Designing Responsible AI: The Power of User-Selected Metrics

Tuva Falk

Department of Computing Science
Umeå University, Sweden
id21tfk@cs.umu.se

Abstract. Effective artificial intelligence (AI) system design requires alignment with user expectations, yet the development process often prioritizes technical perspectives over user needs. This study examines the integration of user-selected user experience (UX) evaluation metrics within the participatory design process, emphasizing their role in shape design decisions and system assessment. Through a mixed-methods approach with 66 participants, including AI developers, designers, and end-users, we analyze correlations between design features, design characteristics, and UX evaluation metrics. Additionally, the study examines differences in stakeholder priorities, revealing variations in how participants value personalization, adaptability, and ethical considerations such as privacy. The findings suggest that incorporating user-driven evaluation criteria from the early stages of development can lead to more transparent, inclusive, and user-centered AI systems.

1 Introduction

Responsible artificial intelligence (AI) emphasizes inclusivity, equity, transparency, and respect for human rights [1, 2]. While Responsible AI frameworks provide a strong ethical foundation, defining the core principles of responsible AI, implementing these principles in practice remains a significant challenge, particularly in ensuring AI systems interact fairly and transparently with diverse user groups [3]. One approach to designing more responsible AI includes involving end-users and other stakeholders in the design process through methodologies such as participatory design [4, 5].

Participatory design is a democratic methodology that empowers users to actively contribute to shaping the systems they interact with [6]. Despite its potential, this approach faces critical gaps in the context of AI development, particularly in its early stages. Current practices often lack clear methodologies for integrating users in the process of selecting or defining evaluation metrics, which are crucial for aligning system design with diverse user needs and preferences. Furthermore, technical stakeholders frequently dominate the design process, which can lead to systems that fail to adequately reflect the priorities of end-users. These gaps risk undermining user trust and the effectiveness of AI systems [4, 7, 8].

The objective of this study is to explore how integrating user-selected UX evaluation metrics into participatory design processes can enhance the development and design of recommender systems that align with both technical goals and user-centered design principles. By examining the preferences of two key stakeholder groups, AI developers and end-users, the study seeks to identify actionable insights for balancing diverse priorities in responsible AI system design. To achieve this, a mixed-methods approach was employed, combining quantitative analysis of correlations between design features and evaluation metrics with qualitative insights from stakeholders. By comparing the priorities of AI developers and end-users, this study seeks to uncover strategies for reconciling these perspectives. The study aims to investigate whether user-selected evaluation metrics can serve as valuable design materials within an expanded participatory design framework.

This paper is organized as follows: Section 2 reviews existing literature on responsible AI development and participatory design, highlighting the theoretical underpinnings of this research. Section 3 outlines the research methodology, detailing the mixed-methods approach used to collect and analyze data. Section 4 presents the study's findings, focusing on the correlations observed between user-selected metrics and design features, as well as differences in priorities between stakeholders. Section 5 discusses the implications of these findings for user-centered AI design and identifies key trade-offs and limitations.

By addressing critical gaps in participatory design practices, this study contributes to the growing discourse on responsible AI by providing empirical evidence on the integration of user-selected UX evaluation metrics into AI design processes. It demonstrates how these metrics can systematically align system features with user preferences, improving both usability and inclusivity. Additionally, the study identifies strategies for balancing the often-divergent priorities of developers and end-users, offering insights into the trade-offs required to foster transparency, fairness, and user trust. By presenting correlations between design features and evaluation metrics, and uncovering key differences in stakeholder priorities, this research lays a foundation for more effective participatory design methodologies that re-democratize the design process, prioritizing both technical and ethical considerations.

2 Earlier Work

Understanding how user-selected evaluation metrics influence participatory design processes requires grounding the study in existing research on responsible artificial intelligence (AI), user experience (UX) design, and participatory design methodologies. This section reviews foundational frameworks and prior studies that inform the integration of UX metrics into participatory design, highlighting key gaps and opportunities addressed by this research.

