# Insights into Building Fairness in AI-XR Systems

**Clara Maathuis[1], Dragos Datcu[2] and Dorin Mircea Popovici[3]**
[1]Open University of the Netherlands, Heerlen, The Netherlands
[2]Independent Researcher, The Netherlands
[3]Ovidius University of Constanta, Romania

clara.maathuis@ou.nl
email@dragosdatcu.eu
dmpopovici@univ-ovidius.ro

**Abstract:** Seeing the ongoing developments in the AI-XR (Artificial Intelligence-Extended Reality) domain, immersive systems are created with transformative potential across various domains, yet ensuring fairness remains an important challenge for responsible development and deployment. To this end, this research captures and provides expert and field insights and technical recommendations for building fair AI-XR systems by embedding fairness throughout the AI-XR life cycle. Key strategies include curating diverse and representative datasets, integrating fairness-aware algorithms, continuously auditing models to detect, mitigate, and avoid bias, multidisciplinary collaboration engaging various experts and user, and the use of fairness metrics such as demographic parity, equal opportunity, and user satisfaction across groups. At the same time, the findings advocate for a responsible and human-centred approach that is supported by ethical guidelines, technical best practices, and continuous stakeholder engagement in order to ensure that AI-XR systems deliver equitable, accessible, and culturally sensitive experiences for all users.

**Keywords**: AI, AI fairness, AI-XR, AI-XR fairness, Responsible AI, Trustworthy AI

## 1. Introduction

Technological advancements are rapidly reshaping society, with Artificial Intelligence (AI) becoming deeply integrated across diverse domains such as healthcare, education, security, entertainment, and manufacturing (Baum, 2021; Oladoyinbo et al., 2024). Among these, Extended Reality (XR) encompassing augmented, virtual, and mixed reality, stands out as a transformative force through which AI allows the creation and enhancement of immersive, interactive, and highly personalized environments (Hirzle et al., 2023; Huynh-The et al., 2023). The convergence of AI and XR is thus enabling applications ranging from hyper-realistic training simulations and adaptive educational tools to intelligent healthcare diagnostics and dynamic retail experiences, fundamentally altering how individuals interact, learn, and work (Qayyum et al., 2024). Nevertheless, as AI-powered XR systems become increasingly advanced and adaptive, they also introduce complex ethical, legal, and technical challenges. Issues such as algorithmic transparency, fairness, data privacy, and accountability must be addressed from the earliest stages of design and development to ensure that these technologies are deployed responsibly and in alignment with ethical and societal norms and values (Ratican, Hutson & Wright, 2023; Kiourtis et al., 2024; Maathuis et al., 2025).

While the development and deployment of AI-XR systems have prompted various studies dedicated to tackling issues such as transparency, safety, and security, each being essential for building user trust and safeguarding sensitive information, there remains a gap in attention to the concept of fairness within these technologies (Qayyum et al., 2024; Stacchio et al., 2024). Current research and industry guidelines emphasize the need for transparent data practices, robust cyber security measures, and privacy preserving mechanisms to address the risks associated with the vast collection and processing of personal data in immersive environments (Alkaeed et al., 2024)]. Nevertheless, the challenge of embedding fairness into AI-XR systems, i.e., as to systematically identify, mitigate, and prevent bias throughout their entire life cycle has received comparatively less attention. This dimension is as important since unchecked algorithmic bias can perpetuate discrimination and inequity, undermining the ethical development and deployment of AI-XR applications (Fox & Thornton, 2022). On this behalf, addressing fairness requires not only technical interventions, but also a comprehensive understanding of how bias can arise at different stages, from data collection and model training to real-time system operation and user interaction, ensuring that these systems serve all users equitably, and are thus responsible and trustworthy (Wheeler & Carroll, 2022; Maathuis, 2024).

To address the identified gap concerning fairness in the context of AI-XR systems, a mixed-methods study is conducted for integrating both expert insight and comprehensive literature analysis. Specifically, a workshop with thirteen domain experts is conducted, whose diverse perspectives and practical experiences provided various nuanced recommendations on embedding fairness throughout the AI-XR development life cycle to enable a robust synthesis of current academic and practitioner knowledge. This approach produces

understanding on the multifaceted meaning of fairness in this context, spanning stages from data collection and algorithm design to deployment and user interaction. At the same time, the recommendations provide actionable insights for building AI-XR systems that proactively mitigate and avoid bias, thereby supporting the creation of more equitable, responsible, and trustworthy immersive technologies.

The findings of this research contribute to both academic and practical efforts aimed at building and fostering fairness in AI-XR systems. This advances the theoretical understanding of fairness as a dynamic and context-dependent construct that must be addressed at every stage of the AI-XR system life cycle. At the same time, this research provides a structured framework and a set of empirically grounded insights that can inform future research on bias mitigation, ethical design, and fairness evaluation in immersive technologies. For practitioners, the actionable recommendations derived from expert consensus offer guidance for integrating fairness considerations into real-world AI-XR development processes, from data governance and algorithmic transparency to inclusive user experience design. The implications extend to policy makers and industry leaders, highlighting the necessity of multidisciplinary collaboration and continuous assessment to ensure that AI-XR systems do not inadvertently perpetuate inequities.

