Inclusive and Explainable AI Systems: A Systematic Literature Review

Amelie Girard
Data61
amelie.girard@data61.csiro.au

Didar Zowghi
Data61
didar.zowghi@data61.csiro.au

Muneera Bano
Data61
muneera.bano@data61.csiro.au

Marian-Andrei Riziou
University of Technology Sydney
marian-andrei.rizoiu@uts.edu.au

Abstract

Explainable AI (XAI) plays a crucial role in enhancing transparency and providing rational explanations to support users of AI systems. Inclusive AI actively seeks to engage and represent individuals with diverse attributes who are affected by and contribute to the AI ecosystem. Both inclusion and XAI advocate for the active involvement of the users and stakeholders during the entire AI system lifecycle. However, the relationship between XAI and Inclusive AI has not been explored. In this paper, we present the results of a systematic literature review with the objective to explore this relationship in the recent AI research literature. We were able to identify 18 research articles on the topic. Our analysis focused on exploring approaches to (1) the human attributes and perspectives, (2) preferred explanation methods, and (3) human-AI interaction. Based on our findings, we identified potential future XAI research directions and proposed strategies for practitioners involved in the design and development of inclusive AI systems.

Keywords: Inclusion, Explainable AI, Transparency, Human-centered

1. Introduction

The deployment of artificial intelligence (AI) systems inherently carries certain risks (Vasileva, 2020). Bias and disparities have surfaced as significant issues in the context of AI applications (Shi et al., 2020). Individuals are increasingly relying on AI to guide their decision-making processes in order to improve their performance. This is why it becomes imperative to explore the relationship between XAI and Inclusive AI. To illustrate the importance of this exploration, we consider two examples: healthcare decision support systems and autonomous vehicles in diverse urban settings. In the healthcare sector, decision support systems are increasingly being used to assist medical professionals in diagnosing diseases and recommending treatments. While these systems can be highly accurate, their complexity often makes them inaccessible to healthcare providers or patients from non-technical or marginalized communities (Musen et al., 2021). For example, a machine learning model that predicts the likelihood of a patient developing a specific condition may use a multitude of variables and complex algorithms. If the model's decision-making process is not explainable, healthcare providers may find it challenging to evaluate the system recommendations' worthiness, particularly those who are not AI savvy. This could be more pronounced in marginalized communities that have historically been subject to medical discrimination (Procter et al., 2023).

Similarly, autonomous vehicles in diverse urban settings make real-time decisions based on a myriad of sensors and algorithms. However, their decision-making process is often a 'black box,' making it difficult for the public to understand how decisions are made. This lack of transparency can be a significant barrier to public trust, especially among communities that have been historically subject to discrimination in transportation planning. For instance, if an autonomous vehicle is programmed to avoid areas with high crime rates, it may inadvertently reinforce existing societal biases by not serving marginalized communities (Wang et al., 2021; Guan et al., 2021). Therefore, explainability in autonomous vehicles is crucial not only for public trust but also for ensuring that these technologies are inclusive and do not perpetuate existing inequalities.

Researchers have revealed a human tendency to

URI: https://hdl.handle.net/10125/106537
978-0-9981331-7-1

excessively rely on AI recommendations (Bansal et al., 2021; Buçinca et al., 2020), where individuals may be superficially processing the information provided by AI without critically analyzing it with their own knowledge and expertise. In particular, automation bias is a type of cognitive bias where users overly depend on automation recommendations (Kathleen and Linda, 1996). Empirical evidence on the effectiveness of explanation in improving human decision-making performance has yielded mixed results (Bertrand et al., 2022; Vasconcelos et al., 2023). For instance, a study demonstrates that placebo explanations can generate a comparable degree of trust as genuine explanations (Eiband et al., 2019).

Biases present in the real world and historical data can perpetuate statistical biases, thereby reinforcing societal biases (Schneider, 2020). These biases can permeate every stage of the data generation and machine learning pipeline (Ntoutsi et al., 2020). AI systems may learn incorrect correlations from the real world, leading to erroneous classifications. In order to effectively address the specific needs of humans, it is crucial to have a deep understanding of their attributes and their interaction with the AI systems (Arrieta et al., 2020; Meske et al., 2022). Consequently, AI supported decision-making, when tainted by biased input data or algorithms, have been observed to perpetuate and reinforce discriminatory outcomes, such as racial and gender biases (Zhao et al., 2017).