2.1 User Experience in responsible AI Development

In recent years, AI tools such as algorithm-based recommender systems have revolutionized UX by enabling personally tailored interactions with technology. However, AI-induced user experiences involve unique design processes that differ significantly from conventional UX approaches. One major challenge is the uncertainty of AI system outputs, as users can only evaluate whether a recommender system works for them on a case-by-case basis [9]. Despite these unique challenges, current UX-design practices used in AI largely mirror those used for non-AI systems, prompting researchers to explore whether AI-specific design practices are needed. Yang et al. [10] proposed a framework to aid human-computer interaction practitioners address these challenges, emphasizing two attributes that make AI design uniquely difficult: capability uncertainty and output complexity.

The UX of digital systems is pivotal for building trust and reducing errors, but traditional UX principles, like Nielsen Norman's 10 [11], fall short for AI design. While these guidelines emphasize usability and user control, they neglect key principles of responsible AI, such as transparency, privacy, and feedback loops. As Bodegraven [12] argues, traditional UX frameworks are insufficient for addressing the complexities of AI systems, which demand new heuristics for truly human-centered design. However, the fundamental UX-design practices have demonstrated their value in shaping AI systems around user needs, minimizing misalignment with workflows, and operationalising principles like fairness in real-world contexts [9, 13, 14, 15].

Balancing user-friendly design with responsible AI principles, such as transparency, interpretability, and privacy protection, is a central challenge in AI system development. As automation and algorithmic decision-making advance, ensuring trust and user agency while mitigating bias and promoting fairness remains critical. Addressing these challenges requires interdisciplinary collaboration and diverse stakeholder engagement throughout the AI life-cycle [16, 5, 17, 18].

2.2 A New Dimension to Participatory Design

Participatory design is a co-design methodology emphasizing democratic involvement, equality, and inclusivity, where diverse stakeholders collaborate to shape technologies [6].

Hansen et al. [19] introduced a Program Theory framework to analyze how participatory design connects inputs (resources and knowledge) to outputs (including design products and long-term impacts) through key mechanisms and activities in the design process. These mechanisms, conceptualized as fundamental principles, include fostering balanced power relations among stakeholders, promoting mutual learning, and creating a sense of ownership over the design process. By empowering users and ensuring design outputs align with stakeholder values, participatory design sustains engagement and has the potential to addresses challenges in responsible AI, such as fairness, transparency, and

accountability. The outcomes of participatory design extend beyond immediate products to include lasting, systemic changes that benefit participants and their communities, aligning closely with the goals of responsible AI to foster inclusivity and equity [19].

A critical challenge in AI development is the exclusion of stakeholders early in the design process, often before key decisions about evaluation metrics or the system's necessity are made [20]. This oversight risks creating systems that prioritize narrow objectives over broader societal needs [4]. Participatory Design offers a potential solution by engaging stakeholders throughout the life-cycle, ensuring that the system reflects users preferred criteria and societal values.

To better incorporate and measure diverse stakeholder preferences, Zheng and Huang propose integrating UX evaluation metrics as design inputs [14]. Their study demonstrates that incorporating these metrics during the input phase of participatory design enhances AI system development. By revealing correlations between evaluation metrics, design characteristics, and features, UX evaluation metrics help designers make more informed, user-driven decisions. They also empower users to articulate key design values early in the process, fostering inclusivity and responsiveness to diverse needs. For example, their findings indicate that preferences for features like empathy or emoji use vary depending on the evaluation metrics applied, underscoring the need for adaptable evaluation frameworks. By expanding the Program Theory framework [19] to integrate UX evaluation metrics, their work lays the foundation for more inclusive and effective participatory design in AI.

3 Methodology

This section describes the study structure, the design materials included in the survey, and the method of data analysis.

3.1 Study Overview

This study builds upon the work of Zheng and Huang [14], who integrated user experience (UX) evaluation metrics into the Program Theory model [19] to enhance participatory design for AI systems. Drawing on their adaptation, this research applies UX evaluation metrics specifically to recommender systems, investigating their role in aligning system design with stakeholder needs. Additionally, this study examines how two key stakeholder groups, AI developers and end-users, differ in their design preferences. By comparing how technical and non-technical participants prioritize UX metrics, this research provides insights into the diverse expectations shaping AI system development, contributing to more user-centered and balanced design approaches.

A mixed-method approach was employed, combining quantitative and qualitative data collection through a structured survey. Quantitative data aims to uncover correlations between design features, design characteristics, and evaluation metrics, while qualitative data explores thematic patterns across different stakeholder groups.

