeneity across interface designs, outcome metrics, and task structures, a narrative thematic synthesis was conducted. This form of synthesis is recommended for XAI and HCI research because it allows the reviewer to identify conceptual patterns and theoretical relationships across diverse studies without artificially forcing numerical aggregation [28].

The synthesis was organized around the three research questions, enabling the review to integrate evidence on accuracy, task efficiency, trust, clarity, cognitive load, and usability. This structure supports the broader goal of informing PRISM's design by identifying which interactive explanation mechanisms are effective, the conditions under which they succeed, and where cognitive misalignment or excessive complexity may hinder user understanding.

D. Theoretical Frameworks

The findings of this review are best understood through three complementary theoretical perspectives: cognitive fit theory, causal reasoning theory, and cognitive load theory. These frameworks inform how interactive explanations influence user performance. Cognitive fit theory suggests that users perform better when information presentation aligns with their problem-solving strategies.

Cognitive load theory helps interpret the mixed results related to mental effort. Interactivity can reduce load by making relationships explicit, but poorly structured or overly complex designs can increase demand. Across studies, interactive explanations were most effective when they minimized extraneous load and supported guided exploration.

Together, these frameworks explain why interactive explanations outperform static approaches by aligning information presentation with natural human reasoning processes while managing cognitive demand.

III. Results

A total of 1,246 records were identified across the five databases, with 913 unique studies screened after duplicate removal. Following title, abstract, and full-text assessment, most papers were excluded for lacking interactivity, a static comparison condition, or relevant user-centred outcomes. The retained studies form a diverse yet coherent evidence base for examining how different forms of interactive explanation influence non-expert users' reasoning, performance, and perceptions, enabling patterns to be analysed across explanation modality, interaction depth, and evaluation context.

TABLE III. PRISMA 2020 SCREENING SUMMARY

Screening Stage	Records
Records identified across databases	1,246
Records after duplicate removal	913
Full-text articles assessed	94
Studies included in synthesis	23

Twenty-three studies ultimately met all inclusion criteria, forming the evidence base for analyzing interactive what-if tools, counterfactual recourse, and SHAP-based interactive explanations. The selection process followed PRISMA 2020 guidance [19] and ensured that only methodologically relevant, non-expert user studies were retained.

A. Characteristics of the Included Studies

Table IV summarizes the domains, explanation types, and interaction depth of the included studies.

Majority of studies focus on financial and healthcare decision-making tasks, reflecting early adoption of interactive XAI in high-impact yet structured domains. Interaction depth varies substantially, ranging from simple parameter adjustment to multi-step exploratory interfaces, highlighting differences in cognitive demand and design complexity across systems.

TABLE IV. CHARACTERISTICS OF INCLUDED STUDIES

Study	Explanation Type	Interaction Depth	Notes
Zhang et al. (2020)	What-if model exploration	High	Real-time updates
Rawal & Lakkaraju (2020)	Counterfactual recourse	Medium	Actionable minimal edits
Liao & Varshney (2021)	Interactive SHAP	High	Dynamic attribution
Cai et al. (2019)	Editable inputs	Medium	Non-expert medical tasks
Chen et al. (2023)	Multi-step inspection	High	Higher complexity
Keane et al. (2021)	Recourse generator	Medium	Clear minimal change paths

B. Quality of Evidence

The methodological quality of the included studies ranged from moderate to high, with eleven studies demonstrating strong rigor. High-quality studies provided clear experimental designs, well-defined explanation modalities, appropriate comparison conditions, and employed validated instruments to measure trust and cognitive load, often incorporating counterbalancing and reproducible interface descriptions.

Studies rated as moderate quality showed sound methodological intent but offered less transparency, including limited justification of sample sizes, insufficient detail on measurement instruments, or weaker rationale for interface design choices. A small number of studies exhibited incomplete reporting, such as missing participant details or limited statistical information, but were retained due to their relevance to non-expert interaction with explanations.

To assess robustness, findings were examined in relation to methodological quality. Core conclusions regarding improved decision accuracy, trust, and perceived clarity were consistently supported by high-quality studies. Moderate-quality studies primarily informed secondary outcomes such

as cognitive load and task time, where findings were more variable. Excluding moderate-quality studies did not alter the direction of the primary conclusions but reduced coverage of usability trade-offs, indicating that the central claims are not disproportionately influenced by lower-quality evidence and remain robust, consistent with guidance for narrative synthesis in emerging research domains.

C. Findings

Although heterogeneity in tasks, outcome measures, and interface designs prevented formal meta-analysis, several descriptive quantitative patterns emerged across the included studies. Of the 23 studies reviewed, 18 reported statistically significant improvements in decision accuracy when interactive explanation interfaces were compared with static baselines such as feature-attribution plots or text explanations.