The remainder of this article is structured as follows. Section 2 discusses the background and related studies conducted in this domain. Section 3 presents the methodological approach considered in this research. Section 4 presents the findings and recommendations that this research proposes for building fair AI-XR systems. At the end, in Section 5 are provided concluding remarks and future research directions.

## 2. Related Research

Building fair AI-XR systems is a pressing challenge as these technologies increasingly mediate access to information, social interaction, and opportunity in digital environments (Qayyum et al., 2024; Stacchio et al., 2024). At the same time, it is important to mention the fact that both AI and XR domains have unique characteristics and challenges which further causes unique fairness concerns for AI-XR systems: AI systems risk amplifying biases present in training data or algorithms (Albahri et al., 2023; Khera et al., 2023; Ueda et al., 2024), while XR technologies have historically struggled with accessibility, representation, and the risk of replicating societal inequalities in immersive contexts (Guthula & Basu, 2024; Hadan et al., 2024). This calls for a socio-technical and multidisciplinary approach that spans data collection, algorithmic design, user experience, and societal impact by including AI bias detection, auditing, and mitigation methods with inclusive XR design principles that prioritize accessibility, cultural sensitivity, and anti-discrimination measures (Kajla et al., 2024; Maathuis et al., 2025). Achieving this means drawing on expertise from AI, computer science, software engineering, ethics, law, and human-computer interaction domains to ensure that fairness is embedded at every stage of the AI-XR life cycle.

In the context of AI, fairness refers to the principle that AI systems should make decisions that are unbiased, equitable, and just across different individuals and groups, particularly with respect to sensitive characteristics such as race, gender, age, or disability (Mehrabi et al., 2021; Zhang, 2024). This concept is inherently multidisciplinary, drawing from law where fairness involves protection from discrimination, social sciences where it relates to power dynamics and social equity, philosophy where it is related to moral rightness and justice, and quantitative fields where it is formalized through mathematical criteria such as equal allocation or error rates across groups (Kattnig et al., 2024; Pfeiffer et al., 2023; John-Mathews et al., 2022). In this context, fairness is operationalized both at the individual level (ensuring similar individuals are treated similarly) and at the group level (ensuring equitable treatment and outcomes across demographic groups). Hence, the implications of fair AI systems are multi-faceted: they help prevent the reinforcement of existing social inequalities, protect marginalized or disadvantaged groups of population, and foster public trust by aligning with ethical and regulatory standards. On the contrary, unfair AI systems risk perpetuating bias and injustice, leading to reputation, legal, and societal harm (Calvi & Kotzinos, 2023; Angerschimd et al., 2022). This allows that the technologies are developed and operate in a responsible and trustworthy manner. Embedding fairness is a core requirement for trustworthy AI, alongside attributes such as transparency, accountability, safety, and robustness, as recognized in international frameworks and guidelines (Kaur et al., 2022; Varona & Saurez et al., 2022; Cools & Maathuis, 2024). In the context of AI-XR systems, which blend AI with immersive experiences, implementing fairness becomes even more important due to the high potential for bias and exclusion in virtual environments (Christ-Bredendemuhl & Schaarschmidt, 2022). By coupling fairness principles from both the AI and XR domains, developers can create systems that are not only technically robust but also socially responsible and aligned with legal, social, and ethical principles, norms, and values.

Such systems are built for instance in the education domain where responsible AI-XR platforms embed fairness-aware algorithms, privacy preserving data practices, and transparent governance to create adaptive learning environments that are accessible and culturally sensitivity (Tabassum & Zadir, 2024). In healthcare, AI-driven XR systems such as virtual therapy avatars and diagnostic assistants utilize diverse training data, real-time bias detection, and explainable AI techniques to ensure equitable treatment and safeguard patient privacy, in line with emerging industry best practices (Shaikh et al., 2022; Akhtar, 2024; Bang & Kim, 2023). Major technology stakeholders, such as those detailed in Microsoft's 2025 Responsible AI Transparency Report (Shorotriya et al., 2025; Kallina et al., 2025) have implemented layered risk management, pre-deployment reviews, and stakeholder engagement in order to ensure compliance with regulatory standards and to proactively address fairness, safety, and transparency in AI-enabled XR applications. Additionally, various research initiatives and workshops emphasize the importance of designing AI agents and XR interfaces that accommodate users with varying abilities through multi-modal input, adaptive content, and continuous user feedback, thereby expanding access and reducing exclusion.