In addition to the design materials the survey collected information about participants experience with AI systems asking about experience with recommender systems, frequency of interaction with AI-based applications and experience with AI design or development. As well as demographic information about level of education, years of work experience, and country of origin.

Participants were divided into two groups:

- **AI Developers:** Individuals with experience in AI system development or design, including industry professionals and students in technical fields workimng with AI such as Computer Science.
- **End-users:** Non-technical participants who were likely to interact with the AI system but lacked experience in AI development or design.

The study includes 28 participants with prior experience in AI development or design and 38 participants who do not have such experience, providing a balanced representation. The participants were recruited through professional networks, online communities, local organizations, and on-campus resources at Umeå University.

3.2 Design Materials

Following Zheng and Huang's adaptation of the Program Theory model, UX evaluation metrics were included as part of the design materials [14]. The design inputs were collected using a survey with questions presented in clear and accessible language to accommodate participants of all technical backgrounds. Following Zheng and Huang's methodology, participants rated each inputs importance on a scale of 1 to 5, with 5 being "Essential" and 1 being "Not important". Participants were encouraged to explain their reasoning behind selecting specific metrics, offering qualitative insights into their choices. The following section outlines the parts included in the three major design inputs.

- **Design Characteristics (DC).** This section explores user preferences for high-level system attributes and includes 6 questions adapted from previous research on recommender systems. Free-form questions about preferred platform integration [21], dynamic personalization [21, 22, 23], recommendation transparency [21, 23], diversity in recommendations [21, 24, 25], for example "*Would you like a recommender system to suggest diverse and novel items, and how frequently should this occur?* ". It also includes interval scale questions for cross-platform synchronization of suggestions [21] and the preferred roles of recommender systems on platforms.
- **Design Features (DF).** This section focuses on specific functional elements of recommender systems. The respondents were prompted to rate the importance of 12 design features on a scale from 1 (Not Important) to 5 (Essential). The design features included in the survey were derived from prior research on recommender systems. The included design features were: recommendations based on previous behavior [21, 22, 23], ability to provide feedback on recommendations [26, 27], suggestions of diverse and novel items

[25], customization of recommendation criteria [27] adaptation to feedback or emotional cues (e.g., dislikes), secure handling of sensitive information [28], quick and accurate updates based on user input, explanations of how and why recommendations are made [23, 29], mechanisms to report errors or issues, balancing preferences with ethical norms, clear and easy- to-navigate interface, unbiased and accurate recommendations.

– **User Experience Evaluation Metrics (EV).** This section evaluates the importance of specific UX evaluation metrics for recommender systems. The respondents were prompted to rate the importance of 15 evaluation metrics on a scale from 1 (Not Important) to 5 (Essential). Metrics include *user and algorithm behavioral data* derived from Zangerle and Bauer's [23] paper on recommender system evaluation metrics such as prediction accuracy, rate of correct recommendations, recommendation novelty, recommendation diversity, unexpectedness of recommendations, fairness of recommendation quality across users, fairness of recommendation frequency among items, number of redirects to action (e.g., purchases), time spent reporting errors, time engaging with the app. Additionally, EV include *user perceptions data* such as perceived effectiveness of the recommendations, perceived appropriateness of the recommendations, perceived usefulness of the recommendations, privacy concerns, perceived ease of use.

3.3 Data Analysis

The data analysis involved both qualitative and quantitative methods to gain a comprehensive understanding of participant responses. We began by analyzing the open-ended responses to capture an overall understanding of participant experiences and insights. Thematic analysis was conducted on the open-ended responses to identify key themes across participant feedback and compared them across both groups.

– **Correlation Analysis**: To examine the relationship between DC, DF and EV, we applied Spearman's correlation coefficient. This test assessed the strength and direction of associations between these variables, providing insights into which characteristics and features were most aligned with participant evaluations.
– **Group Comparisons**: We used the Mann-Whitney U test to compare the ratings between the participant groups, analyzing if participants with AI development or AI design experience and those without showed significant differences in their ratings for design features and evaluation metrics.