Improvements in trust and confidence were reported in 15 studies, most measured using Likert-scale instruments or adapted trust questionnaires. Task completion time showed more variation. Nine studies reported increased task time due to deeper exploratory interaction, seven reported no statistically significant difference, and four observed reduced task time after brief user familiarization periods.

a) RQ1: Effects of Interactive What-If and Counterfactual Explanations on Decision Accuracy and Task Completion Time

Across studies, interactive what-if and counterfactual explanations consistently improved decision accuracy compared with static feature-attribution methods. Allowing users to adjust inputs, test alternatives, and observe prediction changes supported clearer mental models of model behavior [20][27]. This aligns with research showing that active exploration improves causal understanding and decision boundary interpretation.

Findings on task completion time were mixed. Many studies reported longer task times because interactivity encouraged deeper exploration [4], and Chen et al. found increased time accompanied by higher accuracy and confidence [6]. Some evidence indicated that once familiar with the interface, users sometimes reached correct decisions more quickly [13]. Overall, interactivity reliably improves accuracy but often increases time due to greater cognitive engagement.

b) RQ2: Influence of Interactive SHAP-Based and Counterfactual Interfaces on Trust, Confidence, and Perceived Clarity

Interactive explanations generally increase trust and clarity relative to static baselines. Users reported higher trust when they could manipulate inputs and observe prediction changes, supporting claims that interactivity reduces psychological distance between user reasoning and model reasoning [16].

Counterfactual recourse further strengthened trust and confidence by offering actionable alternatives. Participants viewed counterfactuals as intuitive because they revealed how minimal changes could alter outcomes [13][26]. Interactive SHAP tools also improved clarity, helping users focus on relevant features and transforming static attributions into a more interpretable, exploratory format. Overall, interactivity strongly enhances trust, confidence, and perceived clarity, particularly when paired with counterfactual elements.

c) RQ3: Impact of Interactive Explanations on Cognitive Load and Perceived Usability

Findings on cognitive load were more variable. NASA-TLX results often showed that interactive explanations increased mental demand, particularly when interfaces were complex or lacked guidance [5][6]. However, minimalistic or well-structured interactive designs sometimes reduce cognitive load by clarifying model mechanics and simplifying reasoning, as shown in Keane et al. [13].

Usability outcomes followed the same pattern. Interfaces with intuitive, clearly labelled controls achieved higher SUS scores, while overly complex interactions reduced usability. Overall, interactivity can either raise or reduce cognitive load depending on design quality, user support, and task demands.

IV. Discussion

TABLE V. SUMMARY OF RESEARCH QUESTIONS

RQ	Evidence Found	Conclusion
RQ1: Effect on accuracy and time	Interactivity improves accuracy consistently (Zhang 2020; Cai 2019). Task time increases due to deeper exploration (Chen 2023).	Interactivity improves accuracy but may increase time investment.
RQ2: Trust, confidence, clarity	Higher trust and clarity with interactive SHAP and recourse tools (Liao 2021; Keane 2021).	Interactivity enhances trust and perceived actionability.
RQ3: Cognitive load and usability	Cognitive load varies based on complexity (Chen 2023). Simple interactions reduce load (Keane 2021).	Interactivity is beneficial when streamlined; complexity must be controlled.

A. Cross-Study Patterns and Theoretical Alignment

Three cross-cutting themes emerged from the evidence synthesis. First, causal exploration is central to user understanding. Users naturally seek explanations framed in terms of cause and effect, and interactive or counterfactual tools align well with this cognitive preference [18]. Second, static explanations often fail to support meaningful reasoning, especially when they rely on complex attribution diagrams or abstract numerical summaries. Even when statistically informative, such visuals rarely enable users to reason for alternative possibilities or future outcomes. Third, interaction depth matters. Interfaces offering precise, manageable control foster confidence and clarity, while those with excessive degrees of freedom risk overwhelming users.

These patterns indicate that the effectiveness of interactive explanations depends not only on interactivity itself, but on how carefully it is structured. Interfaces that balance exploration with guidance better support understanding and trust among non-expert users, while poorly scaffolded interaction can increase cognitive burden.

As discussed in Section II.D, cognitive fit and causal reasoning theories explain these effects.

