On the Relationship Between Explanations, Fairness Perceptions, and Decisions

Jakob Schoeffer
Karlsruhe Institute of Technology
Karlsruhe, Germany
jakob.schoeffer@kit.edu

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

Niklas Kuehl*
Karlsruhe Institute of Technology
Karlsruhe, Germany
niklas.kuehl@kit.edu

* denotes equal contribution

Abstract

It is known that recommendations of AI-based systems can be incorrect or unfair. Hence, it is often proposed that a human be the final decision-maker. Prior work has argued that explanations are an essential pathway to help human decision-makers enhance decision quality and mitigate bias, i.e., facilitate human-AI complementarity. For these benefits to materialize, explanations should enable humans to *appropriately rely* on AI recommendations and override the algorithmic recommendation when necessary to increase distributive fairness of decisions. The literature, however, does not provide conclusive empirical evidence as to whether explanations enable such complementarity in practice. In this work, we (a) provide a conceptual framework to articulate the relationships between explanations, fairness perceptions, reliance, and distributive fairness, (b) apply it to understand (seemingly) contradictory research findings at the intersection of explanations and fairness, and (c) derive cohesive implications for the formulation of research questions and the design of experiments.

Author Keywords

Algorithmic decision-making; human-AI complementarity; explanations; fairness; perceptions; reliance

ACM CHI 2022 Workshop on Human-Centered Explainable AI (HCXAI), May 12–13, 2022, New Orleans, LA, USA

Contributions

1. Proposing a framework: We propose a framework to make explicit the relationships between explanations, fairness perceptions, reliance on AI advice, and distributive fairness.

2. Enabling dialogue between prior works: Our framework enables us to articulate a dialogue between (seemingly contradictory) prior works and identify research gaps.

3. Deriving research questions: Given explanations can mislead perceptions, we ask whether this can in turn mislead reliance on AI advice in detriment of distributive fairness.

Introduction

Among many other desiderata [26], it is often assumed in the XAI literature that explanations should enable humans to assess the fairness of AI recommendations, and to ultimately make better and fairer decisions [2, 8, 10, 11, 14–16, 23, 35, 36].

Prior findings are inconclusive However, as of today, there is no conclusive empirical evidence showing that explanations facilitate human-AI complementarity. Prior work has found that explanations can influence people's fairness perceptions towards AI models and their predictions in positive or negative ways (e.g., [4, 10, 25, 34]). Other findings suggest that explanations may (e.g., [7, 24]) or may not (e.g., [1,3,18,32]) lead to enhanced human-AI performance.

Explanations can be misleading Deriving conclusions from existing findings is further complicated by evidence that explanations can mislead people's beliefs [6, 25, 33], even in cases where there is no intention to manipulate [12]. Lakkaraju and Bastani [25] construct high-fidelity explanations to deceive people into trusting models that make decisions based on sensitive information (e.g., race or gender) by leveraging correlations between legitimate and sensitive features. This way, people can be nudged into perceiving a model as procedurally fair, when in reality it is *not* fair. However, to the best of our knowledge there is a lack of research studying how such miscalibrated perceptions influence people's reliance on AI recommendations as well as potential effects on distributive fairness.

A holistic view is needed We aim to make sense of scattered findings at the intersection of explanations and fair decision-making. Specifically, we propose a conceptual framework (see Fig. 1) to better understand the mechanisms through which explanations may affect decisions. To that end, we make explicit the relationships between explanations, procedural fairness perceptions, reliance on AI, and distributive fairness. We show that the proposed framework enables us to articulate a dialogue between prior works and identify gaps that require further research. In particular, we show that prior literature has focused on different individual parts of a bigger system but that a more comprehensive lens—such as the one enabled by our proposed framework—is required to understand the effects of explanations on AI reliance and distributive fairness.

Pathway from explanations to distributive fairness
Based on the application of our framework, we identify a pathway from explanations to distributive fairness, mediated by perceptions of procedural fairness and their effect on AI reliance. We show that previous research has only studied portions of this path, and argue for its importance in the appropriate characterization of explanations' role in fair decision-making. In particular, we conjecture that miscalibrated fairness perceptions (e.g., due to misleading explanations) may influence reliance on AI in undesirable ways, by making people adopt incorrect or override correct AI recommendations. This lends support to the hypothesis that there is a disconnect between what explanations provide and the fairness benefits they claim. As there is little knowledge on this interplay, we are interested in answering the following research question:

RQ: Given that explanations can mislead perceptions of fairness, (how) does this, in turn, mislead adoption/overriding behavior of AI recommendations in detriment of distributive fairness?

In this workshop paper, we introduce our conceptual framework and apply it to better understand previous findings.