4 Results

The correlation analysis examined participants' ratings of design characteristics (DCs), design features (DFs), and evaluation metrics (EVs) using a 5-point scale ranging from 1 (Not Important) to 5 (Essential). For example, participants answered questions such as, "*Do you think a recommender system should be able*

to explain why it suggested a particular item to you? " This analysis quantified the strength and direction of correlations between DCs, DFs, and EVs, providing insights into how specific design elements influence evaluation priorities. Additionally, a thematic analysis was conducted to identify key themes in participant feedback, offering deeper insights into the reasoning behind their responses.

4.1 Correlation of Evaluation Metrics and Design Alternatives

As shown in Figure 1, the design feature with the highest number of correlations were *Quick and accurate updates based on user input* (DF7) showing strong correlations with 6 EVs. The design characteristic *Personalizing recommendations based on user preferences* (DC6-b) exhibited the highest number of intersections with evaluation metrics, correlating with 5 EVs. The design feature with the highest number of correlations were *Quick and accurate updates based on user input* (DF7) showing strong correlations with 6 EVs.

	DF1 Recommendations based on previous behavior	DF2 Ability to provide feedback on recommendations	DF3 Suggestions of diverse and novel items	DF4 Customization of recommendation criteria	DF5 Adaptation to feedback or emotional cues	DF6 Secure handling of sensitive information	DF7 Quick and accurate updates based on user input	DF8 Explanations of how recommendations are made	DF9 Mechanisms to report errors or issues	DF10 Balancing preferences with ethical norms	DF11 Easy-to-navigate interface	DF12 Unbiased and accurate recommendations	DC1 Platform Integration	DC2 Adaptive Algorithm	DC3 Explainability	DC4 Frequency of Diversity	DC5 Multi-Platform Integration	DC6 -a Finding new items	DC6 -b Personalizing recommendations based on user preferences	DC6 -c Improving user engagement with the platform	DC6 -d Suggesting connections or shared preferences	DC6 -e Supporting decision-making	Count
EV1 - Prediction accuracy																						0,32*	1
EV2 - Precision of Recommendations	0,36**						0,27*	0,27*											0,28*				4
EV3 - Recommendation novelty			0,44**	0,27*			0,29*											0,26*					4
EV4 - Recommendation diversity			0,52**																				1
EV5 - Serendipity			0,32**																				1
EV6 - Equal quality across users								0,37**				0,41**											2
EV7 - Fairness of recommendations				0,26*		0,26*				0,39**		0,33**									0,27*		5
EV8 - Number of redirects to action																					0,25*		1
EV9 - Time spent reporting errors																							0
EV10 - Time engaging with the app																			0,30*				1
EV11 - Perceived effectiveness					0,35**		0,40**				0,42**	0,28*			0,26*			0,31*	0,31*				7
EV12 - Perceived appropriateness	0,25*				0,36**		0,27*			0,34**					0,34**			0,26*	0,36**				7
EV13 - Perceived usefulness	0,36**				0,40**		0,33**			0,33**	0,30*				0,31*				0,38**			0,30*	8
EV14 - Privacy concerns						0,39**			0,31											0,27*			3
EV15 - Perceived ease of use	0,29*						0,33**			0,39**	0,38**									0,26*			4
Count	4	0	3	2	3	2	6	2	1	4	3	3	0	0	3	0	0	3	5	2	2	2	

**p<0.01 *p<0.05

Fig. 1. Significant correlations of Design Features (DF) and Design Characteristics (DC) with Evaluation Metrics (EV).

4.2 Correlations Between Design Features and Evaluation Metrics

The design feature *Quick and accurate updates based on user input* (DF7) demonstrated significant positive correlations with several evaluation metrics, including *Perceived Effectiveness* (EV11, $r = 0.40$), *Perceived Usefulness* (EV13, $r = 0.33$), and *Perceived Ease of Use* (EV15, $r = 0.33$). This indicates that adaptiveness, as valued by users, is not strictly a technical measure of system responsiveness but rather a subjective experience, how well the system seems to react and align with user expectations. Notably, this finding highlights an important distinction: rather than relying solely on traditional technical benchmarks for evaluation, DF7 may be more meaningfully assessed through user perception metrics, aligning with how users subjectively experience responsiveness in AI systems [23].