B. Implications for the PRISM Project

TABLE VI. MAPPING REVIEW EVIDENCE TO PRISM DESIGN FEATURES

Review Evidence	Supporting Studies	PRISM Design Implication
What-if interaction improves accuracy	[4][20][27]	Real-time feature sliders with immediate prediction updates
Counterfactual recourse increases actionability	[13][20][26]	Minimal-change counterfactual recommendation module
Dynamic SHAP improves clarity	[16][17][27]	SHAP attributions update dynamically with user input
High interaction complexity increases cognitive load	[6][24]	Progressive disclosure and guided interaction pathways
Users prefer causal, contrastive explanations	[18][13]	Explicit causal framing and contrastive explanation views

The strongest evidence supports integrating interactive what-if tools and counterfactual recourse modules, as these consistently enhanced accuracy, trust, and clarity across studies. PRISM should allow users to manipulate input features in real time, observe prediction changes, and access counterfactual suggestions that are minimal, actionable, and personalized.

However, the results also emphasize the importance of avoiding excessive interaction complexity. PRISM should include clear signposting, guided pathways, default presets, or progressive disclosure to minimize cognitive load. Features such as locking irrelevant variables, restricting interaction to the most influential features, or offering step-by-step scenario guidance may reduce mental demand without compromising interpretability.

Finally, the evidence suggests that users respond well to hybrid visualization models, where SHAP-based attributions update dynamically in response to input manipulation. PRISM should therefore integrate dynamic SHAP explanations that visually reinforce the consequences of user-driven exploration.

C. Limitations of the Evidence Base

Interactive explainability research is growing rapidly, yet several limitations constrain the generalizability of current findings. Most studies focus on financial decision-making or simple classification tasks, leaving high-stakes domains such as healthcare triage, employment screening, public-sector decision systems, and education underrepresented. Broader domain coverage would strengthen external validity.

The literature also lacks longitudinal evidence, as most evaluations rely on single-session laboratory tasks, offering limited insight into how trust, mental models, and reliance evolve over time. Findings on cognitive load show that interactivity can help or hinder users depending on design quality [5][6], which highlights the need for adaptive explanation systems that adjust complexity to user expertise or cognitive state. Measurement practices remain inconsistent, with trust, clarity, and understanding assessed using varied instruments, limiting comparability across studies. In addition, little work investigates multi-modal explanations even though combining textual, visual, and interactive elements may better support diverse users.

Developing adaptive explanation systems that tailor interaction depth to user needs represents an important direction. Further work should also explore multi-modal explanation combinations to identify which configurations best support non-expert understanding.

D. Limitations of This Review

The search was restricted to a defined set of academic databases, increasing the possibility that relevant studies published in specialized venues or emerging outlets were not captured [23]. The exclusion of grey literature, including industry evaluations and preprints, may introduce publication bias and limit the breadth of evidence. Considerable heterogeneity across study designs and evaluation metrics prevented the use of meta-analytic techniques, necessitating reliance on narrative synthesis [28].

V. CONCLUSION

Interactive explanation methods substantially improve non-expert understanding, trust, and decision accuracy compared with static approaches. What-if exploration and counterfactual recourse provide the strongest benefits because they align closely with how humans naturally reason about causes and alternatives. However, overly complex or visually dense interfaces can increase cognitive load, highlighting the need for carefully structured, guided, and minimal interaction design. For PRISM, the findings support integrating intuitive what-if controls, dynamic SHAP components, and concise, actionable counterfactual pathways to improve clarity while managing cognitive effort. Effective explainability depends not on increasing the amount of information presented, but on aligning explanation mechanisms with human reasoning patterns.

Furthermore, the review underscores the importance of adaptive, user-responsive interfaces capable of accommodating varied levels of expertise and cognitive demand. Systems that dynamically adjust interaction depth, visual complexity, or available explanation pathways are more likely to sustain user engagement and support accurate decision-making over time. These insights emphasize the broader need for human-centered explanation design, ensuring that interactive XAI tools remain accessible, meaningful, and impactful across diverse decision contexts.

Despite these contributions, the evidence base remains constrained by limited domain diversity and a lack of longitudinal evaluations. Most studies examine short-term interactions within controlled settings, offering restricted insight into how user understanding, trust calibration, and reliance on explanations evolve over time. Future research should therefore prioritise longitudinal and ecologically valid studies, as well as adaptive explanation systems that respond to users' cognitive states and contextual needs. Addressing these gaps will be essential for advancing interactive explainable AI systems that are not only interpretable, but also robust, trustworthy, and effective in real-world deployment.

REFERENCES

[1] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2870052

[2] U. Bhatt *et al.*, "Explainable machine learning in deployment," in *Proc. FAT*, 2020, pp. 648–657. [Online]. Available: https://doi.org/10.1145/3351095.3375624

[3] J. Brooke, "SUS: A retrospective," *J. Usability Studies*, vol. 14, no. 2, pp. 113–123, 2019.