Other studies mapped the sources of bias in AI, such as unrepresentative datasets, flawed problem framing, and inadequate validation, and have proposed actionable checklists and toolkits such as the AI Fairness 360 for bias detection and mitigation throughout data collection, model development, validation, and deployment in immersive environments (Bellamy et al., 2019; Menezes et al., 2021). In healthcare contexts, frameworks such as the Translational Evaluation of Healthcare AI (TEHAI) and the WHO's ethical guidance emphasize the importance of diverse, inclusive datasets, transparent reporting, and continuous monitoring to ensure that AI-based systems do not perpetuate health disparities or social inequities (Reddy et al., 2021; Jacob et al., 2025; Rudnicka et al., 2024). Moreover, initiatives like the NIH All of Us Research Program and the Standing Together project exemplify large-scale efforts to create standards for dataset diversity and inclusivity, which are essential for generalizable and fair AI-XR systems (Blegen et al., 2024; Baskir et al., 2025; Alderman et al., 2025). Therefore, while these scientific studies demonstrate that embedding fairness in AI-XR requires not only technical rigor and methodological transparency, but also ongoing engagement with diverse stakeholders and the adoption of ethical best practices across the entire system lifecycle, direct recommendations that refer to both data and model aspects as well as human and procedural aspects for building and enhancing fairness in AI-XR systems are needed and are thus tackled in this research.

## 3. Research Methodology

To investigate the meaning and embedding of fairness into AI-XR systems, this research employs a dual phase methodological approach. Firstly, the research performs an extensive literature review focused on fairness challenges, mitigation strategies, and evaluation metrics within AI-XR environments in scientific databases such as IEEE Xplore, ACM Digital Library, and Springer using combinations of keywords like AI, AI fairness, AI-XR, fair AI-XR, XR fairness. Secondly, the research implies the organization of a workshop with thirteen domain experts in XR and AI in order to gather practitioner insights and experiences related to fairness in AI-XR development and deployment in April 2025. In this process, a set of six research questions (RQs) concerning fairness in AI-XR systems were formulated and addressed, as follows:

- RQ1: How can AI-XR systems be designed to ensure equitable access and experiences across diverse user groups including those with different physical abilities or socio-economic backgrounds?
- RQ2: How can developers mitigate algorithmic bias in AI-XR systems that may disproportionately affect certain user groups based on factors like race, gender, or age?
- RQ3: What strategies can be employed to ensure fair representation in the data used to train AI models for XR applications?
- RQ4: How can AI-XR systems be designed to adapt to different cultural contexts while maintaining fairness across diverse user groups?
- RQ5: What mechanisms should be in place to detect and address unfair outcomes or discriminatory behaviours in AI-XR environments?
- RQ6: What metrics can be used to measure and evaluate fairness in AI-XR systems, particularly in immersive environments?

While the literature review component synthesizes existing knowledge on building and enhancing fairness in AI-XR systems, the workshop uncovers practical attributes, methods, and techniques that define a taxonomy of fairness-centred approaches in an actionable manner through direct actionable recommendations. The responses were systematically gathered, clustered, and analysed according to their thematic relevance and relation to specific fairness categories using NLP (Natural Language Processing) methods such as topic

modelling and word clouds. The findings are presented in this article, captured as a word cloud in Figure 1 and bi-grams in Figure 2, and discussed in the context of current state-of-the-art frameworks, methodologies, and applications aimed at embedding fairness within AI-XR environments.