The thematic analysis further supports this, revealing that participants highly value systems that feel responsive and intuitive. One participant expressed this sentiment explicitly: "If I dislike something, I would hope that similar items are not shown" emphasizing the importance of immediate feedback in shaping the user experience. This preference underscores an experience heavy perspective, users evaluate responsiveness based on their perception of adaptability rather than strictly measurable technical parameters. Similarly, participants conveyed frustration with static or unresponsive systems, as illustrated by one comment: "I want to see changes based on my actions rather than being forced to engage in irrelevant content." These insights reinforce the broader expectation that adaptive, user-centered design enhances trust and engagement in recommender systems.

4.3 Correlations Between Design Characteristics and Evaluation Metrics

The results reveal moderate positive correlation between *Personalizing recommendations based on user preferences* (DC6-b) and *Time engaging with the app* (EV10, $r = 0.30$) suggesting that personalization may foster increased user engagement. This aligns with findings from the thematic analysis, where users described personalized systems as more intuitive and engaging.

Furthermore, stronger correlations with *Perceived Effectiveness* (EV11, $r = 0.31$) and *Perceived Appropriateness* (EV12, $r = 0.36$) underscore that trust and contextual relevance are critical dimensions for evaluating the user experience of personalized systems. Notably, the strongest correlation with *Perceived Usefulness* (EV13, $r = 0.38$) highlights that personalization is perceived as a key driver of system utility. Importantly, all of these evaluation metrics are based on user perception, reinforcing that assessing personalization through user-driven evaluation criteria, rather than solely technical performance, can offer valuable insights into system usability and alignment with user expectations from the early stages of design.

These correlations, while non-causal, provide insights into evaluation metrics for assessing the implementation of personalization features in recommender systems. The thematic analysis complements these findings, illustrating user expectations for systems that balance familiarity and novelty, foster trust through transparency, and deliver practical benefits. Quotes such as "[The system suggesting diverse and novel items is] very important because it saves me time looking for new products" and "If I show interest in a certain item, it means I am likely to get more suggestions on better similar items" further emphasize the importance of precision, engagement, and utility.

4.4 Differences in Stakeholder Preferences

The Mann-Whitney U test revealed significant differences between participants with AI development or AI design experience and non-developers in their evaluation of DF, DC, and EV. Those with experience of AI development or design

	Median		Mean		Mann-Whitney	
Metric	AI Developer or Designer (26)	Non AI Developer (36)	AI Developer or Designer (26).	Non AI Developer (36).	U	p-value
EV12 Perceived appropriateness**	5	4	4.58	3.94	683.5	0.001**
EV13 Perceived usefulness**	5	4	4.54	3.78	705.5	0.000**
EV14 Privacy concerns**	5	5	4.85	4.31	614.5	0.009**
DF1 Recommendations based on previous behavior*	4	4	4.19	3.42	616.0	0.029*
DF5 Adaptation to feedback or emotional cues**	5	4	4.42	3.64	652.0	0.006**
**p<0.01 *p<0.05						

Fig. 2. Signifcant Differences in Design Feature (DF), Design Charachteristics (DC) and Evaluation Metric (EV) Preferences between participants with AI design or development experience and those without.

consistently rated each variable with a significant difference higher than those without.

Significant differences were identified between stakeholder groups in their ranking of design features and evaluation metrics. For example, participants with AI development or design experience rated features such as *Recommendations based on previous behavior* (DF1, $p = 0.029$), and *Perceived appropriateness* (EV12, $p = 0.001$) to be of greater importance than the participants without. Additionally, developers and designers assigned higher importance to *Perceived usefulness* (EV13, $p < 0.001$) and *Privacy concerns* (EV14, $p = 0.009$), suggesting a heightened sensitivity to both the functional effectiveness and ethical considerations of AI systems.

5 Discussion

The findings of this study highlight the crucial role of integrating user-selected evaluation metrics (EV) at the outset of the participatory design of AI systems. The observed correlations imply that early incorporation of EVs helps streamline the design process, ensuring user preferences are aligned with the development of AI systems. Notably, this approach also provides valuable insights into how these design choices can be effectively evaluated.

First, integrating EVs into the participatory design enables a deeper understanding of users' values and preferences. As shown in *Section*4.2, significant correlations were observed between design features (DF) like *quick and accurate updates based on user input* (DF7) and EVs such as *perceived effectiveness* (EV11) and *perceived usefulness* (EV13). Identifying the evaluation metrics that are strongest associated with specific design options allows designers to gain valuable insights into what users prioritize most in those design features. This not only enhances the relevance of the design but also empowers users by ensuring the nuances of their preferences are considered from an early stage.