In this paper, we present a systematic literature review (SLR) that we conducted to examine research in the field of XAI that focused on practices of inclusion and published from 2018 to 2023. To the best of our knowledge, no previous review has explored the interplay between XAI and inclusion. Our SLR was guided by two research questions:

(RQ1) What are the latest explanation methods used during the design of inclusive AI systems?

(RQ2) How were the explanation methods integrated into the development of inclusive AI systems?

The main contributions of this research are:

- An analysis of inclusive AI systems that have utilized XAI from 2018 to 2023.
- An examination of relationship between XAI and inclusive AI
- Insights and lessons learned when aiming to use XAI as a prerequisite for Inclusive AI.

2. Background and Related Work

2.1. Explainable AI

XAI is a timely field of research, and currently, there isn't a universally agreed-upon definition of the term "XAI" and its practical implementations (Hussain et al., 2021; Meske et al., 2022). XAI is rather referring to *"the movement, initiatives, and efforts made in response to AI transparency and trust concerns, more than to a formal technical concept"* (Adadi and Berrada, 2018, p.52140). Arrieta et al., 2020 defined XAI as: *"Given an audience, an explainable Artificial Intelligence is the one that produces details or reasons to make its functioning clear or easy to understand"* (Arrieta et al., 2020; p.6). The main goal of XAI is to generate explanations that enable humans to comprehend the decision-making process, understand the reasons behind specific predictions, and provide guidance on achieving desired outcomes (Singh et al., 2023).

Explainable AI require a solid problem formulation (Lipton, 2018; Meske et al., 2022) and robust evaluation methods (Doshi-Velez and Kim, 2017; Gilpin et al., 2018; Guidotti et al., 2018). When the problem definition is flawed, neither algorithms nor experiments can adequately address the core issue. XAI encompasses a variety of motivations for explainability, such as enhancing trust, fairness, and comprehension, that need to be replaced by precise objectives (Lipton, 2018). In evaluating the effectiveness of explanations and comparing various techniques for providing explanations, there are three potential evaluation standards: application, human, and functionally grounded explainability. The first two standards involve conducting studies with human participants, while the third standard focuses on the formal interpretability of the models. The evaluation method should match the nature of the research hypothesis being proposed (Doshi-Velez and Kim, 2017). There are cases where complex post-hoc explanations can deceive users. Interpretable models don't necessarily create or enhance trust instantly; instead, they empower users to make informed decisions regarding their trustworthiness. Therefore, the adoption of inherently interpretable models is advocated (Rudin, 2019.

The diverse XAI methods serve specific objectives and require customization based on the users and stakeholders' attributes (Arrieta et al., 2020; Meske et al., 2022). Acknowledging the inherent subjectivity of explanations, it is vital to account for the interests, demands, and requirements of the diverse stakeholders who interact with AI systems (Rocchi et al., 2004). Different stakeholders, such as AI developers, regulators, managers, and users, have distinct requirements for AI explanations based on their roles and responsibilities. For example, developers focus on improving performance and debugging, regulators need explanations for testing and certification, managers seek explanations for supervision and control, and users desire understandability to assess the system's reasoning. Additionally, individuals affected by AI-based decisions also have an interest in explainability to evaluate fairness (Meske et al., 2022). To accommodate the diverse needs of stakeholders, personalized XAI approaches are necessary, as different methods serve different purposes (Arrieta et al., 2020).

2.2. Inclusion in AI

The European Commission Ethics Guidelines for Trustworthy AI (Smuha, 2019) and AI ethics principles (Fjeld et al., 2020) advocate for lawful AI technology that is, among other principles, more inclusive. Within the AI literature, there are only a few definitions of inclusion that go beyond considering inclusion solely as a means to ensure fairness (Chi et al., 2021). Although the research community and leading tech companies like Google and Microsoft acknowledge the importance of inclusion in AI (Google, 2022; AI, 2022), concerns have arisen regarding the potential drift of the inclusion concept towards an exclusive focus on personalizing and context. This shift raises apprehensions that companies may adopt diversity and inclusion practices without adequately addressing broader societal inclusion needs (Chi et al., 2021). By integrating principles of diversity and inclusion, AI systems can be developed to better align with comprehensive societal needs, uphold human rights, and reflect contemporary societal values (Fosch-Villaronga and Poulsen, 2022).