**Figure 1: AI-XR systems**

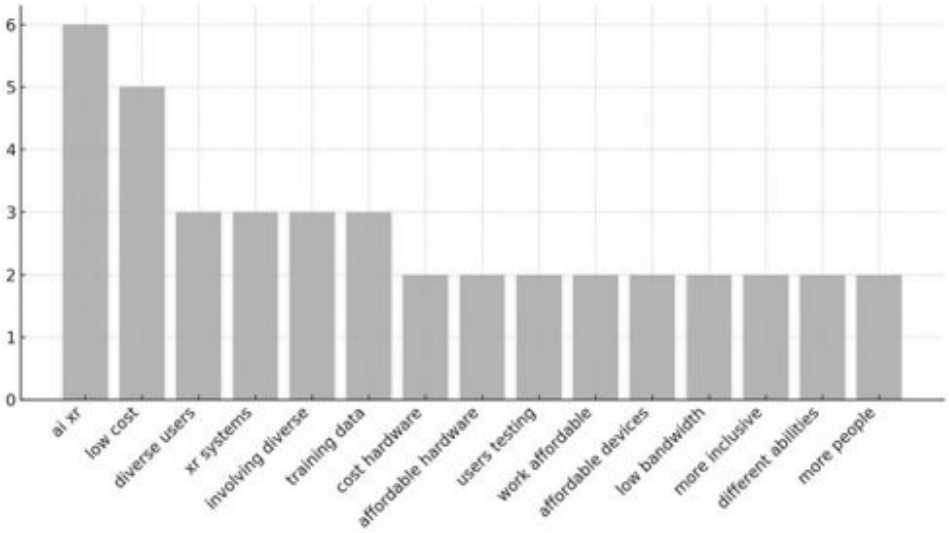


**Figure 2: AI-XR fairness core elements**

## 4. Fairness Recommendations

To build fair AI-XR systems calls for a proper understanding of fairness in the XR domain through a multifaceted stance. Firstly, data used to train AI models for XR must be representative, inclusive, and free from historical or systemic biases, which can be addressed through careful dataset curation, augmentation, and continuous auditing. Secondly, model development should incorporate fairness-aware algorithms and regularization techniques that explicitly mitigate bias, such as adversarial debiasing or reweighting methods. Thirdly, the deployment of AI-XR systems needs to include adaptive mechanisms that personalize experiences to diverse user needs and cultural contexts, ensuring accessibility and inclusivity. Moreover, in order to detect and address unfair outcomes, robust fairness detection methods such as disparate impact analysis, subgroup performance evaluation, and counterfactual fairness testing need to be integrated into the system life cycle. At the same time, the use of well-established fairness metrics, including demographic parity, equalized odds, and user-centric experiential measures, is critical for ongoing evaluation and accountability in immersive environments. Further, each of these elements are elaborated below.

To ensure equitable experiences in AI-XR systems across diverse user groups (RQ1) including those with varying physical abilities and socioeconomic backgrounds, researchers and developers need to adopt a human-centred and inclusive design approach. Central to this is the implementation of low cost hardware compatibility and affordable software solutions, enabling participation from users with limited resources. On this behalf, AI-XR interfaces need to be adaptive and accessible by incorporating multi-modal interaction methods such as voice commands, gesture recognition, and gaze tracking to accommodate different abilities. At the same time, personalization features, including customizable interfaces and adaptive content, allow the

system to respond to individual user needs and preferences. In addition, multilingual and culturally sensitive content, as well as support for offline or low-bandwidth operation, further broaden accessibility for users in diverse regions and with varying levels of connectivity. At the same time, by involving a wide spectrum of users in the design and testing phases through a human-centred and co-design approach, ensures that the system addresses real world challenges and needs while accounting potential risks. As important is the incorporation of regular auditing measures for the models included for bias, adherence to accessibility ethical and legal standards, and the adoption of open-source development practices foster transparency by embedding XAI (explainable AI) methods and inclusivity. These aspects are captured in Figure 3 and show that by integrating these strategies, affordable hardware, flexible and accessible interfaces, adaptive AI, and continuous user engagement, AI-XR systems can deliver equitable, user-centred experiences for their users. For example, fairness evaluations should compare task completion rates, error rates, system latency, and user comfort across demographic groups, disability categories, device types, and connectivity conditions to identify disparities in system performance. When such disparities are detected, developers need to apply targeted mitigation measures such as retraining models on more representative datasets, redesigning inaccessible interaction methods, or reducing hardware and bandwidth requirements, and subsequently verify their effectiveness through repeated testing with the affected user groups.




**Figure 3: Equitable user experiences**

To mitigate algorithmic bias in AI-XR systems (RQ2) that may disproportionately affect specific user groups considering sensitive attributes, researchers and developers need to adopt a comprehensive and multi-layered approach. Central to this is the use of diverse and representative training datasets that accurately reflect the demographic and cultural heterogeneity of the target user population, thereby reducing the risk of exclusionary model behaviour. This calls for regular and rigorous fairness testing mechanisms throughout the AI model life cycle for detecting and correcting disparities in outcomes, employing fairness metrics such as demographic parity and equalized odds. Incorporating fairness-aware algorithms that embed constraints to promote equitable treatment during model training can further reduce bias. At the same time, the involvement of multidisciplinary and diverse development teams, alongside inclusive user testing with affected communities, ensures that multiple perspectives inform the design, development, and evaluation processes, helping to identify subtle or emergent biases early in the life cycle. In addition to this, both the developers and users need to be able to understand and provide feedback on system behaviour. Moreover, continuous monitoring and updating the datasets and models used are necessary to address bias that may arise over time due to data or model drift, or due to evolving user interactions. Techniques such as generating synthetic data to balance underrepresented groups can also be employed cautiously to enhance dataset inclusivity. These aspects are depicted in Figure 4 and reflect the importance of the responsibility that the researchers and developers have on this behalf to implement technical and procedural strategies to foster fairness and equitable outcomes in AI-XR systems.