However, not all design features exhibit strong correlations with evaluation metrics. For instance, *ability to provide feedback on recommendations* (DF2) showed no correlation, suggesting that its independent effect on user satisfaction in this case may be marginal indicating the need for a holistic design approach.

Alternatively, the lack of correlation may imply challenges in assessing these features quantitatively.

Second, including EVs early on helps rationalize the design choices made by participants. As seen in this study, the correlation analysis provides a clear guide for understanding which elements of design are most likely to meet user expectations. This aids in justifying certain design decisions over others, ensuring that user needs are directly aligned with the resulting design solutions. Such an approach also fosters a more transparent and user-centered design process, directly reflecting users' design inputs.

The study also revealed interesting differences between the priorities of AI developers and end-users. Developers tended to emphasize systemic considerations such as *privacy concerns* (EV14) and *perceived appropriateness* (EV12), which can be attributed to their deeper understanding of AI systems and their associated regulatory and ethical responsibilities. This contrast does not necessarily suggest conflicting priorities between stakeholders but might reflect the distinct perspectives and strengths of developers and designers, who must balance technical robustness with user trust and ethical guidelines. Their focus on privacy, for example, underscores the importance of safeguarding user data and ensuring compliance with legal frameworks, which are critical in AI system development.

The broader implications of integrating user-selected metrics into the early stages of the participatory design process of AI are significant. These metrics not only help rationalize design decisions but also promote inclusivity by capturing a range of user preferences. By aligning design features with community values and expectations, AI systems can be designed to be not only user-friendly but also ethically responsible. The integration of such metrics operationalizes key principles of responsible AI, such as fairness, privacy, and transparency, providing measurable standards to guide the design and refinement of AI systems.

5.1 Study Limitations

The findings must be interpreted within the context of its limitations. The reliance on self-reported data introduces potential biases, such as social desirability or inaccurate recall. Additionally, the study's broad methodological scope may limit its generalizability, as participants' interpretations and experiences with specific AI systems may have varied significantly. Challenges in simplifying AI-specific terminology during survey design may have further impacted the clarity and reliability of responses. The comparison between user and developer priorities highlights this issue, as qualitative data shows that participants with technical AI experience provided more detailed answers, reflecting a deeper understanding of context and terminology. Finally, while the structured survey format was effective for data collection, it may have constrained participants' ability to articulate nuanced perspectives, particularly regarding trade-offs in design decisions.

5.2 Future Work

Building on these findings, future research could refine the participatory design methodology by systematically mapping stakeholder preferences to evaluation metrics, allowing for more tailored and transparent assessment frameworks. Given the observed differences in user priorities, further studies could explore strategies for reconciling divergent perspectives, ensuring that AI systems accommodate both technical feasibility and experiential expectations. Additionally, expanding the methodology to incorporate structured mechanisms for measuring and implementing ethical principles, such as fairness, transparency, and privacy, could enhance its applicability in responsible AI development. Longitudinal studies examining how user preferences evolve over time in real-world deployments would also provide valuable insights into the long-term impact of the participatory design framework in dynamic AI systems.

References

[1] European Union (EU): Eu ai act: First regulation on artificial intelligence (2023) Accessed: 2024-09-25.
[2] European Union (EU): High-level expert group on artificial intelligence (2023) Accessed: 2024-09-25.
[3] Hagendorff, T.: Blind spots in AI ethics. AI and Ethics **2**(4) (November 2022) 851–867
[4] Delgado, F., Yang, S., Madaio, M., Yang, Q.: The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. EAAMO '23, New York, NY, USA, Association for Computing Machinery (October 2023) 1–23
[5] Shams, R.A., Zowghi, D., Bano, M.: AI and the quest for diversity and inclusion: a systematic literature review. AI and Ethics (November 2023)
[6] Muller, M.J., Kuhn, S.: Participatory design. Communications of the ACM **36**(6) (June 1993) 24–28
[7] Wolf, C.T., Zhu, H., Bullard, J., Lee, M.K., Brubaker, J.R.: The Changing Contours of "Participation" in Data-driven, Algorithmic Ecosystems: Challenges, Tactics, and an Agenda. In: Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing, Jersey City NJ USA, ACM (October 2018) 377–384
[8] Shneiderman, B.: Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. ACM Trans. Interact. Intell. Syst. **10**(4) (October 2020) 26:1–26:31
[9] Ruiz, C., Quaresma, M.: The Participation of UX Designers in Artificial Intelligence Projects: Recommender Systems. Ergodesign & HCI **10**(1) (June 2022) 87–99 Number: 1.
[10] Yang, Q., Steinfeld, A., Rose, C., Zimmerman, J.: Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20, New York, NY, USA, Association for Computing Machinery (April 2020) 1–13

