Explainability and Contestability for the Responsible Use of Public Sector AI

Timothée Schmude
Faculty of Computer Science, Research Network Data Science, Doctoral School Computer Science
University of Vienna
Vienna, Austria
timothee.schmude@univie.ac.at

Abstract

Public institutions have begun to use AI systems in areas that directly impact people's lives, including labor, law, health, and migration. Explainability ensures that these systems are understandable to the involved stakeholders, while its emerging counterpart contestability enables them to challenge AI decisions. Both principles support the responsible use of AI systems, but their implementation needs to take into account the needs of people without technical background, *AI novices*. I conduct interviews and workshops to explore how explainable AI can be made suitable for AI novices, how explanations can support their agency by allowing them to contest decisions, and how this intersection is conceptualized. My research aims to inform policy and public institutions on how to implement responsible AI by designing for explainability and contestability. The Remote Doctoral Consortium would allow me to discuss with peers how these principles can be realized and account for human factors in their design.

CCS Concepts

• **Human-centered computing** → **Field studies**.

Keywords

explainable AI, contestable AI, algorithmic fairness, interdisciplinary research, qualitative methods

ACM Reference Format:
Timothée Schmude. 2025. Explainability and Contestability for the Responsible Use of Public Sector AI. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3706599.3721096

1 Foreword and overview

I am a Ph.D. student in explainable AI (XAI) at the University of Vienna in the fourth year of study during the consortium. My thesis is part of the research project "Interpretability and Explainability as Drivers to Democracy", supervised by Sebastian Tschiatschek, Torsten Möller, and Laura Koesten, and projected to be completed by autumn 2026.

My thesis' main overarching research question is **how explanations can help AI novices understand and contest AI systems that are implemented in public institutions**. Explainability and contestability are seen to support the responsible use of AI by improving transparency, human autonomy, and accountability [1, 17]. But their implementation and evaluation remain challenging due to the difficulty of adapting designs to different stakeholders and use case settings. To date, I conducted three interview studies examining the design of explanations for AI novices and one interview study on the intersection of explanation and contestation. All projects focused on end-users and people who might be affected by the decisions of ADM systems, including the system's users and 'decision subjects' [25]. By focusing on these stakeholder groups, I aim to gain insight into how explanation and contestation methods can award decision subjects 'democratic control' [8] and alleviate information asymmetry [5], power imbalances [14], and the 'algorithmic imprint' [13] that AI systems are known to produce.

In the following, I will outline my research's motivation and describe two published studies (Section 2). The first study was conducted in the summer of 2022 and published at FAccT'23 [36] (Section 3). The second study was conducted in the summer of 2023 and published at the International Journal of Human-Computer Studies [38] (Section 4). I then outline two current projects, the dissertation status (Section 5) and the expected benefits and contributions (Section 6).

2 Introduction

The risks of using AI systems in high-stakes settings have been discussed at large [8, 32, 34]. To counteract the resulting risks, principles were compiled to ensure trustworthy [18] and responsible [7] use of AI systems. These principles were incorporated into legal texts such as the GDPR, DSA, and the AI Act to govern the development and deployment of AI systems in the EU through safeguarding fundamental rights and conducting risk-based assessments [26]. However, realizing these conceptual principles in practice proves challenging. Both explainability and contestability are realized differently for different stakeholders, since developers [21, 24, 29, 33] have different aims and prior knowledge than non-technical stakeholders [2, 19, 27]. Further, as current regulations do not specify their implementation, the operationalization of both principles will manifest through the actions of regulated actors and extrajudicial processes [30], including through public institutions, standardization bodies, and civil society organizations. This opens up a space of interpretation for how these principles should be realized and what it means to comply with them [1].

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3721096

In practice, explanations can help numerous functions, explaining either the system's inner workings (*descriptive*) or the norms and reasons governing its use, connecting to justifications [17] (*normative)*. For example, developers can use explanations to understand if an AI system accurately predicts the employability of job-seekers, while decision subjects can use them to understand how a decision was made and which options of contestation are available to them. Due to this heterogeneity in aims and information needs, explanations should be adaptive and personalized to the individual [11, 42], being not only a means of information but also of empowerment for decision subjects [4, 8]. Similar challenges arise for the design of contestation mechanisms, depending on whether a single person contests or a collective and whether they use judicial or non-judicial channels.