Zowghi and da Rimini (2023) defined Inclusion as *"the process of proactively involving and representing the most relevant humans with diverse attributes; those who are impacted by, and have an impact on, the AI ecosystem context"* (Zowghi and da Rimini, 2023, p.4). The link between XAI and Inclusion has not been sufficiently explored. Thus, the focus of this study is to explore the extant literature to investigate the relationship between XAI and Inclusion.

3. Methodology

In this section, we describe our methodology for conducting the SLR. We have followed the well-established guidelines for all stages: planning, conducting, and reporting (Kitchenham et al., 2010).

3.1. Planning the review

Following a top-down approach, this review underwent multiple iterations of improvements during pilot testing to address the research questions. The objective of the pilot phase was to determine the appropriate keywords for the search engine queries as well as the adequate strategy to achieve this. We also used the unified XAI taxonomies to cover papers from diverse disciplines (Graziani et al., 2023). The search query was designed to encompass three keywords: AI systems, Explainability, and Inclusion. To maintain relevance and capture recent developments in the field, we filtered the search results to include papers published from January 2018 - March 2023. The following keywords and their alternatives were thus selected:

- AI: ("artificial intelligence", "machine learning", algorithm*)
- XAI: (explainab*, explanation*, interpretab*, transparen*, XAI)
- Inclusion: (inclu*)

We established specific inclusion criteria for selecting Peer-reviewed papers that (1) Discuss Inclusion in the field of XAI, (2) Offer a procedure or design approach to inclusion in XAI, (3) Describe ways to evaluate inclusion related to AI explanations. We excluded papers that did not provide primary insights on inclusion in XAI. We also removed literature review papers, although we scanned their reference list to see if there was any relevant paper for our study. This approach allowed us to sample a wide range of literature pertaining to inclusion in XAI. However, it is important to note that certain XAI articles may have addressed inclusion using different terminology, which we may not have found during search and selection. The same criteria for selecting studies in the primary search were applied to the potentially eligible papers identified during the secondary search using backward and forward snowballing techniques.

3.2. Search and Selection

We conducted direct searches in Scopus, IEEE, EBSCOhost and ACM on titles and abstracts for the selected duration of 2018-2023. The search on online databases yielded a total of 154 results, comprising 78 from Scopus, 18 papers from ACM, 16 from IEEE, 28 From EBSCOhost, and an additional 14 papers obtained from a Secondary search. After removing 23 duplicates, we were left with 131 unique records (see Figure 1).

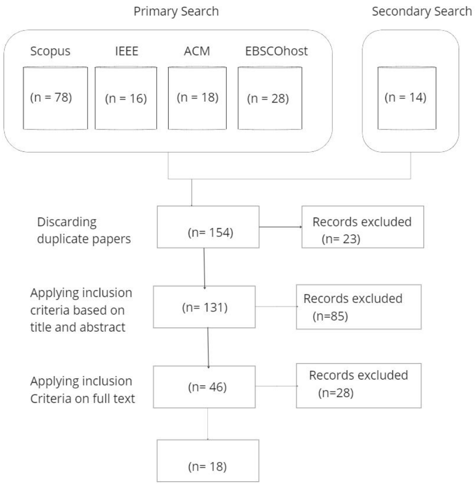

Figure 1. steps on how the final corpus was curated

Each paper's title and abstract were reviewed by one of the authors, applying the inclusion and exclusion criteria to determine whether the paper should proceed to the next phase of eligibility. In the eligibility stage, one author read the remaining articles in their entirety. Based on the inclusion and exclusion criteria, a decision was made collectively by three authors on whether to proceed with the article for the final phase. At this stage, 85 articles were excluded as they did not adequately address the proposed research questions outlined in the introduction. Finally, 18 articles were retained and advanced to the final phase of the review.

3.3. Analysis phase

For data extraction, the lead author extracted all the relevant information from the papers. This information primarily pertained to the research question. To ensure the quality of coding, two authors peer-reviewed the results and had a discussion to build consensus. The resulting codes encompassed the following elements: the human attributes, the provided explanation, and the result of the interaction between the human and the AI system. The diversity of subject areas is represented in Table 1, which categorizes the corpus based on domain/task.

