# Translation/Dialogue Tutorial: Disentagling the Relationship between Explainable AI and Fairness

Luca Deck
luca.deck@uni-bayreuth.de
University of Bayreuth & Fraunhofer FIT
Bayreuth, Germany

Jakob Schoeffer
j.j.schoeffer@rug.nl
University of Groningen
Groningen, The Netherlands

Maria De-Arteaga
dearteaga@utexas.edu
University of Texas at Austin
Austin, USA

Niklas Kühl
kuehl@uni-bayreuth.de
University of Bayreuth & Fraunhofer FIT
Bayreuth, Germany

## Abstract

Proponents of explainable artificial intelligence (XAI) commonly assume an implicit link between explanations and fairness. A plethora of XAI approaches and methods have been claimed to "promote" or even "ensure" fairness. However, the exact relationship often remains unclear. In this tutorial, we present a critical view of common claims on the alleged fairness benefits of XAI, as well as its drawbacks, anchored on a systematic review of 175 recent articles on the topic. By organizing the scattered debate into meaningful sub-debates around seven archetypal claims on the alleged fairness benefits of XAI, we provide an entry point for future discussions on the suitability and limitations of XAI for fairness. To foster more productive research, design, and application of XAI methods for fairness purposes, we provide guidelines for researchers and practitioners to be specific about *what* kind of XAI method is used, *which* fairness desideratum it addresses, *how* exactly it promotes fairness, and *who* is the stakeholder that benefits from XAI.

**ACM Reference Format:**
Luca Deck, Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2025. Translation/Dialogue Tutorial: Disentagling the Relationship between Explainable AI and Fairness. In *ACM FAccT '25, June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Description & Impact Statement

Explainable AI (XAI) in its various forms is commonly conceived as a remedy to algorithmic unfairness [2, 3, 20]. However, the implicit link between XAI and fairness has been challenged due to a lack of specificity, normative reasoning, and evidence [4, 12, 19].

In this tutorial, we introduce seven archetypal claims on the alleged fairness benefits of XAI, distilled from a large-scale survey of the literature conducted by the tutorial organizers [12]. In Figure 1, these claims are organized along three high-level fairness dimensions: *(i) General Fairness* refers to claims where it is left unspecified what kind of fairness is pursued, *(ii) Formal Fairness* refers to mathematical and statistical operationalizations of fairness [5, 6], and *(iii) Perceived Fairness* refers to subjective and context-specific human attitudes that are often measured as psychological constructs [7, 29].

Based on this systemization, we disentangle the diverse roles of XAI and discuss how they may or may not serve *epistemic* and *substantial* goals. *Epistemic* refers to the capability of humans to *observe* fairness properties of a model (e.g., XAI providing insights into a model's reliance on sensitive features), whereas a substantial goal actively aims to *alter* fairness properties (e.g., mitigating formally unfair model characteristics). By distinguishing epistemic from substantial goals of XAI, we structure the discourse and make it accessible to audiences beyond computer science, where the majority of claims are originating from.

We proceed to identify three fundamental critiques of common claims on the fairness benefits of XAI: *First*, despite being highly optimistic, many claims on the relationship between XAI and fairness are vague and simplistic. This calls for more specificity about the relationship between concrete XAI mechanisms and fairness desiderata. *Second*, many fairness desiderata pursued with XAI methods are lacking normative reasoning. For example, some research treats "reliance on sensitive features" [2] as a form of unfairness without offering a normative rationale for why such reliance might be problematic. This notion also disregards cases where the use of sensitive features may be warranted or even necessary for certain purposes, seemingly ignoring the literature on the limitations of "fairness through unawareness." *Third*, even in cases of specifying and motivating a valid fairness desideratum, some claims are poorly aligned with the actual capabilities of XAI. For example, if the goal is to achieve formal distributive fairness, it is unclear how exactly XAI should promote this [24].

To foster more effective applications of XAI for fairness, this tutorial provides useful tools and terminology to XAI researchers and practitioners to articulate *what* kind of XAI method is being considered, *which* fairness desideratum it refers to, *how* exactly it promotes fairness, and *who* is the stakeholder that benefits from XAI. Additionally, we aim to spark discussions on practical opportunities and limitations of XAI for specific fairness desiderata, as well as requirements for novel XAI methods to pursue these desiderata in the future. Taken together, tutorial attendees can expect to learn about the state of XAI for fairness, how to make precise claims about the potential of XAI for fairness and acquire foundations