Investigating how explanation and contestation mechanisms can be made suitable for different audiences is complicated by various factors: First, AI lifecycles encompass large amounts of information [12] that need to be structured and presented in a sensible manner; second, evaluating if stakeholders *understand* a given explanation is difficult as it involves mental processes that XAI research does not always address [16, 31]; third, effects of explanations are influenced by perceptions of the deploying institutions, necessitating considerations of context [39, 46]; and fourth, while explainability is posited as a necessary precondition for contestation, their intersection has only begun to be mapped out [47]. To gain insight into resolving these challenges, my thesis addresses the following research questions.

- RQ1: How do different explanation modalities impact participants' understanding? [36]
- RQ2: What are affected stakeholders' information needs when deciding on adopting an ADM system? [38]
- RQ3: How are explainability and contestability linked in their design and regulation?

Two published papers address RQ1 and RQ2, a third addresses their implementation in an explanation design, and a fourth addresses RQ3 in a just completed and submitted paper. My contributions to this point include a conceptualization and empirical examination of participant understanding, using the "six facets of understanding" framework by Wiggins and McTighe [45] (Section 3) and a collection of affected stakeholders' information needs about high-risk AI systems (Section 4). Current projects further provide insights into developing explanation designs for AI novices in groups and outlining the intersection of explainability and contestability (Section 5).

3 First study (RQ1): Participant understanding

The first study was published under the title "On the Impact of Explanations on Understanding of Algorithmic Decision-Making" at FAccT'23 [36]. Its goal was to better conceptualize what it means for people to *understand* explanations. Understanding enables the satisfaction of epistemic and substantial desiderata [20], such as assessing whether a model makes fair and transparent decisions. Explanations are a means to increase understanding, and thus, knowing how understanding is constituted is crucial to designing better explanations. While prior research in XAI often used understanding as the success metric of explanations [40, 41, 44], the term was rarely specified to describe how understanding develops cognitively or didactically. Research in the learning sciences argues that understanding is more than pure knowledge retrieval, but rather the ability to use acquired knowledge flexibly and in different contexts [6, 9, 45]. This study examined whether concepts from the learning sciences could be used to analyze participants' understanding of explanations (in textual, dialogue, and interactive modalities). We took the Austrian *AMS Algorithm* as an example for high-risk AI systems in public institutions [23]: a system developed to predict job-seekers' employability based on statistical comparison with past data [3].

3.1 Methods

We conducted in-person interview studies with 30 participants, sampled from the employment agency and public locations in Vienna. Participants were presented with one of three explanation modalities, taking the form of slide shows detailing the decision process of the *AMS Algorithm*, which differed in how additional information was conveyed (per textual or verbal comments or interactive controls). Interviews were recorded, transcribed, and subsequently thematically analyzed both inductively and deductively using the "six facets of understanding" framework [45]. Articulations of participants were coded if they showed a specific process of understanding: explaining, interpreting, applying, taking perspective, empathizing, and self-reflecting on knowledge.[1] Further, articulations were coded inductively regarding fairness perceptions, producing code categories which were then compared with the four dimensions of fairness perceptions defined by Starke et al. [43] (algorithmic predictors, human predictors, comparative effects, and consequences of ADM).

3.2 Findings

Our explanation setting allowed participants to share their understanding processes more thoroughly than a written test or assessment would have likely captured them [35]. Examining these processes, we observed that participants who felt more personally affected by the *AMS Algorithm* tended to use more "emotional" (interpret, empathize, self-reflect) than "rational" (explain, apply, take perspective) facets of understanding – highlighting the importance of emotional ways of understanding which are not usually considered in explanations. Further, participants who used multiple facets at once, e.g., explained what was happening, gave an example, and took a different perspective all in one statement, were able to articulate more detailed fairness assessments and often reflected on their knowledge. Regarding the modality, we found that participants welcomed the dialogue explanation to exchange and share their thought processes. In contrast, the textual modality led to understanding barriers (forgetting, inability to apply and reflect on knowledge [45]). We concluded that the individuality of understanding processes demanded a better adaptation of explanations to different stakeholders' information needs and ways of thinking, motivating the second study's design and research objectives.

[1] According to the six facets framework, the more of these facets someone can cover, the better their understanding of a topic.