Table 1. Categorization of papers based on the Domain/Task.

Domain/Task	Papers
Art	emotions (Lieto et al., 2022), values (Kadastik et al., 2022).
Business/Finance	Shot-term lending (Gadzinski and Castello, 2022), credit scoring (Lyu et al., 2023).
Human resources	employees satisfaction (Lyu et al., 2023), candidates hiring (Hofeditz et al., 2022; Sánchez-Monedero et al., 2020), job matching (Delecraz et al., 2022), employees performance (Park et al., 2022), disability (Tran et al., 2021).
Policy/Regulation	fine-tech lending (Chou, 2019, civil right (Chi et al., 2021).
Education	incidental learning (Gajos and Mamykina, 2022), online teaching (Nazaretsky et al., 2022; Conati et al., 2021), learning history and facial recognition (Kusuma et al., 2022).
Urban planning	resources allocation (Lyu et al., 2023).
Healthcare	diabetes risk monitoring onset (Bhattacharya et al., 2023).

4. Results

The data extraction and analysis of the selected papers resulted in classifying the papers into four distinct categories: (D1) focuses on the attributes or perspectives of individuals who are affected by the AI system, and (D2) centers on the attributes or perspectives of individuals who influence the AI system. The selection of these categories is grounded in the concept of Inclusion as defined by Zowghi and da Rimini (2023). Additionally, (D3) delves into the preferences of these individuals, and (D4) examines the nature of interactions between humans and AI, both of which are influenced by Bederson and Shneiderman (2003)'s theories on human-AI interaction (Bederson and Shneiderman, 2003). Categories (D1), (D2), and (D4) aim to address RQ2, exploring how inclusion is tackled in the Explainable AI (XAI) literature. Category (D3) focuses on answering RQ1, investigating the methods of explainability discussed in the literature. Table 2 outlines how the selected papers have addressed these categories.

Table 2. Categorization of the papers based on their focus area (D1) attributes or perspectives of the Humans who have an impact on the AI system (D2) attributes or perspectives of the Humans who are impacted by the AI system (D3) explanation preferences of these Humans, and (D4) the type of Human-AI interaction.

Articles	D1	D2	D3	D4
Conati et al., 2021	X	X	X	X
Kusuma et al., 2022	X	X	X	X
Lyu et al., 2023	X		X	X
Singh et al., 2023		X	X	
Hofeditz et al., 2022		X	X	
Lieto et al., 2022	X		X	
Gajos and Mamykina, 2022				X
Cai and Canales, 2022			X	
Bhattacharya et al., 2023	X		X	X
Tran et al., 2021	X	X	X	
Nazaretsky et al., 2022	X	X	X	X
Dankwa-Mullan and Weeraratne, 2022	X			
Chi et al., 2021	X		X	
Park et al., 2022	X		X	
Gadzinski and Castello, 2022	X		X	
Kadastik et al., 2022	X			
Lopes et al., 2021	X		X	X
Sánchez-Monedero et al., 2020	X		X	X

4.1. Audience attributes or perspectives

Current XAI methods are predominantly designed for machine learning professionals rather than for users who lack AI expertise. However, the research seems to shift this focus, emphasizing these non-AI expert users recognizing the identified need to cater to non-AI experts. The shift towards designing XAI methods for non-AI experts is not just a trend but a necessity for inclusion. Hofeditz et al., 2022 provides a compelling example: their focus on the impact of candidates' sensitive attributes on HR decisions.

Domain experts. Studies that explored the audience attributes investigated the characteristics of individuals who interact or are impacted by the AI system. These articles typically presuppose the existence of an accurate and fair computational model; in this context, Hofeditz et al. (2022) explored the impact of sensitive attributes (age, foreign race, and gender) of Job applicants on the decision-making process of HR-professional. The target audience is AI experts, most of whom have limited HR experience. Similarly, Lyu et al., 2023 focused on domain experts but selected six urban planners from different countries, each with varying geographical expertise and experience ranging from 3 to 10 years, to study the fair allocation of resources in the city of New York. Nevertheless, the residents represented in the data stopped at a categorical level and acknowledged the need for further consideration of their attributes.