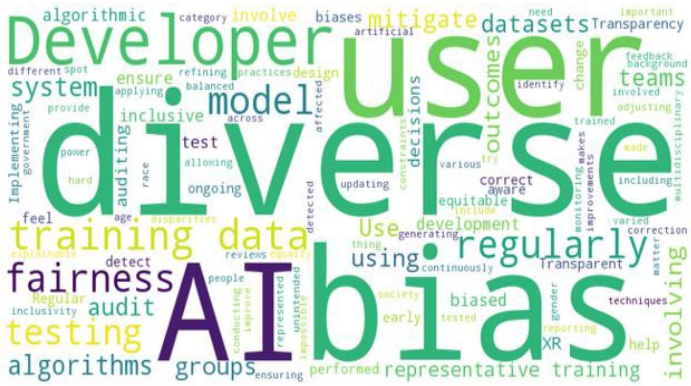


**Figure 4: Bias mitigation**

To ensure a fair representation in the data used to train AI models for XR applications (RQ3), researchers and developers need to design and adopt strategies that prioritize diversity, balance, and continuous validation throughout the data life cycle. This begins with collecting data from a wide range of sources that encompass aspects such as varied demographics, physical abilities, cultural backgrounds, and socioeconomic contexts, thereby avoiding over-reliance on any single group. On this behalf, intentional efforts to include underrepresented and marginalized communities through collaboration and ethical engagement are fundamental to capturing authentic and equitable user experiences. To this end, the application of bias detection techniques to identify imbalances or gaps in datasets, applying augmentation or re-sampling methods when necessary to achieve demographic balance. Further, by involving domain experts and inclusivity specialists in the data validation process, ensures that the collected data accurately reflects real world diversity and addresses potential cultural or contextual nuances. Moreover, regular updates need to be considered for the datasets to keep pace with evolving populations and usage patterns, maintaining fairness over time. At the same time, transparency in data sourcing, along with adherence to ethical guidelines and best practices for responsible AI, further supports trustworthiness and accountability. These aspects are illustrated in Figure 5 and further considerations could be given to the use of large-scale curated datasets for both educational an design purposes as they exemplify these principles by covering diverse XR platforms, application types, and user interactions, facilitating robust and inclusive AI model training for XR environments.

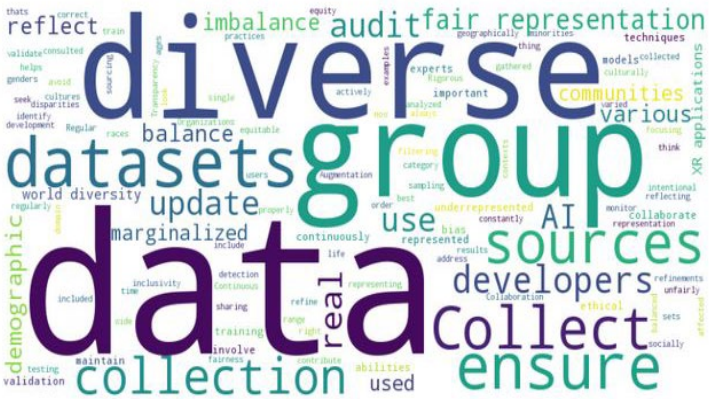


**Figure 5: Fair representation**

In order to design AI-XR systems that can adapt to different cultural contexts (RQ4), culturally aware design principles that embed multilingual support, localization, and user customization need to be accounted. This involves developing interfaces and content that respect and reflect local cultural norms, symbols, gestures, and communication styles, which vary widely across regions, e.g., adjusting interaction patterns to align with culturally preferred gestures or voice commands and adapting visual elements like colours and icons to avoid misinterpretation or offense. Incorporating diverse, culturally representative datasets during AI model training ensures that system behaviours are relevant and sensitive to various user populations. At the same time, by engaging local experts and diverse user groups throughout design, development, and testing phases, it facilitates the identification and mitigation of cultural biases and existing or potential stereotypes. In addition, real-time language translation and adaptive feedback mechanisms enable personalized experiences that honour individual cultural preferences while preserving equitable access. Also here, regular audits and

iterative feedback loops further support continuous improvement in cultural fairness. These aspects are shown in Figure 6 and further show that Ai-based immersive cultural sensitivity training demonstrates how AI and XR combined can foster deeper cross-cultural understanding and empathy, enhancing user engagement and inclusivity.




**Figure 6: Cultural dimensions**

To detect and address unfair outcomes or discriminatory behaviours in AI-XR environments (RQ5), a comprehensive mechanism combining real-time monitoring, bias detection algorithms, user feedback systems, and transparent reporting need be developed and deployed. On this behalf, real-time monitoring tools could be used to continuously analyse AI model outputs to identify anomalies and disparities related to protected attributes such as race, gender, or age, enabling prompt detection of biased behaviours before they propagate in the system. At the same time, bias detection modules could embed ensemble learning frameworks and adversarial training to uncover hidden and intersectional biases by comparing outputs across primary and auxiliary evaluators. User feedback channels empower individuals to report perceived unfairness, facilitating human-in-the-loop oversight and enriching bias identification with lived experiences. In addition, it is recommended to consider regular fairness audits conducted by independent experts to ensure systematic evaluation and correction of discriminatory patterns. Moreover, adaptive mitigation strategies such as reinforcement learning driven dynamic reweighting has the potential to adjust model parameters in real time to counteract unfairness while preserving system performance. On this behalf, ethical AI guidelines and clear antidiscrimination protocols that guide these processes, ensuring accountability and transparency need to be accounted as well added to privacy preserving techniques like differential privacy and data anonymization which safeguard user data during bias analysis. These aspects are captured in Figure 7 and further assure the promotion of continuous fairness, inclusivity, and trustworthiness in AI-XR systems.




