# Careful Explanations: A Feminist Perspective on XAI

Laura State[1,2], Miriam Fahimi[3]

[1]Department of Computer Science, University of Pisa, Largo B. Pontecorvo, 3, 56127 Pisa, Italy
[2]Scuola Normale Superiore, Piazza dei Cavalieri, 7, 56126 Pisa, Italy
[3]Digital Age Research Center (D[i]ARC), University of Klagenfurt, Universitätsstraße 65-67, 9020 Klagenfurt/Celovec, Austria

Abstract

Explainable artificial intelligence (XAI) is a rapidly growing research field that has received a lot of attention during the last few years. An important goal of the field is to use its methods to detect (social) bias and discrimination. Despite these positive intentions, aspects of XAI can be in conflict with feminist approaches and values. Therefore, our conceptual contribution brings forward both a *careful* assessment of current XAI methods, as well as visions for *carefully* doing XAI from a feminist perspective. We conclude with a discussion on the possibilities for *caring* XAI, and the challenges that might lie along the way.

**Keywords**

explainable AI, feminism, care work

The question "If I had been a man (or a woman), would I have received the same treatment?" is common when people reason about their own experience when being discriminated against, to gauge whether the discrimination is originating from their own gender, or not. Interestingly, this exact question is now re-appearing in the field of *explainable artificial intelligence* (XAI), a field that is constructing (technical) explanations for otherwise intransparent automated decision-making (ADM) systems. Next to providing these methods, one of the main goals of this new field is to uncover (social) bias that is emerging in and through ADM [1, 2]. One way to achieve this is by answering the question presented in our introductory example, a so-called "what if" question.

However, only because XAI methods allow answering questions about the interdependence between discrimination and ADM, it does not necessarily mean that these methods are in agreement with feminist approaches and values. For instance, XAI can also be perceived as producing specific *knowledge* about a situation that XAI seeks to address and explain, and that is hard to be contested outside of computer science. This can be considered as an act of reproducing *power* relations [3, 4].

It is exactly that intersection of power and (X)AI that we want to understand with this work: we assess existing explainability methods from a *feminist* viewpoint, to understand whether they are the tools they are claimed to be, something else, or something beyond. Thus, we propose both to *carefully* reason about explanations, and to investigate how they can be considered as a *caring* practice.

We start by developing a layered definition of what feminism means in and for our work [5, 6, 7]. Some central parts in this definition will develop around 1) understanding power

*EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland*



relations and how (structural) discrimination emerges through them; 2) a special focus on the relationship between gender identities, and the necessity to think beyond the gender binary; 3) acknowledging care and care work in our society, its (often intentional) invisibility, and implications for the redestribution of this work.

Following these key aspects of feminism, and central work that is already existing at the intersection between feminism, feminist epistemology, care work, data and AI [8, 9, 10, 11, 12, 13, 14, 15, 16], we seek to critically assess and contribute towards a feminist perspective on XAI.

To do this, our work is asking the following two research questions (RQ): *i*) How can we assess *current* methods of XAI from a feminist viewpoint? and *ii*) How do we imagine *future* versions of XAI, and which *changes* to the current methodologies does this entail?

We can dissect RQ 1) using a couple of follow-up questions: Which aspects of existing XAI methods interfere with our working definition of feminism? How does existing XAI interact with power relations: do they perpetuate, or foster to challenge them? Which aspects of existing XAI methods can be considered feminist? Why? Regarding the RQ 2), we will ask questions such as: how could a feminist explanation look like? Which technical and social conditions are needed along the way? How do we imagine a feminist future towards *caring* XAI?

Answering the posed questions is an experimental and explorative endeavor, set out to forward a meaningful and critical contribution to the XAI community over techno-determinist solutionism [17]. As we will give answers to these questions, the issues that are raised might not be solved. Instead, they allow for further questioning and research.

This work *in progress* is of interdisciplinary character, bringing together perspectives from the social sciences, the computer sciences, and our shared interest in feminist theories, concepts and methodologies. While it is important to understand the global dimension of XAI and AI, we acknowledge our positionality as researchers that are based in Europe, and with European institutions.

**Outlook** A feminist perspective is an important and necessary addition to XAI and AI - it brings a view to the table that is historically neglected or marginalized and offers possibilities of "studying up" instead of "studying down" [18, 19]. This is especially important if we reason about (social and historical) discrimination, and about how this discrimination can be counter-acted.