Lay users. Singh et al. (2023) conducted two studies using American participants from Amazon Turk in the domains of credit scoring and employee satisfaction to understand user preferences for directive explanations compared to non-directive explanations. They argued that machine learning-generated explanations could be enhanced by not only explaining why a decision was made but also providing guidance on how individuals can achieve their desired outcome. Similarly, Lieto et al. (2022) focused on the perspective of the museum users who received recommendations for cultural items that evoke not only familiar emotions from previous experiences or preferences but also introduce new items that evoke different emotional responses. Deaf users evaluated the system.

The case and user studies outlined the specific audiences, which are specified in Table 3. However, the empirical studies only mentioned regulators and other users.

Table 3. Case and user studies target audience

Articles	Domain experts	Lay users	AI experts
Kusuma et al., 2022	X	X	X
Lyu et al., 2023	X		
Singh et al., 2023		X	
Hofeditz et al., 2022	X		
Lieto et al., 2022		X	
Gajos and Mamykina, 2022		X	
Cai and Canales, 2022	X	X	X
Bhattacharya et al., 2023	X	X	
Tran et al., 2021		X	

Based on the audience attributes, an explanation method needs to be selected (Arrieta et al., 2020,Gilpin et al., 2018). The following section explores the explanation methods chosen in the case studies.

4.2. Explanation methods

This section presents the explanation methods explored by the selected corpus of studies. Similar to the studies focusing on target audience attributes or perspectives, research focusing on exploring or comparing explanation methods also ensured an accurate model before exploring users' explanation preferences.

Actionable explainability. Singh et al. (2023) introduced the concept of actionable explainability. They argued that directive explanations, specifically directive-generic explanations, were favoured when participants sought autonomy and had their own problem-solving ideas. Non-directive explanations were found to be more suitable when outcomes were favourable, particularly in the credit scoring domain. The key findings highlighted a significant preference for directive explanations. Directive-specific explanations were preferred in scenarios with unfavourable outcomes, while directive-generic explanations were favoured when participants desired autonomy. Along the same lines, Bhattacharya et al. (2023) employed directive explanations to monitor the risk of diabetes onset in conjunction with what-if exploration. They presented an explanation dashboard that predicts the onset of diabetes and clarifies these predictions using three distinct methods: data-centric, feature-importance, and example-based explanations. The evaluation of these methods focused on their understandability, usefulness, actionability, and trustworthiness. The results revealed a preference among participants for data-centric explanations, which provide local explanations supplemented by a global overview, over the other methods.

Causal explainability: Attribution, Contrastive. Lyu et al. (2023) utilized explanations based on causal attribution (Why), contrastive (Why Not) and counterfactual reasoning (What If, How To) to assist urban planners in identifying and addressing unfairness in resource allocation problems.

High-level explanation. Hofeditz et al. (2022) participants received a high-level explanation of the recommendation system input-output process. The findings suggest that the high-level explanation did not moderate the effect of the system's recommendations on the selection of older and female candidates. However, it did positively influence the selection of foreign-race candidates. The authors propose that the lack of explanation impact on age and gender selection could be due to the need for different types of explanation methods. More suitable types might be needed to support HR professionals.

4.3. Human AI Interaction

The nature of human-AI interaction can either facilitate or hinder inclusion and explainability. Effective User interfaces or personalised explanations make AI systems more accessible and inclusive. All papers highlighted the importance of the interaction between the users and the automated system. Lyu et al. (2023) Found that contrastive explanations were crucial for explainability, but the design iteration timeline was not extensively used. While domain experts tended to explore through trial and error before turning to automatic recommendations, non-expert planners might benefit more from prioritizing these recommendations. These findings are comparable to the result of Gajos and Mamykina (2022) study, which formulated two main hypotheses: first, that presenting individuals with a recommendation and an explanation would improve immediate decision-making but not lead to significant learning. Second, alternative designs aimed at promoting deeper processing of AI-provided information would not only offer immediate benefits but also result in incidental learning. These findings suggest that merely including explanations alongside AI-generated recommendations may not ensure careful engagement with AI-provided information. The research introduces an alternative design that encourages incidental learning and more thoughtful processing of AI recommendations and explanations. This design, where individuals are responsible for reaching decisions themselves based on AI explanations, resulted in both immediate decision benefits and knowledge acquisition.