**Figure 7: Unfair outcomes**

The evaluation of fairness in AI-XR systems (RQ6) can be done using a combination of quantitative and qualitative metrics that capture both group-level equity and individual experience. To this end, the use of core metrics such as demographic parity is recommended, as it assesses whether outcomes are distributed equally across sensitive groups, and equalized odds or equal opportunity, which assess whether prediction accuracy

and error rates are consistent across these groups. At the same time, disparate impact quantifies the extent to which outcomes disproportionately affect certain demographics, while individual fairness ensures that similar users receive similar system responses. In addition, user satisfaction surveys and accessibility scores provide valuable insights into perceived fairness, capturing experiential disparities that may not be evident in algorithmic outputs. Other recommended metrics are error rate gaps, representation balance, and inclusion indexes in order to help monitoring the system performance across diverse user populations, while bias impact scores and counterfactual fairness can identify and quantify subtle or intersectional biases. To these aspects captured in Figure 8 are added continuous monitoring and transparent reporting of these metrics, in line with ethical and regulatory standards for building trust and ensuring that AI-XR systems deliver equitable and accessible experiences for all users. For example, an AI-XR training system can compare recognition accuracy, task-completion time, interaction failures, motion discomfort, and satisfaction scores across gender, age, disability, language, and socio-economic groups under different devices and network conditions. Any statistically or practically significant gap needs then to trigger a documented investigation, targeted mitigation, and subsequent re-evaluation with affected users to confirm that the intervention has reduced the disparity without creating new disadvantages elsewhere.

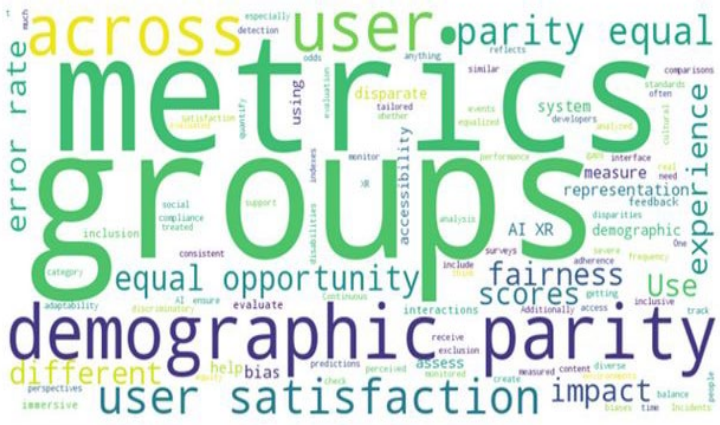


**Figure 8: Fairness evaluation.**

## 5. Conclusions

The findings of this research show that building fairness in AI-XR systems requires a multidisciplinary and multidimensional approach that spans data collection, model development, system design, and ongoing evaluation. To these are added, ensuring equitable experiences across diverse user groups involves designing for low-cost hardware compatibility, adaptive and accessible interfaces, and personalized content that respects different physical abilities, socioeconomic backgrounds, and cultural contexts. At the same time, the mitigation of algorithmic bias needs to be done by curating diverse and representative datasets, employing fairness-aware algorithms, and conducting regular audits, while involving multidisciplinary teams and affected communities in both design and testing. Moreover, fair representation in training data can be done through intentional inclusion of marginalized groups, systematic auditing for imbalance, continuous dataset updates, real-time monitoring, and user feedback channels. Moreover, unlike traditional AI systems, fairness in AI-XR needs to account for both disparities in algorithmic outputs and for biases arising from embodied, spatial, and multimodal interaction. These include embodiment bias in avatar representation, spatial interaction unfairness caused by differences in mobility or physical reach, and sensor bias resulting from variable performance of gaze, gesture, voice, and motion-tracking technologies across user groups. Hence, fairness evaluation in AI-XR should incorporate both computational outcomes and the accessibility, accuracy, comfort, and inclusiveness of users' immersive experiences.

Hence, building fair AI-XR systems demands both technical innovation and sustained human-centred engagement. From here, future research could focus on developing standardized, context-sensitive fairness metrics tailored to the unique challenges of immersive, interactive environments, enabling more precise and actionable evaluations. Another future research path could be dedicated to building scalable, real-time bias mitigation and detection frameworks that can operate within the dynamic and adaptive nature of XR systems using XAI methods and reinforcement learning mechanisms. These directions could help ensure that AI-XR systems not only avoid replicating societal biases, but actively contribute to more responsible, trustworthy, and equitable digital experiences.

## Acknowedgement

We are grateful for the valuable involvement and insights provided by the experts involved in this research and for the important infrastructure that was made available to us in this process.

**Ethics and AI Declaration**: For this research, no ethical clearance is required and no AI tools were used in the creation of this article.

## References

Akhtar, Z.B., 2024. Artificial intelligence (ai) and extended reality (xr): A biomedical engineering perspective investigation analysis. Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics, 6(3).

Alderman, J.E., Palmer, J., Laws, E. and McCradden, 2025. Tackling algorithmic bias and promoting transparency in health datasets: The standing together consensus recommendations. The Lancet Digital Health, 7(1), pp. e64-e88.