4 Second study (RQ2): Affected stakeholders' information needs

The second study was published under the title "Information That Matters: Exploring Information Needs of People Affected by Algorithmic Decisions" at IJHCS. This study investigated how explanations could adapt to AI novices in content [28], form [10], and purpose [15] by examining which information AI novices who were also decision subjects deemed relevant. We used an approach similar to Liao et al. [21]'s collection of AI practitioners' information needs to collect AI novices' information needs and create the "XAI Novice Question Bank" (Figure 1). The collection is meant to be both a guide for future explanation design and aid in direct stakeholder interaction, increasing AI systems' intelligibility [22] by giving an informational overview. In addition to the *AMS Algorithm*, we used a health wristband as a second use case to examine whether a change in domain and system would also impact participant inquiry.

4.1 Methods

We conducted in-person interview studies with 24 participants (12 participants for each use case). Participants were required to have no previous knowledge about algorithmic systems (e.g., software developers would not qualify) and to be somehow affected by the system (i.e., present or past job-seekers, retirees, or family of people in care). Participants were presented with one use case and given 30 minutes to ask verbal questions about the system to the study examiner before deciding on its adoption. The process was framed as a simulated public vote to incentivize information acquisition. The 30-minute inquiry was split into two 15-minute phases: first, participants inquired freely, then they received the XAI Question Bank [21] as inspiration to examine its usefulness for AI novices. Self-reports of understanding, decision confidence, and perceptions of risks and benefits were elicited before and after the inquiry. Interviews were transcribed and thematically analyzed, using an inductive approach for creating question categories and a deductive approach for comparison with the XAI Question Bank [21]. Self-reports were visualized and compared with participant inquiries to gain insight into factors influencing information needs.

4.2 Findings

The XAI Novice Question Bank (Figure 1) depicts categories of participants' information needs and highlights the relevance of system context and system usage. Crucially, existent explanation approaches do not address questions in these categories, such as *What is the intention of deploying the system?* and *How will the system impact personal relations?* The technical focus of the XAI Question Bank [21] proved valuable for participants by offering questions that they did not consider themselves but still found relevant. Further, participants who perceived the system's risk to be high focused their inquiries on intention and consequence, whereas participants who perceived low risk rather focused on the system's operation. In conclusion, we provide a list of six key implications arguing that explanations for affected stakeholders must incorporate more and different information than for technical audiences and that further research is needed to ensure that explanations support affected stakeholders' agency.

5 Dissertation status and next steps

To date, I have published two first-author papers. A third paper reports on implementing the findings from RQ1 and RQ2 into a concrete explanation design for AI novices in groups and as individuals, featuring an interview and workshop study with 43 participants. It is currently being prepared for re-submission [37]. A fourth one examines RQ3, the intersection between explainability and contestability in design and regulation, by reporting findings from 14 interviews with AI regulation experts, having just been submitted for review.

In the next steps of my thesis, I aim to develop and test a digital explanation interface that can adapt to different stakeholders based on the insights of RQ1 and RQ2. I also plan to explore more closely how to design descriptive and normative explanations that can support contestation and how this can influence the deployment of responsible public AI systems. Further, as providing contestation can also lead to an excessive administrative load, finding participatory approaches that allow for democratic control by citizens without overwhelming the deploying institutions is a key challenge that warrants further research. Finding a way to implement explainability and contestability according to regulation and best practices in design while adjusting these measures to the capacities and limitations of public institutions will thus be one of my main areas of interest.

6 Benefits and contributions

The Remote Doctoral Consortium will be the first consortium that I attend, allowing me to critically discuss my research outlined above with other Ph.D. students and obtain new perspectives on its realization in terms of design and human factors. Additionally, receiving feedback on my overall research direction from senior researchers in the field of HCI would be highly valuable. As I am further considering applying for postdoctoral grants or research positions at EU institutions, I would appreciate guidance on how to proceed on this career path.

Further, networking and discussing AI research is both enjoyable and a valuable source of collaboration for me. My research stay at Télécom Paris and my latest project were made possible through connections with peers and colleagues whom I met at CHI'23 and FAccT'23. Together, we established a research project that brought together HCI, design, sociology, and law researchers from three different countries, creating a model for interdisciplinary collaboration. My future projects will likely be situated in similar research spaces and would greatly benefit from continued exchange and networking with colleagues from diverse fields.