5. Discussion

The discussion consolidates our findings in relation to two primary research questions: RQ1, which investigates the dominant methods of explainability in existing literature, and RQ2, which examines how the principle of inclusion is articulated in the field of Explainable AI (XAI). To systematically explore these questions, we've organized the literature into four principal dimensions—D1, D2, and D4, which are particularly aligned with RQ2, while D3 addresses RQ1. These dimensions focus on the attributes and perspectives of individuals either impacted by or influencing AI systems, as well as the dynamics of human-AI interactions. Our research reveals that while most studies focus on the attributes of individuals affected by AI systems, they often overlook those who influence these systems. Notable exceptions include Conati et al. (2021) and Kusuma et al. (2022), who serve as crucial counterpoints and indicate new directions for future research. Moreover, the importance of understanding structural and historical biases is emphasized, a perspective largely missing in existing literature except for the work by Kusuma et al. (2022). In terms of explainability methods, most papers used general explanations but recognized this as a limitation. A few studies ventured into more nuanced forms of explainability, such as ”directive explainability” introduced by Singh et al. (2023) and employed by Bhattacharya et al. (2023). These nuanced approaches indicate a growing awareness of the need for more tailored and actionable explanations in different contexts. When it comes to human-AI interaction, the nature of this interaction can either promote or inhibit inclusion, and explainability serves as the linchpin for ensuring the former. User-friendly interfaces or personalised explanations make AI systems more accessible and inclusive. For example, Lyu et al. (2023) found that contrastive explanations were crucial for explainability, but the design iteration timeline was not extensively used. This aligns with the findings of Gajos and Mamykina (2022), who posited that merely including explanations alongside AI-generated recommendations may not ensure careful engagement with AI-provided information. Their research introduces an alternative design that encourages incidental learning and more thoughtful processing of AI recommendations and explanations, resulting in both immediate decision benefits and knowledge acquisition. Several studies, such as those by Singh et al. (2023) and Bhattacharya et al. (2023), acknowledged the potential ramifications of understanding individual attributes for the broader applicability of their research findings. This suggests an expansive scope for upcoming research in XAI and inclusion. The field of XAI is indeed growing, but significant gaps remain, particularly in the balanced consideration of individuals who are either affected by or influence AI systems. Bridging these gaps is crucial for the evolution of more inclusive and explainable AI frameworks.

6. Recommendations for future research

Our analysis unequivocally establishes that explainability is not just an add-on but a prerequisite for Inclusive AI. The intersection of XAI and Inclusive AI is not merely a research gap but a critical area that demands immediate and sustained attention. Given that diversity is a prerequisite for inclusion, we recommend the following to researchers and practitioners:

- Incorporating those affected by AI into the design process and ensuring the inclusion of under-represented groups in the evaluation process.
- Comparing the advantages and disadvantages of different XAI methods considering contextual and Human attributes that impact the AI System.
- There is a need to personalise the explanation to the human attributes such as cognitive abilities as well as investigate how different approaches to designing human-AI interactions can impact the Human learning process.

7. Threats to Validity

Despite our meticulous adherence to the Evidence-Based Systematic Literature Review (SLR) guidelines, which ensured a rigorous search and selection of our sample studies, there remains a chance that some documents may not have been included in our data collection. This exclusion might occur due to their unavailability or absence on digital platforms.

Internal Validity. There could be a risk due to the limited number of papers selected and the narrow time span considered. Given that XAI and Inclusive AI are relatively new research fields, we limited our search to the past five years. There could also be a risk of biases in the selection of studies and data extraction. To mitigate these issues, we utilized the investigator triangulation technique.

Construct Validity. A possible risk could be the irrelevance of many papers that surfaced due to our search string. We initially selected a considerable number of papers by reviewing their abstracts, looking for insights on inclusive XAI to ensure not to miss the relevant studies. However, many of them were discarded after a full read, as they were not directly related to our focused topic. Another risk could stem from the subjective interpretation of the extracted data. We addressed both these concerns using the investigator triangulation technique.

8. Conclusion

We conduct a systematic review with the aim of gaining insights into the latest explainable methods used as a foundation for building inclusive AI. The dimensions identified include the attributes and perspectives of individuals, preferences for different types of explanations, and the dynamics of interaction between individuals and AI systems.

Our analysis reveals that while the influence of audience attributes and perspectives on the effectiveness of XAI systems is recognized, factors such as AI expertise, cultural background, and personal experiences have not been thoroughly considered in the selection of explanation methods. This finding highlights the need for personalisation in AI explanation design to accommodate the diverse range of individuals who are impacted by or have an impact on the AI system.