*ACM FAccT '25, Athens, Greece*

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

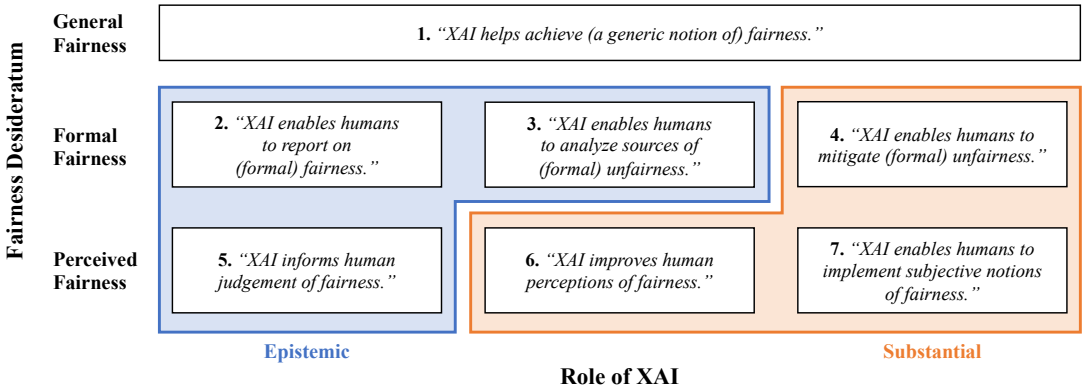


**Figure 1: Seven archetypal claims on the role of XAI for fairness desiderata.**

that can enable more principled development of XAI methods for concrete fairness applications.

## 2 Tutorial Team

**Luca Deck** is a research associate in the Business & Information Systems Engineering branch of the Fraunhofer FIT and a PhD student in Information Systems and Human-Centric AI at the University of Bayreuth in Germany. His research is focused on the limitations and potentials of XAI for fairness through a stakeholder-centered lens [e.g., 12, 13]. Together with interdisciplinary experts from the fields of ethics and law, he is also studying specific facets of fairness such as the relationship between algorithmic fairness and non-discrimination law [e.g., 11]. In his role at the Fraunhofer FIT, he is regularly hosting workshops and trainings on the effective and responsible use of AI systems in companies.

**Jakob Schoeffer** is an Assistant Professor at the University of Groningen, Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence. Previously, he was a Postdoctoral Research Fellow at the University of Texas at Austin, working closely with Maria De-Arteaga. He holds a PhD in Information Systems and Human-Computer Interaction from the Karlsruhe Institute of Technology in Germany. His research is focused on responsible and explainable AI [e.g., 26, 27], as well as on human-AI collaboration in critical decision-making [e.g., 24, 25], such as healthcare. He is especially interested in developing socio-technical interventions that empower human experts to complement AI systems, leading to more effective and equitable decision-making.

**Maria De-Arteaga** is an Assistant Professor at the Information, Risk, and Operations Management Department at the McCombs School of Business at the University of Texas at Austin. She is also a core faculty member in the interdepartmental Machine Learning Laboratory and a Good Systems researcher. She holds a joint PhD in Machine Learning and Public Policy from Carnegie Mellon University's Machine Learning Department and Heinz College. Her research is focused on algorithmic fairness and human-AI complementarity. As part of her work, she characterizes how societal biases encoded in historical data may be reproduced and amplified by ML models [e.g., 1, 10], and develops algorithms to mitigate these risks [e.g., 16, 22]. Moreover, effective human-AI collaboration is often complicated by other factors, such as the need for humans and algorithms to complement one another. In her research, she studies humans' ability to make productive use of algorithms [e.g., 8, 12], and to develop human-centered ML that can improve expert decision-making [e.g., 9, 15].

**Niklas Kühl** is a Full Professor of Information Systems and Human-Centered Artificial Intelligence at the University of Bayreuth. He is also a Group Leader at the Fraunhofer FIT and a Senior Expert in Artificial Intelligence at IBM. He holds a PhD in Information Systems from the Karlsruhe Institute of Technology, where he also led the Applied AI in Services Lab. His research focuses on human-AI collaboration, reliable AI, and the development of robust and scalable AI systems. In his work, he explores how AI can support expert decision-making [e.g., 18, 28] while ensuring fairness [e.g., 25, 26], interpretability [e.g., 21, 30], and appropriate reliance [e.g., 14, 23]. As part of his research, he develops methodologies to structure fairness in AI decisions and to design human-centered AI that aligns with real-world needs [e.g., 13, 17]. His work is regularly published in leading journals and conferences, and he collaborates with international institutions such as CMU, MIT, and the University of Texas at Austin.

## 3 Timeline

The tutorial will last 60 minutes, with 45 minutes dedicated to a lecture-style format and 15 minutes for interactions and discussions with the audience. The first part consists of four chapters:

(1) Introduction
(2) Claims on fairness benefits of XAI
(3) Current shortcomings
(4) Guidelines for future applications of XAI for fairness

The discussion will make room for questions from the audience and collect input guided by these questions:

(1) (How) have you employed XAI for fairness?
(2) How can we be more specific about *what* kind of XAI method is used, *which* fairness desideratum it refers to, *how* exactly it enables fairness, and *who* is the stakeholder that benefits from XAI?
(3) How could XAI be designed to effectively adress specific fairness desiderata?