In terms of contributions, I will be happy to share my experience in conducting empirical studies with qualitative methods, insights about working with different groups of people (e.g., job-seekers, retirees, and employees of public institutions), and recruiting strategies (e.g., street sampling, establishing contact to organizations, compensation). As I hold a Master's degree in professional writing and have experience in writing and editing journalistic, fictional, and academic texts, I will gladly offer writing advice or feedback to any other Ph.D. students.

XAI Novice Question Bank

System context

- ① ② What is the intention of deploying it?
- Ⓐ Ⓑ How does the deployment process work?
- Ⓓ Ⓔ What are the consequences after deployment?
- ② Ⓑ How is the system developed?
- ⑤ Who is the system's intended target group?
- ① ② Who is responsible for the system's deployment?
- ⑤ ⑥ What do other people think about the system?
- ① ⑥ How are moral values considered in the system?

System usage

- ④ Ⓒ How is the system operated?
- Ⓒ Ⓓ How will the system impact interpersonal relations?
- Ⓐ Ⓑ How is the system integrated into ex. structures?
- Ⓓ Ⓔ How can the system be misused?
- ③ ⑤ How would the system handle [this case]?
- ③ What are the costs?

Data

- ③ What kind of data was the system trained on?*
- ③ What is the source of the training data?*
- ③ Are the data correct / representative / safe?

System specifications

- ③ What features does the system consider and why?*
- ③ What is the scope of the system's capability?*
- ③ What kind of output does the system give?*
- ③ How does the system learn?*
- ③ What is the system's overall logic?*
- ③ What kind of algorithm was used?*
- ③ What kind of mistakes is the system likely to make?*
- ③ What are the limitations of the system?*
- ③ How reliable are the predictions?*
- ③ What does [a machine learning concept] mean?