Definitions

Procedural Fairness Perceptions: Whether people think that the underlying AI's decision-making procedures are fair. (e.g., [39])

Distributive Fairness: The magnitude of disparities in error rate distributions across demographic groups. (e.g., [5])

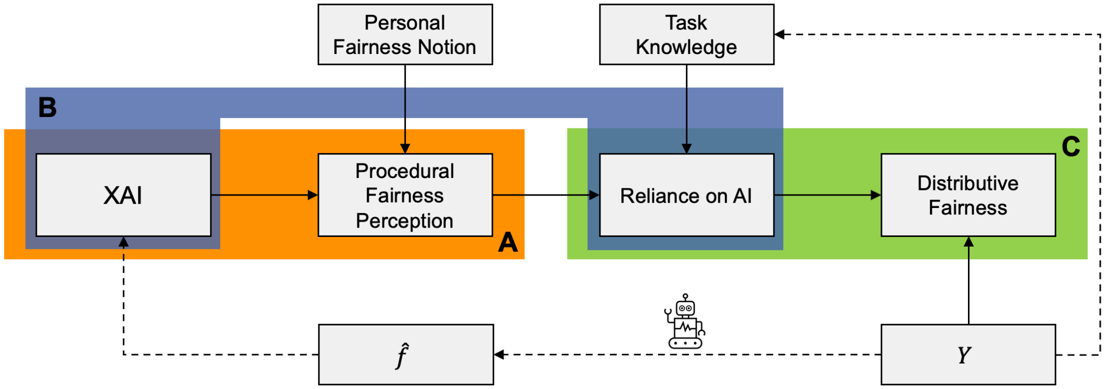

Figure 1: Conceptual framework on the interplay of explanations (XAI), procedural fairness perceptions, reliance on AI, and distributive fairness. Dashed lines indicate "brittle" relationships, Y is the gold standard, and $\hat{f}$ is the functional representation of the AI model.

Addressing the above research question through a human subject study is part of our in-progress work.

Conceptual Framework

We propose a conceptual framework (see Fig. 1) to make explicit the relationships between explanations *(XAI)*, procedural fairness perceptions, reliance on AI recommendations, and distributive fairness. While our framework may not capture *every* factor possibly at play, it aims to capture the primary factors considered in the literature.

Hypothesis

Unwarranted high (low) perceptions will lead to unwarranted adherence (overrides) of AI recommendations, which make explanations an unreliable mechanism towards improving distributive fairness.

Known relationships From prior work (e.g., [4, 10, 19–21, 25]), we know that explanations affect people's procedural fairness perceptions (i.e., whether people think that the underlying AI's decision-making procedures are fair). Especially the revelation of sensitive features (e.g., gender or race) being used in the process appears to have significant effects [21, 25, 38]. We further know that there are several human-specific predictors of fairness perceptions [10, 37], which we subsume under *Personal Fairness Notion*. This may include, e.g., individuals' stance towards affirmative action [22], but may also vary across demographics [20, 31]. Finally, by *distributive fairness* we mean the magnitude of disparities in error rate distributions across demographic groups (e.g., males and females) [5].

"Brittle" relationships Our framework also includes several "brittle" relationships, indicated by dashed lines. First, the relationship between Y and the functional representation of the AI model $(\hat{f})$ depends on the architecture, performance, and underlying data of the employed AI model (indicated by a robot icon in Fig. 1). Second, the relationship between $\hat{f}$ and *XAI* is ambiguous because an AI model (or its predictions) can be explained in a multitude of ways, or even independent of $\hat{f}$ [13]—even when explanations are honest [2, 25]. Third, task knowledge may or may not represent Y (i.e., knowledge can be "good" or "bad").

Hypothesized relationships The relationship between fairness perceptions and reliance on AI (i.e., whether to adopt or override AI recommendations) is seldom touched upon in the XAI literature. We assume a relationship to exist. In particular, we conjecture that higher procedural fairness perceptions may be associated with increased adoption of AI recommendations—even if unwarranted. This would be problematic insofar as perceptions can be manipulated through explanations (as discussed earlier): *inappropriate reliance* [27] might be a consequence.

Takeaways

For researchers: When proposing novel XAI techniques, researchers should show how their methods do *not* lead to unwarranted reliance on AI recommendations.

For auditors: Auditors should take a holistic perspective and assess the effects of explanations along *all* dimensions of our framework (i.e., perceptions, reliance, and distributive fairness).

For policy makers: Knowing that explanations can be misleading, demanding a "right to explanations" is *not* a sufficient condition for ensuring fair, accountable, and transparent use of AI in decision-making.

Applying Our Framework

We apply our framework to previous findings, inferring that they can be divided into three groups, based on the subset of relationships that were examined (A, B, or C in Fig. 1).

A: XAI and fairness perceptions A first set of works have studied the relationship between explanations and people's perceptions. Lakkaraju and Bastani [25], e.g., construct explanations based on sensitive vs. relevant features and show that they can be used to mislead people into trusting untrustworthy models. Similarly, Pruthi et al. [33] manipulate attention-based explanations such that people can be deceived into thinking that a model does not rely on sensitive information (e.g., gender) when in fact it does. Binns et al. [4] compared fairness perceptions across different explanation styles and scenarios—with inconclusive findings. Dodge et al. [10] find that people perceive global and local explanations differently, but also conclude that the effect of explanations depends on "the kinds of fairness issues and user profiles." Similarly, Shulner-Tal et al. [34] found that some explanations "are more beneficial than others," but perceptions mainly depend on "the outcome of the system."