If we do not integrate a feminist perspective, we will build a world of (X)AI that leaves some people outside of it.

## Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS - Artificial Intelligence without Bias" (*nobias-project.eu*). This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

## References


[1] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.

[2] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernández, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, S. Staab, Bias in data-driven artificial intelligence systems - an introductory survey, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 10 (2020).

[3] S. Harding, Whose Science? Whose Knowledge?: Thinking from Women's Lives, Cornell University Press, 2016. URL: https://www.degruyter.com/document/doi/10.7591/9781501712951/html. doi:10.7591/9781501712951, publication Title: Whose Science? Whose Knowledge?

[4] I. Lowrie, Algorithmic Rationality. Epistemology and Efficiency in the Data Sciences, Big Data & Society 4 (2017) 2053951717700925. URL: https://doi.org/10.1177/2053951717700925. doi:10.1177/2053951717700925, publisher: SAGE Publications Ltd.

[5] J. Wajcman, Feminism confronts Technology, Polity Press, Cambridge, UK, 1991.

[6] D. Haraway, Situated Knowledges. The Science Question in Feminism and the Privilege of Partial Perspective, Feminist Studies 14 (3) (1988) 575–599.

[7] P. Hill Collins, Black feminist thought: knowledge, consciousness, and the politics of empowerment, Routledge classics, 2nd ed. ed., Routledge, New York, 2009.

[8] T. Gebru, Oxford handbook on AI ethics book chapter on race and gender, CoRR abs/1908.06165 (2019).

[9] L. Hancox-Li, I. E. Kumar, Epistemic values in feature importance methods: Lessons from feminist epistemology, in: M. C. Elish, W. Isaac, R. S. Zemel (Eds.), FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021, ACM, 2021, pp. 817–826.

[10] C. Perez, Invisible Women: Data Bias in a World Designed for Men, Random House, 2019.

[11] C. D'Ignazio, L. Klein, Data Feminism, MIT Press, 2020.

[12] C. Bath, Searching for methodology: Feminist technology design in computer science, http://elib.suub.uni-bremen.de/ip/docs/00010427.pdf (2009).

[13] M. Ruckenstein, L. L. M. Turunen, Re-Humanizing the Platform. Content Moderators and the Logic of Care, New Media & Society 22 (2020) 1026–1042. URL: http://journals.sagepub.com/doi/10.1177/1461444819875990. doi:10.1177/1461444819875990.

[14] N. Seaver, CARE AND SCALE: Decorrelative Ethics in Algorithmic Recommendation, Cultural Anthropology 36 (2021) 509–537. URL: https://onlinelibrary.wiley.com/doi/abs/10.14506/ca36.3.11. doi:10.14506/ca36.3.11, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.14506/ca36.3.11.

[15] K. J. Rohlfing, P. Cimiano, I. Scharlau, T. Matzner, H. M. Buhl, H. Buschmeier, E. Esposito, A. Grimminger, B. Hammer, R. Häb-Umbach, I. Horwath, E. Hüllermeier, F. Kern, S. Kopp, K. Thommes, A.-C. N. Ngomo, C. Schulte, H. Wachsmuth, P. Wagner, B. Wrede, Explanation as a Social Practice. Toward a Conceptual Framework for the Social Design of AI Systems,

IEEE Transactions on Cognitive and Developmental Systems (2020) 1–1. doi:10.1109/TCDS.2020.3044366, conference Name: IEEE Transactions on Cognitive and Developmental Systems.

[16] A. Kasirzadeh, Reasons, Values, Stakeholders: A Philosophical Framework for Explainable Artificial Intelligence, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 14. URL: http://doi.org/10.1145/3442188.3445866. doi:10.1145/3442188.3445866.

[17] E. Morozov, To Save Everything, Click Here: The Folly of Technological Solutionism, PublicAffairs, New York, 2013.

[18] L. Nader, Up the Anthropologist: Perspectives Gained From Studying Up, Technical Report, 1972. URL: https://eric.ed.gov/?id=ED065375, eRIC Number: ED065375.

[19] C. Barabas, C. Doyle, J. Rubinovitz, K. Dinakar, Studying up: reorienting the study of algorithmic fairness around issues of power, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 167–176. URL: https://dl.acm.org/doi/10.1145/3351095.3372859. doi:10.1145/3351095.3372859.