Alkaeed, M., Qayyum, A. and Qadir, J., 2024. Privacy preservation in artificial intelligence and extended reality (ai-xr) metaverses: A survey. Journal of Network and Computer Applications, p. 103989. Available at: https://doi.org/10.1016/j.jnca.2024.103989.

Albahri, A.S., Duhaim, M.M., Fadhel, M.A., Alnoor, A., Baqer, N.S., Alzubaidi, L. and Deveci, M., 2023. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion, 96, pp. 156-191.

Angerschmid, A., Zhou, J., Theuermann, K., Chen, F. and Holzinger, A., 2022. Fairness and explanation in ai-informed decision making. Machine Learning and Knowledge Extraction, 4(2), pp. 556-579.

Bang, J. and Kim, J.-Y., 2023. Metaverse ethics for healthcare using ai technology: Challenges and risks. In: International Conference on Human-Computer Interaction. Cham: Springer Nature Switzerland, July 2023, pp. 367-378.

Baskir, R., Lee, M., McMaster, S.J., Lee, J., Blackburne-Proctor, F., Azuine, R. et al., 2025. Research for all: Building a diverse researcher community for the all of us research program. Journal of the American Medical Informatics Association, 32(1), pp. 38-50.

Baum, S.D., 2021. Artificial interdisciplinarity: Artificial intelligence for research on complex societal problems. Philosophy & Technology, 34(Suppl 1), pp. 45-63.

Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R. and Zhang, Y., 2019. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), pp. 4-1.

Blegen, M., Wirkus, A.L., Wagner, S.J., Meyer, V.A., Cicek, J.G., Cicek, M.S., All of Us Research Demonstration Project Teams, Choi, S.H., Wang, X., Rosenthal, E.A. et al., 2024. Genomic data in the all of us research program. Nature, 627(8003), pp. 340-346.

Calvi, A. and Kotzinos, D., 2023. Enhancing ai fairness through impact assessment in the european union: A legal and computer science perspective. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, June 2023, pp. 1229-1245.

Christ-Brendemühl, S. and Schaarschmidt, M., 2022. Customer fairness perceptions in augmented reality-based online services. Journal of Service Management, 33(1), pp. 9-32.

Cools, K., & Maathuis, C. (2024, October). Trust or bust: Ensuring trustworthiness in autonomous weapon systems. In *MILCOM 2024-2024 IEEE Military Communications Conference (MILCOM)* (pp. 182-189). IEEE.

Fox, D. and Thornton, I.G., 2022. White paper - the IEEE global initiative on ethics of extended reality (xr) report: Extended reality (xr) ethics and diversity, inclusion, and accessibility. Tech. Rep., The IEEE Global Initiative on Ethics of Extended Reality (XR), white paper.

Guthula, J.S.N. and Basu, A., 2024. Navigating gender biases in xr: Towards equitable technological future. In: 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). IEEE, March 2024, pp. 326-329.

Hadan, L., Choong, L., Zhang-Kennedy, L. and Nacke, L.E., 2024. Deceived by immersion: A systematic analysis of deceptive design in extended reality. ACM Computing Surveys, 56(10), pp. 1-25.

Hirzle, T., Müller, F., Draxler, F., Schmitz, M., Knierim, P. and Hornbæk, K., 2023. When xr and ai meet - a scoping review on extended reality and artificial intelligence. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Hamburg, Germany: Association for Computing Machinery, pp. 1-45.

Huynh-The, T., Pham, Q.-V., Pham, X.Q., Nguyen, T.T., Han, Z. and Kim, D.S., 2023. Artificial intelligence for the metaverse: A survey. Engineering Applications of Artificial Intelligence, 117, p. 105581.

Jacob, C., Brasier, N., Laurenzi, E., Heuss, S., Mougiakakou, S.-G., Çöltekin, A. and Peter, M.K., 2025. Ai for impacts framework for evaluating the long-term real-world impacts of ai-powered clinician tools: Systematic review and narrative synthesis. Journal of Medical Internet Research, 27, p. e67485.

John-Mathews, J.-M., Cardon, D. and Balagué, C., 2022. From reality to world. a critical perspective on ai fairness. Journal of Business Ethics, 178(4), pp. 945-959.

Kajla, T., Kansra, P. and Singh, N. eds., 2024. Multidisciplinary Applications of Extended Reality for Human Experience. IGI Global.

Kallina, E., Bohné, T. and Singh, J., 2025. Stakeholder participation for responsible ai development: Disconnects between guidance and current practice. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, June 2025, pp. 1060-1079.

Kattnig, M., Angerschmid, A., Reichel, T. and Kern, R., 2024. Assessing trustworthy ai: Technical and legal perspectives of fairness in ai. Computer Law & Security Review, 55, p. 106053.

Kaur, D., Uslu, S., Rittichier, K.J. and Durresi, A., 2022. Trustworthy artificial intelligence: A review. ACM Computing Surveys (CSUR), 55(2), pp. 1-38.