Moreover, the review underscores the lack of consideration for societal and historical discrimination in the context of human interaction with AI systems. This oversight points to a significant gap in current practices, suggesting that these factors warrant more attention to structural inclusive AI systems. By addressing these issues, we can better ensure that AI systems are not only understandable but also equitable and fair, thereby reflecting the diverse realities of all audiences.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE access*, *6*, 52138–52160.

AI, M. (2022). *2022 ai principals progress update*. https://www.microsoft.com/en-us/ai/responsible-ai-resources (accessed: 25.03.2023)

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, *58*, 82–115.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Bederson, B. B., & Shneiderman, B. (2003). *The craft of information visualization: Readings and reflections*. Morgan Kaufmann.

Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How cognitive biases affect xai-assisted decision-making: A systematic review. *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society*, 78–91.

Bhattacharya, A., Ooge, J., Stiglic, G., & Verbert, K. (2023). Directive explanations for monitoring the risk of diabetes onset: Introducing directive data-centric explanations and combinations to support what-if explorations. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 204–219.

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. *Proceedings of the 25th international conference on intelligent user interfaces*, 454–464.

Cai, J., & Canales, J. I. (2022). Dual strategy process in open strategizing. *Long Range Planning*, *55*(6), 102177.

Chi, N., Lurie, E., & Mulligan, D. K. (2021). Reconfiguring diversity and inclusion for ai ethics. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 447–457.

Chou, A. (2019). What's in the black box: Balancing financial inclusion and privacy in digital consumer lending. *Duke LJ*, *69*, 1183.

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized xai: A case study in intelligent tutoring systems. *Artificial intelligence*, *298*, 103503.

Dankwa-Mullan, I., & Weeraratne, D. (2022). Artificial intelligence and machine learning technologies in cancer care: Addressing disparities, bias, and data diversity. *Cancer Discovery*, *12*(6), 1423–1427.

Delecraz, S., Eltarr, L., Becuwe, M., Bouxin, H., Boutin, N., & Oullier, O. (2022). Making recruitment more inclusive: Unfairness monitoring with a job matching machine-learning algorithm. *2022 IEEE/ACM International Workshop on Equitable Data & Technology (FairWare)*, 34–41.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The impact of placebic explanations on trust in intelligent systems. *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, 1–6.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*, (2020-1).

Fosch-Villaronga, E., & Poulsen, A. (2022). Diversity and inclusion in artificial intelligence. *Law* *and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice*, 109–134.

Gadzinski, G., & Castello, A. (2022). Combining white box models, black box machines and human interventions for interpretable decision strategies. *Judgment and Decision Making*, *17*(3), 598–627.

Gajos, K. Z., & Mamykina, L. (2022). Do people engage cognitively with ai? impact of ai assistance on incidental learning. *27th International Conference on Intelligent User Interfaces*, 794–806.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80–89.

Google. (2022). *2022 ai principals progress update*. https : / / ai . google / static / documents / ai - principles - 2022 - progress - update . pdf (accessed: 25.03.2023)

Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J. P., Yordanova, K., Vered, M., Nair, R., Abreu, P. H., Blanke, T., Pulignano, V., et al. (2023). A global taxonomy of interpretable ai: Unifying the terminology for the technical and social sciences. *Artificial intelligence review*, *56*(4), 3473–3504.

Guan, J., Zhang, S., D'Ambrosio, L. A., Zhang, K., & Coughlin, J. F. (2021). Potential impacts of autonomous vehicles on urban sprawl: A comparison of chinese and us car-oriented adults. *Sustainability*, *13*(14), 7632.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1–42.

Hofeditz, L., Clausen, S., Rieß, A., Mirbabaie, M., & Stieglitz, S. (2022). Applying xai to an ai-based system for candidate management to mitigate bias and discrimination in hiring. *Electronic Markets*, 1–27.

Hussain, F., Hussain, R., & Hossain, E. (2021). Explainable artificial intelligence (xai): An engineering perspective. *arXiv preprint arXiv:2101.03613*.

Kadastik, N., Pedersen, T., Bruni, L., Damiano, R., Lieto, A., Striani, M., De Giorgis, S., Kufli, T., & Wecker, A. (2022). Exploring values in museum artifacts in the spice project: A preliminary study. *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (Barcelona, Spain)(UMAP'22). Association for Computing Machinery, New York, NY, USA*.