[11] Nielsen, J.: Enhancing the explanatory power of usability heuristics. In: Proceedings of the ACM CHI'94 Conference on Human Factors in Computing Systems, Boston, MA, USA (April 24–28 1994) 152–158
[12] Van Bodegraven, J.: How Anticipatory Design Will Challenge Our Relationship with Technology. (2017)
[13] Liao, Q.V., Vorvoreanu, M., Subramonyam, H., Wilcox, L.: UX Matters: The Critical Role of UX in Responsible AI. interactions **31**(4) (June 2024) 22–27
[14] Zheng, Q., Huang, Y.: "Begin with the End in Mind": Incorporating UX Evaluation Metrics into Design Materials of Participatory Design. In: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg Germany, ACM (April 2023) 1–7
[15] Wallach, D.P., Flohr, L.A., Kaltenhauser, A.: Beyond the Buzzwords: On the Perspective of AI in UX and Vice Versa. In Degen, H., Reinerman-Jones, L., eds.: Artificial Intelligence in HCI, Cham, Springer International Publishing (2020) 146–166
[16] Hall, P., Ellis, D.: A systematic review of socio-technical gender bias in AI algorithms. Online Information Review **47**(7) (November 2023) 1264–1279
[17] Tacheva, J., Ramasubramanian, S.: AI Empire: Unraveling the interlocking systems of oppression in generative AI's global order. Big Data & Society **10**(2) (July 2023)
[18] Aizenberg, E., Van Den Hoven, J.: Designing for human rights in AI. Big Data & Society **7**(2) (July 2020)
[19] Hansen, N.B., Dindler, C., Halskov, K., Iversen, O.S., Bossen, C., Basballe, D.A., Schouten, B.: How Participatory Design Works: Mechanisms and Effects. In: Proceedings of the 31st Australian Conference on Human-Computer-Interaction, Fremantle WA Australia, ACM (December 2019) 30–41
[20] Bossen, C., Dindler, C., Iversen, O.S.: Evaluation in participatory design: a literature survey. In: Proceedings of the 14th Participatory Design Conference: Full papers - Volume 1, Aarhus Denmark, ACM (August 2016) 151–160
[21] Gunawardana, A., Shani, G.: A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. Journal of Machine Learning Research (December 2009)
[22] Konstan, J.A., Riedl, J.: Recommender systems: from algorithms to user experience. User Modeling and User-Adapted Interaction **22**(1-2) (April 2012) 101–123
[23] Zangerle, E., Bauer, C.: Evaluating Recommender Systems: Survey and Framework. ACM Computing Surveys **55**(8) (August 2023) 1–38
[24] Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A.V., Turrin, R.: Looking for "Good" Recommendations: A Comparative Evaluation of Recommender Systems. In: Human-Computer Interaction - INTERACT 2011. Volume 6948. Springer Berlin Heidelberg, Berlin, Heidelberg (2011) 152–168
[25] Schedl, M., Gomez, E., Lex, E.: Trustworthy Algorithmic Ranking Systems. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, Singapore Singapore, ACM (February 2023) 1240–1243
[26] Pu, P., Chen, L.: A User-Centric Evaluation Framework of Recommender Systems. **612** (2010)
[27] Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C.: Explaining the user experience of recommender systems. User Modeling and User-Adapted Interaction **22**(4-5) (October 2012) 441–504
[28] Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., Kashef, R.: Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities. Applied Sciences **10**(21) (November 2020) 7748

[29] Gunawardana, A., Shani, G., Yogev, S.: Evaluating Recommender Systems. In Ricci, F., Rokach, L., Shapira, B., eds.: Recommender Systems Handbook. Springer US, New York, NY (2022) 547–601