B: XAI and reliance on AI Another set of works have examined how explanations may impact people's reliance on AI. Poursabzi-Sangdeh et al. [32] analyzed human-AI decision-making for the case of house price estimation and found that performance did *not* increase in the presence of explanations—likely due to information overload. Green and Chen [18] confirmed that explanations did not improve human performance, and Liu et al. [28] found that interactive explanations did not remedy this. A similar study by Alufaisan et al. [1] found no conclusive evidence of explanations' influence on decision accuracy either and showed that explanations did not enable humans to detect when the AI was correct or incorrect. Bansal et al. [3] did observe complementarity improvements in the presence of AI augmentation, but explanations only led to over-reliance on AI advice. On the other hand, Lai and Tan [24] found that providing explanations and AI predictions can enhance human decision-making for the task of deception detection.

C: Reliance on AI and distributive fairness Several prior works have addressed the interplay of humans' reliance on AI recommendations and fairness of outcomes. Peng et al. [30] identify different types of biases in AI-based hiring decisions and find that balancing gender representations when showing potential hires to human decision-makers can correct biases in instances where humans do not exhibit persistent preferences. Peng et al. [29] also investigated how an AI model's predictive performance and biases may transfer to humans; one of the core findings being that different model architectures have different effects on team performance and potential mitigation of biases. In the realm of child maltreatment screening, De-Arteaga et al. [9] found that call workers changed behavior in the presence of an AI recommendation, and that they were less likely to adopt incorrect AI advice. Green and Chen [17], however, in a different risk assessment case, found that humans under-performed the AI even when presented with its advice, were unable to evaluate both their own and the AI's performance, and biases against Black people were amplified through the use of AI recommendation.

In our in-progress work, we jointly consider relationships A, B, and C, in order to empirically examine explanations' effects on AI reliance and distributive fairness.

REFERENCES

[1] Yasmeen Alufaisan, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantarcioglu. 2020. Does explainable artificial intelligence improve human decision-making? *arXiv preprint arXiv:2006.11194* (2020).

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

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[4] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[5] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

[6] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark patterns of explainability, transparency, and user control for intelligent systems. In *IUI Workshops*, Vol. 2327.

[7] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? A case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248* (2020).

[8] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *arXiv preprint arXiv:2006.11371* (2020).

[9] 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.

[10] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 275–285.

[11] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2020. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems* 36, 4 (2020), 25–34.

[12] Upol Ehsan and Mark O Riedl. 2021. Explainability pitfalls: Beyond dark patterns in explainable AI. *arXiv preprint arXiv:2109.12480* (2021).

[13] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[14] Juliana Jansen Ferreira and Mateus de Souza Monteiro. 2020. Evidence-based explanation to promote fairness in AI systems. *arXiv preprint arXiv:2003.01525* (2020).

[15] Julie Gerlings, Arisa Shollo, and Ioanna Constantiou. 2020. Reviewing the need for explainable artificial intelligence (XAI). *arXiv preprint arXiv:2012.01007* (2020).

[16] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 80–89.

[17] Ben Green and Yiling Chen. 2019a. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 90–99.

[18] Ben Green and Yiling Chen. 2019b. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[19] Nina Grgić-Hlača, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*. 903–912.

[20] Nina Grgić-Hlača, Adrian Weller, and Elissa M Redmiles. 2020. Dimensions of diversity in human perceptions of algorithmic fairness. *arXiv preprint arXiv:2005.00808* (2020).

[21] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[22] Harry Holzer and David Neumark. 2000. Assessing affirmative action. *Journal of Economic Literature* 38, 3 (2000), 483–568.

[23] Aditya Kuppa and Nhien-An Le-Khac. 2020. Black box attacks on explainable artificial intelligence (XAI) methods in cyber security. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[24] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.

[25] Himabindu Lakkaraju and Osbert Bastani. 2020. “How do I fool you?” Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85.

[26] 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.

[27] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.

[28] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.

[29] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. 2022. Investigations of performance and bias in human-AI teamwork in hiring. *arXiv preprint arXiv:2202.11812* (2022).

[30] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What you see is what you get? The impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 125–134.

[31] Emma Pierson. 2017. Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124* (2017).

[32] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[33] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2019. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913* (2019).

[34] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. Fairness, explainability and in-between: Understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology* 24, 1 (2022), 1–13.

[35] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 180–186.

[36] Kacper Sokol and Peter A Flach. 2019. Counterfactual explanations of machine learning predictions: Opportunities and challenges for AI safety. In *SafeAI@AAAI*.

[37] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2021. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *arXiv preprint arXiv:2103.12016* (2021).

[38] Niels Van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[39] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.