[4] C. J. Cai *et al.*, "Human-centered tools for coping with imperfect algorithms during medical decision-making," in *CHI Conf. Human Factors Comput. Syst.*, 2019, pp. 1–14. [Online]. Available: https://doi.org/10.1145/3290605.3300234

[5] S. Cao, S. Stumpf, and M. Burnett, "Mental models in explainable AI: A systematic review," *Int. J. Hum.-Comput. Stud.*, vol. 165, p. 102850, 2022. [Online]. Available: https://doi.org/10.1016/j.ijhcs.2022.102850

[6] Z. Chen, J. Dodge, Q. V. Liao, S. Reddy, and J. Stoyanovich, "How much is too much? Exploring the effects of interaction complexity in explainable AI," *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. CSCW2, pp. 1–31, 2023. [Online]. Available: https://doi.org/10.1145/3610229

[7] I. K. Crombie, *The Pocket Guide to Critical Appraisal*. London, U.K.: BMJ Publishing Group, 1996.

[8] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv*, 2017. [Online]. Available: https://arxiv.org/abs/1702.08608

[9] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding explainability: Towards social transparency in AI systems," in *CHI Conf. Human Factors Comput. Syst.*, 2021, pp. 1–19. [Online]. Available: https://doi.org/10.1145/3411764.3445315

[10] U. Ehsan and M. O. Riedl, "Explainability pitfalls: Beyond algorithmic transparency," in *Proc. AAAI/ACM Conf. AI, Ethics, Society*, 2021, pp. 71–78. [Online]. Available: https://doi.org/10.1145/3461702.3462580

[11] R. Guidotti *et al.*, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018. [Online]. Available: https://doi.org/10.1145/3236009

[12] J. P. T. Higgins *et al.*, *Cochrane Handbook for Systematic Reviews of Interventions*. New York, NY, USA: Wiley, 2019.

[13] M. T. Keane, B. Smyth, and R. M. J. Byrne, "Counterfactual explanations: Theoretical foundations, methodological issues, and application guidelines," in *Proc. IJCAI*, 2021, pp. 4406–4413.

[14] H. Lakkaraju and O. Bastani, "'How do I fool you?' Manipulating user trust via misleading black box explanations," in *Proc. AAAI/ACM Conf. AI, Ethics, Society*, 2020, pp. 79–85. [Online]. Available: https://doi.org/10.1145/3375627.3375852

[15] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the AI: Informing design practices for explainable AI user experiences," in *CHI Conf. Human Factors Comput. Syst.*, 2020, pp. 1–15. [Online]. Available: https://doi.org/10.1145/3313831.3376590

[16] Q. V. Liao and K. R. Varshney, "Human-centered explainable AI: Towards a reflective socio-technical approach," in *CHI Conf. Human Factors Comput. Syst.*, 2021, pp. 1–19. [Online]. Available: https://doi.org/10.1145/3411764.3445518

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[18] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019. [Online]. Available: https://doi.org/10.1016/j.artint.2018.07.007

[19] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021. [Online]. Available: https://doi.org/10.1136/bmj.n71

[20] K. Rawal and H. Lakkaraju, "Beyond recourse: Interpreting black-box models via actionable explanations," in *Proc. AAAI*, vol. 34, no. 04, pp. 12088–12095, 2020. [Online]. Available: https://doi.org/10.1609/aaai.v34i04.6106

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discov. Data Min.*, 2016, pp. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[22] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-28954-6

[23] H. Snyder, "Literature review as a research methodology: An overview and guidelines," *J. Bus. Res.*, vol. 104, pp. 333–339, 2019. [Online]. Available: https://doi.org/10.1016/j.jbusres.2019.07.039

[24] J. Sweller, "Cognitive load theory, evolutionary educational psychology, and instructional design," *Educ. Psychol. Rev.*, vol. 28, pp. 251–271, 2016. [Online]. Available: https://doi.org/10.1007/s10648-015-9328-1

[25] I. Vessey, "Cognitive fit: A theory-based analysis of graphical vs. tabular information presentation," *Decision Sciences*, vol. 22, no. 2, pp. 219–240, 2006. [Online]. Available: https://doi.org/10.1111/j.1540-5915.1991.tb00344.x

[26] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harv. J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2018.

[27] Y. Zhang, Q. V. Liao, and R. Bellamy, "Effect of interactivity on users' understanding of AI predictions: A user study with SHAP explanations," in *CHI Conf. Human Factors Comput. Syst.*, 2020, pp. 1–14. [Online]. Available: https://doi.org/10.1145/3313831.3376590

[28] Y. Xiao and M. Watson, "Guidance on conducting a systematic literature review," *J. Plan. Educ. Res.*, vol. 39, no. 1, pp. 93–112, 2019. [Online]