## References


[1] Nil-Jana Akpinar, Maria De-Arteaga, and Alexandra Chouldechova. 2021. The effect of differential victim crime reporting on predictive policing systems. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 838–849.

[2] Guilherme Alves, Vaishnavi Bhargava, Fabien Bernier, Miguel Couceiro, and Amedeo Napoli. 2020. FixOut: An ensemble approach to fairer models. https://hal.archives-ouvertes.fr/hal-03033181/

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[4] Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2022. Challenges in applying explainability methods to improve the fairness of NLP models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*.

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT Press.

[6] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12, 1 (2022), 4209.

[7] Jason A. Colquitt and Jessica B. Rodell. 2015. Measuring Justice and Fairness. In *The Oxford handbook of justice in the workplace*, Russell Cropanzano, Russell S. Cropanzano, and Maureen L. Ambrose (Eds.). Oxford University Press, Oxford.

[8] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[9] Maria De-Arteaga, Vincent Jeanselme, Artur Dubrawski, and Alexandra Chouldechova. 2025. Leveraging expert consistency to improve algorithmic decision support. *Management Science (forthcoming)* (2025).

[10] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*. 120–128.

[11] Luca Deck, Jan-Laurin Müller, Conradin Braun, Domenique Zipperling, and Niklas Kühl. 2024. Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness. In *European Workshop on Algorithmic Fairness (EWAF'24)*. http://arxiv.org/pdf/2403.20089

[12] Luca Deck, Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. A critical survey on fairness benefits of explainable AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1579–1595.

[13] Luca Deck, Astrid Schomäcker, Timo Speith, Jakob Schöffer, Lena Kästner, and Niklas Kühl. 2024. Mapping the Potential of Explainable Artificial Intelligence (XAI) for Fairness Along the AI Lifecycle. In *European Workshop on Algorithmic Fairness (EWAF'24)*. http://arxiv.org/pdf/2404.18736

[14] Sven Eckhardt, Niklas Kühl, Mateusz Dolata, and Gerhard Schwabe. 2024. A Survey of AI Reliance. *arXiv preprint arXiv:2408.03948* (2024).

[15] Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. 2021. Human-AI Collaboration with Bandit Feedback. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 1722–1728. doi:10.24963/ijcai.2021/237 Main Track.

[16] Soumyajit Gupta, Sooyong Lee, Maria De-Arteaga, and Matthew Lease. 2023. Same same, but different: Conditional multi-task learning for demographic-specific toxicity detection. In *Proceedings of the ACM Web Conference 2023*. 3689–3700.

[17] Niklas Kühl. 2024. Human-centric Artificial Intelligence: The road ahead. *Transfer: Zeitschrift für Kommunikation & Markenmanagement* 70, 4 (2024).

[18] Julius Peter Landwehr, Niklas Kühl, Jannis Walk, and Mario Gnädig. 2022. Design knowledge for deep-learning-enabled image-based decision support systems: evidence from power line maintenance decision-making. *Business & Information Systems Engineering* 64, 6 (2022), 707–728.

[19] 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 (2021), 103473.

[20] David Leslie. 2019. Understanding artificial intelligence ethics and safety. *The Alan Turing Institute* (2019). https://www.turing.ac.uk/news/publications/understanding-artificial-intelligence-ethics-and-safety

[21] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The impact of imperfect XAI on human-AI decision-making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–39.

[22] Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4187–4195. doi:10.18653/v1/N19-1424

[23] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[24] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2024. Explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[25] Jakob Schoeffer, Johannes Jakubik, Michael Vössing, Niklas Kühl, and Gerhard Satzger. 2025. AI reliance and decision quality: Fundamentals, interdependence, and the effects of interventions. *Journal of Artificial Intelligence Research* 82 (2025), 471–501.

[26] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1616–1628.

[27] Jakob Schoeffer, Alexander Ritchie, Keziah Naggita, Faidra Monachou, Jessica Finocchiaro, and Marc Juarez. 2023. Online platforms and the fair exposure problem under homophily. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11899–11908.

[28] Philipp Spitzer, Daniel Hendriks, Jan Rudolph, Sarah Schläger, Jens Ricke, Niklas Kühl, Boj Hoppe, and Stefan Feuerriegel. 2025. The effect of medical explanations from large language models on diagnostic decisions in radiology. *medRxiv* (2025), 2025–03.

[29] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society* 9, 2 (2022), 20539517221115189.

[30] Michael Vössing, Niklas Kühl, Matteo Lind, and Gerhard Satzger. 2022. Designing transparency for effective human-AI collaboration. *Information Systems Frontiers* 24, 3 (2022), 877–895.