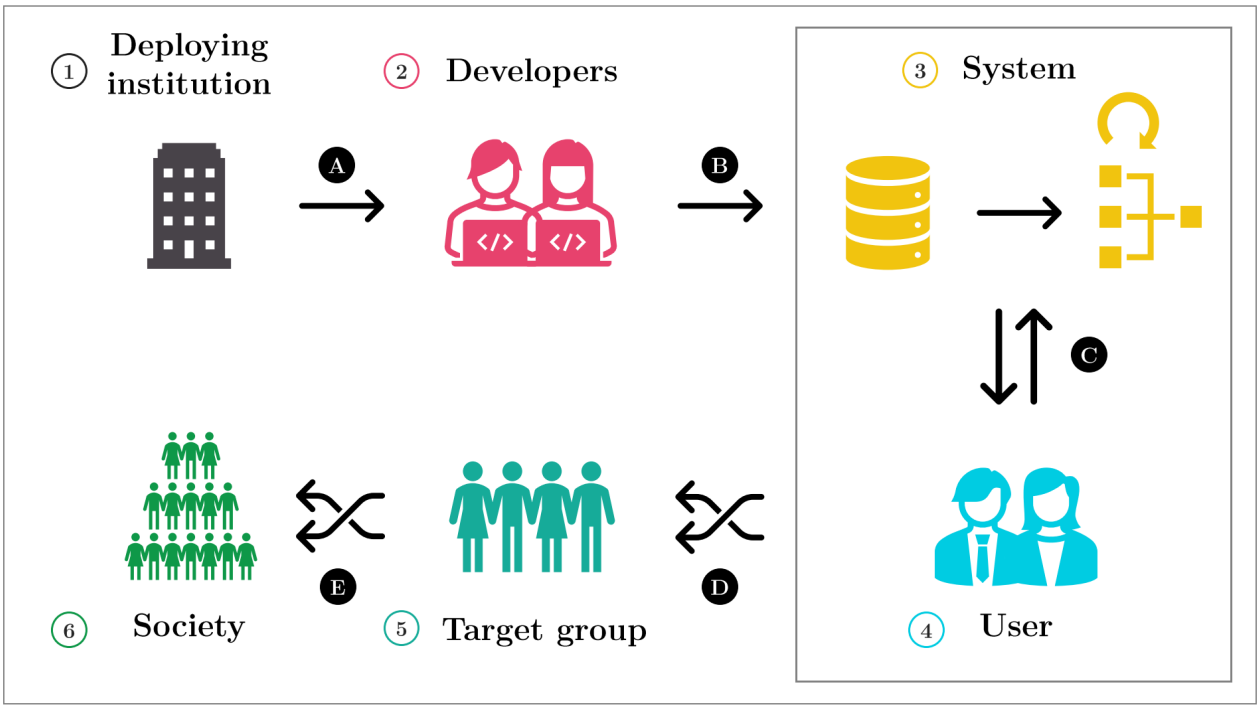

Figure 1: The XAI Novice Question Bank and system-inquiry diagram. Depicted are four categories of questions that subsume inquiries by affected stakeholders about two ADM systems. An asterisk (*) indicates that the question is already present in the XAI Question Bank [21]. Numbers and letters refer to stakeholders and procedures in the system deployment process.

Acknowledgments

This work has been funded by the Vienna Science and Technology Fund (WWTF) [10.47379/ICT20058] as well as [10.47379/ICT20065].

References

[1] Kars Alfrink, Ianus Keller, Mireia Yurrita Semperena, Denis Bulygin, Gerd Kortuem, and Neelke Doorn. 2024. Envisioning Contestability Loops: Evaluating the Agonistic Arena as a Generative Metaphor for Public AI. *She Ji: The Journal of Design, Economics, and Innovation* 10, 1 (2024), 53–93.

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (Nov. 2023), 101805. doi:10.1016/j.inffus.2023.101805

[3] Doris Allhutter, Astrid Mager, Florian Cech, Fabian Fischer, and Gabriel Grill. 2020. *Der AMS-Algorithmus: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS).* Technical Report. Österreichische Akademie der Wissenschaften. epub.oeaw.ac.at.

[4] Marco Almada. 2019. Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. ACM, Montreal QC Canada, 2–11. doi:10.1145/3322640.3326699

[5] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20, 3 (March 2018), 973–989. doi:10.1177/1461444816676645

[6] Lorin W. Anderson and David R. Krathwohl (Eds.). 2001. *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives* (complete ed ed.). Longman, New York.

[7] Ricardo Baeza-Yate and Jeanna Matthews. 2022. Statement on Principles for Responsible Algorithmic Systems. https://www.acm.org/binaries/content/assets/public-policy/final-joint-ai-statement-update.pdf Last accessed on 13th May 2024.

[8] Daniel James Bogiatzis-Gibbons. 2024. Beyond Individual Accountability: (Re-)Asserting Democratic Control of AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 74–84. doi:10.1145/3630106.3658541

[9] John D. Bransford and Marcia K. Johnson. 1972. Contextual Prerequisites for Understanding: Some Investigations of Comprehension and Recall. *Journal of Verbal Learning and Verbal Behaviour 11* (1972).

[10] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. doi:10.1145/3290605.3300789

[11] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence* 298 (Sept. 2021), 103503. doi:10.1016/j.artint.2021.103503

[12] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference 2021*. ACM, Virtual Event USA, 1591–1602. doi:10.1145/3461778.3462131

[13] Upol Ehsan, Ranjit Singh, Jacob Metcalf, and Mark Riedl. 2022. The Algorithmic Imprint. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1305–1317. doi:10.1145/3531146.3533186

[14] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, Inc.

[15] Timo Freiesleben and Gunnar König. 2023. Dear XAI Community, We Need to Talk!. In *Explainable Artificial Intelligence*, Luca Longo (Ed.). Springer Nature Switzerland, Cham, 48–65. doi:10.1007/978-3-031-44064-9_3

[16] Stephen R. Grimm. 2019. Varieties of Understanding. In *Varieties of Understanding*. Oxford University Press, 1–14. doi:10.1093/oso/9780190860974.003.0001

[17] Clément Henin and Daniel Le Métayer. 2022. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY* 37, 4 (Dec. 2022), 1397–1410. doi:10.1007/s00146-021-01251-8

[18] High-Level Expert Group on Artificial Intelligence. 2019. Ethics guidelines for trustworthy AI. (April 2019). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-{AI}

[19] Robert R. Hoffman, Shane T. Mueller, Gary Klein, Mohammadreza Jalaeian, and Connor Tate. 2023. Explainable AI: roles and stakeholders, desirements and challenges. *Frontiers in Computer Science* 5 (Aug. 2023), 1117848. doi:10.3389/fcomp.2023.1117848