Kathleen, M., & Linda, S. (1996). *Human decision makers and automated decision aids: Made for each other?. in automation and human performance: Theory and applications*. CRC Press.

Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O. P., Turner, M., Niazi, M., & Linkman, S. (2010). Systematic literature reviews in software engineering–a tertiary study. *Information and software technology*, *52*(8), 792–805.

Kusuma, M., Mohanty, V., Wang, M., & Luther, K. (2022). Civil war twin: Exploring ethical challenges in designing an educational face recognition application. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 369–384.

Lieto, A., Pozzato, G. L., Striani, M., Zoia, S., & Damiano, R. (2022). Formal methods meet xai: The tool degari 2.0 for social inclusion.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.

Lopes, B. G., Soares, L. S., Prates, R. O., & Gonçalves, M. A. (2021). Analysis of the user experience with a multiperspective tool for explainable machine learning in light of interactive principles. *Proceedings of the XX Brazilian Symposium on Human Factors in Computing Systems*, 1–11.

Lyu, Y., Lu, H., Lee, M. K., Schmitt, G., & Lim, B. Y. (2023). If-city: Intelligible fair city planning to measure, explain and mitigate inequality. *IEEE Transactions on Visualization and Computer Graphics*.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management*, *39*(1), 53–63.

Musen, M. A., Middleton, B., & Greenes, R. A. (2021). Clinical decision-support systems. In *Biomedical informatics: Computer applications in health care and biomedicine* (pp. 795–840). Springer.

Nazaretsky, T., Bar, C., Walter, M., & Alexandron, G. (2022). Empowering teachers with ai: Co-designing a learning analytics tool for personalized instruction in the science classroom. *LAK22: 12th International Learning Analytics and Knowledge Conference*, 1–12.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., et al. (2020). Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1356.

Park, H., Ahn, D., Hosanagar, K., & Lee, J. (2022). Designing fair ai in human resource management: Understanding tensions surrounding algorithmic evaluation and envisioning stakeholder-centered solutions. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–22.

Procter, R., Tolmie, P., & Rouncefield, M. (2023). Holding ai to account: Challenges for the delivery of trustworthy ai in healthcare. *ACM Transactions on Computer-Human Interaction*, *30*(2), 1–34.

Rocchi, C., Stock, O., Zancanaro, M., Kruppa, M., & Krüger, A. (2004). The museum visit: Generating seamless personalized presentations on multiple devices. *Proceedings of the 9th international conference on Intelligent user interfaces*, 316–318.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206–215.

Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to'solve'the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 458–468.

Schneider, V. (2020). Locked out by big data: How big data algorithms and machine learning may undermine housing justice. *Colum. Hum. Rts. L. Rev.*, *52*, 251.

Shi, S., Wei, S., Shi, Z., Du, Y., Fan, W., Fan, J., Conyers, Y., & Xu, F. (2020). Algorithm bias detection and mitigation in lenovo face recognition engine. *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II 9*, 442–453.

Singh, R., Miller, T., Lyons, H., Sonenberg, L., Velloso, E., Vetere, F., & Dourish, P. (2023). Directive explanations for actionable explainability in machine learning applications. *ACM Trans. Interact. Intell. Syst.* https://doi.org/10.1145/3579363

Smuha, N. A. (2019). The eu approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International*, *20*(4), 97–106.

Tran, H. X., Le, T. D., Li, J., Liu, L., Liu, J., Zhao, Y., & Waters, T. (2021). Recommending the most effective intervention to improve employment for job seekers with disability. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 3616–3626.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–38.

Vasileva, M. I. (2020). The dark side of machine learning algorithms: How and why they can leverage bias, and what can be done to pursue algorithmic fairness. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3586–3587.

Wang, Z.-j., Chen, X.-m., Wang, P., Li, M.-x., Ou, Y.-j.-x., & Zhang, H. (2021). A decision-making model for autonomous vehicles at urban intersections based on conflict resolution. *Journal of advanced transportation*, *2021*, 1–12.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Zowghi, D., & da Rimini, F. (2023). Diversity and inclusion in artificial intelligence. *arXiv preprint arXiv:2305.12728*.