Khera, R., Simon, M.A. and Ross, J.S., 2023. Automation bias and assistive ai: Risk of harm from ai-driven clinical decision support. JAMA, 330(23), pp. 2255-2257.

Kiourtis, A., Mavrogiorgou, A., Makridis, G., Kyriazis, D., Soldatos, J., Fatouros, G. and Falsetta, M., 2024. Xr5.0: Human-centric ai-enabled extended reality applications for industry 5.0. In: 2024 36th Conference of Open Innovations Association (FRUCT). IEEE, October 2024, pp. 314-323.

Maathuis, C., 2024. Human-centred ai in military cyber operations. In: J. du Toit and B. van Niekerk, eds. Proceedings of the 19th International Conference on Cyber Warfare and Security (ICCWS 2024). Johannesburg, South Africa: Academic Conferences and Publishing Limited, March 2024.

Maathuis, C., Cidota, M.A., Datcu, D. and Marin, L., 2025. Integrating explainable artificial intelligence in extended reality environments: A systematic survey. Mathematics, 13(2), p. 290.

Maathuis, C., Van Den Steen, K. and van Dijke, R., 2025. Virtual reality training prototype for police-bystander conflict. In: 2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). IEEE, March 2025, pp. 313-315.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), pp. 1-35.

Menezes, H.F., Ferreira, A.S., Pereira, E.T. and Gomes, H.M., 2021. Bias and fairness in face detection. In: 2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE, October 2021, pp. 247-254.

Oladoyinbo, T.O., Olabanji, S.O., Olaniyi, O.O., Adebiyi, O.O., Okunleye, O.J. and Ismaila Alao, A., 2024. Exploring the challenges of artificial intelligence in data integrity and its influence on social dynamics. Asian Journal of Advanced Research and Reports, 18(2), pp. 1-23.

Pfeiffer, J., Gutschow, J., Haas, C., Möslein, F., Maspfuhl, O., Borgers, F. and Alpsancar, S., 2023. Algorithmic fairness in ai: An interdisciplinary view. Business & Information Systems Engineering, 65(2), pp. 209-222.

Qayyum, A., Butt, M.A., Ali, H., Usman, M., Halabi, O., Al-Fuqaha, A., Abbasi, Q.H., Imran, M.A. and Qadir, J., 2024. Secure and trustworthy artificial intelligence-extended reality (ai-xr) for metaverses. ACM Computing Surveys, 56(7), pp. 1-38.

Ratican, J., Hutson, J. and Wright, A., 2023. A proposed meta-reality immersive development pipeline: Generative ai models and extended reality (xr) content for the metaverse. Journal of Intelligent Learning Systems and Applications, 15, pp. 24-35.

Reddy, S., Rogers, W., Mäkinen, V.-P., Coiera, E., Brown, P., Wenzel, M., Weicken, E., Ansari, S., Mathur, P., Casey, A. and Kelly, B., 2021. Evaluation framework to guide implementation of ai systems into healthcare settings. BMJ Health & Care Informatics, 28(1), p. e100444.

Rudnicka, Z., Proniewska, K., Perkins, M. and Pręgowska, A., 2024. Cardiac healthcare digital twins supported by artificial intelligence-based algorithms and extended reality-a systematic review. Electronics, 13(5), p. 866.

Shaikh, T.A., Dar, T.R. and Sofi, S.A., 2022. A data-centric artificial intelligent and extended reality technology in smart healthcare systems. Social Network Analysis and Mining, 12(1), p. 122.

Shrotriya, N., Banu, N., Aulkarni, A. and Aiyangar, S., 2025. Navigating the path to ethical & responsible ai integration in health & life sciences with human and machine collaboration. Journal of the Society for Clinical Data Management, 5(1), p. 5.

Stacchio, L., Pierdicca, R., Paolanti, M., Zingaretti, P., Frontoni, E., Giovanola, B. and Tiribelli, S., 2024. Xrai-ethics: Towards a robust ethical analysis framework for extended artificial intelligence. In: 2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). IEEE, October 2024, pp. 214-219.

Tabassum, A. and Qadir, J., 2024. A critical ai approach for managing the risks in educational metaverses. In: EDULEARN24 Proceedings. IATED, pp. 9578-9586.

Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R. and Naganawa, S., 2024. Fairness of artificial intelligence in healthcare: review and recommendations. Japanese Journal of Radiology, 42(1), pp. 3-15.

Varona, D. and Suárez, J.L., 2022. Discrimination, bias, fairness, and trustworthy ai. Applied Sciences, 12(12), p. 5826.

Wheeler, R. and Carroll, F., 2023. An explainable ai solution: Exploring extended reality as a way to make artificial intelligence more transparent and trustworthy. In: Proceedings of the International Conference on Cybersecurity, Situational Awareness and Social Media: Cyber Science 2022; 20-21 June; Wales. Springer Nature Singapore, pp. 255-276.

Zhang, W., 2024. Ai fairness in practice: Paradigm, challenges, and prospects. AI Magazine, 45(3), pp. 386-395.