[20] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (July 2021), 103473. doi:10.1016/j.artint.2021.103473

[21] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376590

[22] Brian Y. Lim and Anind K. Dey. 2009. Assessing Demand for Intelligibility in Context-Aware Applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing* (Orlando, Florida, USA) *(UbiComp '09)*. Association for Computing Machinery, New York, NY, USA, 195–204. doi:10.1145/1620545.1620576

[23] Paola Lopez. 2019. Reinforcing Intersectional Inequality via the AMS Algorithm in Austria. In *Proceedings of the 18th Annual STS Conference*. Graz, 289–309. doi:10.3217/978-3-85125-668-0-16

[24] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[25] Henrietta Lyons, Tim Miller, and Eduardo Velloso. 2023. Algorithmic Decisions, Desire for Control, and the Preference for Human Review over Algorithmic Review. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 764–774. doi:10.1145/3593013.3594041

[26] Winston Maxwell and Bruno Dumas. 2023. Meaningful XAI based on user-centric design methodology: Combining legal and human-computer interaction (HCI) approaches to achieve meaningful algorithmic explainability. *SSRN Electronic Journal* (2023). doi:10.2139/ssrn.4520754

[27] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. doi:10.1016/j.artint.2018.07.007

[28] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-Driven Decision Support Using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 333–342. doi:10.1145/3593013.3594001

[29] Christoph Molnar. 2022. *Interpretable Machine Learning* (2 ed.). https://christophm.github.io/interpretable-ml-book

[30] Nadia Nahar, Jenny Rowlett, Matthew Bray, Zahra Abba Omar, Xenophon Papademetris, Alka Menon, and Christian Kästner. 2024. Regulating Explainability in Machine Learning Applications – Observations from a Policy Design Experiment. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2101–2112. doi:10.1145/3630106.3659028

[31] Andrés Páez. 2019. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines* 29, 3 (01 Sep 2019), 441–459. doi:10.1007/s11023-019-09502-w

[32] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 959–972. doi:10.1145/3531146.3533158

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. doi:10.1145/2939672.2939778

[34] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (01 May 2019), 206–215. doi:10.1038/s42256-019-0048-x

[35] Brian K. Sato, Cynthia F. C. Hill, and Stanley M. Lo. 2019. Testing the test: Are exams measuring understanding? *Biochemistry and Molecular Biology Education* 47, 3 (May 2019), 296–302. doi:10.1002/bmb.21231

[36] Timothée Schmude, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. 2023. On the Impact of Explanations on Understanding of Algorithmic Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 959–970. doi:10.1145/3593013.3594054

[37] Timothée Schmude, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. 2024. Deliberative XAI: How Explanations Impact Understanding and Decision-Making of AI Novices in Collective and Individual Settings. arXiv:2411.11449 [cs.HC] https://arxiv.org/abs/2411.11449

[38] Timothée Schmude, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. 2024. Information That Matters: Exploring Information Needs of People Affected by Algorithmic Decisions. arXiv:2401.13324 [cs.HC]

[39] Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, and Bettina Berendt. 2022. Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, Seoul Republic of Korea, 2138–2148. doi:10.1145/3531146.3534631

[40] Ruoxi Shang, K. J. Kevin Feng, and Chirag Shah. 2022. Why Am I Not Seeing It? Understanding Users' Needs for Counterfactual Explanations in Everyday Recommendations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1330–1340. doi:10.1145/3531146.3533189

[41] Hong Shen, Haojian Jin, Angel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–22. doi:10.1145/3415224

[42] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. Enhancing Fairness Perception – Towards Human-Centred AI and Personalized Explanations Understanding the Factors Influencing Laypeople's Fairness Perceptions of Algorithmic Decisions. *International Journal of Human–Computer Interaction* (July 2022), 1–28. doi:10.1080/10447318.2022.2095705

[43] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2021. Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. (2021).

[44] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 109–119. doi:10.1145/3397481.3450662

[45] Grant P. Wiggins and Jay McTighe. 2005. *Understanding by design* (expanded 2nd ed.). Association for Supervision and Curriculum Development, Alexandria, VA.

[46] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–14. doi:10.1145/3173574.3174230

[47] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–21. doi:10.1145/3544548.3581161