Towards Explainability for AI Fairness

Jianlong Zhou[1,2], Fang Chen[1], Andreas Holzinger[2,3][0000−0002−6786−5194]

[1] Human-Centered AI Lab, University of Technology Sydney, Australia
{Jianlong.Zhou, Fang.Chen}@uts.edu.au
[2] Human-Centered AI Lab, Medical University Graz, Austria
[3] xAI Lab, Alberta Machine Intelligence Institute, Edmonton, Canada
Andreas.Holzinger@medunigraz.at

Abstract. AI explainability is becoming indispensable to allow users to gain insights into the AI system's decision-making process. Meanwhile, fairness is another rising concern that algorithmic predictions may be misaligned to the designer's intent or social expectations such as discrimination to specific groups. In this work, we provide a state-of-the-art overview on the relations between explanation and AI fairness and especially the roles of explanation on human's fairness judgement. The investigations demonstrate that fair decision making requires extensive contextual understanding, and AI explanations help identify potential variables that are driving the unfair outcomes. It is found that different types of AI explanations affect human's fairness judgements differently. Some properties of features and social science theories need to be considered in making senses of fairness with explanations. Different challenges are identified to make responsible AI for trustworthy decision making from the perspective of explainability and fairness.

Keywords: Fairness, explainable AI, explainability, machine learning

1 Introduction

Artificial Intelligence (AI) including Machine Learning (ML) algorithms are increasingly shaping people's daily lives by making decisions with ethical and legal impacts in various domains such as banking, insurance, medical care, criminal justice, predictive policing, and hiring [43, 44]. While AI-informed decision making can lead to faster and better decision outcomes, however, AI algorithms such as deep learning often use complex learning approaches and even their designers are often unable to understand why AI arrived at a specific decision. Therefore, AI remains a black box that makes it hard for users to understand why a decision is made or how the data is processed for the decision making [8, 45, 44]. Because of the black box nature of AI models, the deployment of AI algorithms especially in high stake domains usually requires testing and verification for reasonability by domain experts not only for safety but also for legal reasons [35]. Users also want to understand reasons behind specific AI-informed decisions. For example, high-stake domains require explanations of AI before any critical decisions, computer scientists use explanations to refine and further improve performance of AI algorithms, and AI explanations can also improve the user experience of a product or service by helping end-users trust that the AI is making good decisions [7]. As a result, the issue of AI explanation has experienced a significant surge in interest from the international research community to various application domains, ranging from agriculture to human health and is becoming indispensable in addressing ethical concerns and fostering trust and confidence in AI systems [42, 43, 20].

Furthermore, AI algorithms are often trained on a large amount of historical data, which may not only replicate, but also amplify existing biases or discrimination in historical data. Therefore, due to such biased input data or faulty algorithms, unfair AI-informed decision making systems have been proven to systematically reinforce discrimination such as racial/gender biases in AI-informed decision making. These drive a distrust in and fear the use of AI in public discussions [41].

In addition, the wide use of AI in almost every aspect of our life implies that with great powers comes great responsibility. Fairness shows that an AI system exhibits certain desirable ethical characteristics, such as being bias-free, diversity-aware, and non-discriminatory. While explanations to an AI system provide human-understandable interpretations of the inner working of the system and decisions. Both fairness and explanation are important components for building “Responsible AI”. For example, the fair treatment and/or fair outcome are important ethical issues that need to be considered in the algorithmic hiring decision making. How the decisions made by an algorithmic process can be explained in a transparent and compliant way is also necessary for ethical use of AI in the hiring [36]. Therefore, both fairness and explanations are important ethical issues that can be used to promote user trust in AI-informed decision making (see Fig. 1).

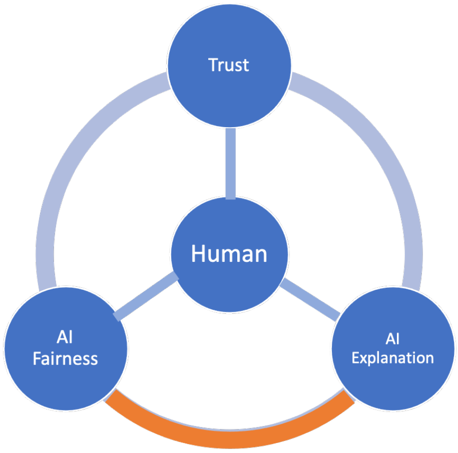

Fig. 1. Relations among AI fairness, AI explanation, and trust.

Previous research found that AI explanations are not only for human to understand the AI system, but also provide an interface for human in the loop, enabling them to identify and address fairness and other issues [12]. Furthermore, differences in AI outcomes amongst different groups in AI-informed decision making can be justified and explained via different attributes in some cases [27]. When these differences are justified and explained, the discrimination is not considered to be illegal [22]. Therefore, explanation and fairness have close relations in AI-informed decision making (as highlighted in orange colour in Fig. 1). Taken the talent recruiting as an example, disproportional recruitment rates for males and females may be explainable by the fact that more males may have higher education, and if males and females are treated equally, it will introduce reverse discrimination, which may be undesirable as well [22]. In another example on the annual income analysis [2], males have a higher annual income than females on average in the data. However, this does not mean that there is a discrimination to females in the annual income because females have fewer work hours than males per week on average. Therefore, the explanation to the difference of the annual income between males and females with the use of work hours per week helps the outcomes of annual income acceptable, legal and fair [22]. It shows that fairness and explanation are tightly related to each other. Therefore, it is significant to understand how AI explanations impact the fairness judgement or how the AI fairness enhances AI explanations. This paper aims to investigate state-of-the-art research in these areas and identifies key research challenges. The contributions of the paper include:

- The relations between explanability and AI fairness are identified as one of significant components for the responsible use of AI and trustworthy decision making.
- A systematic analysis on the explanabillitty and AI fairness to learn the current status of explanability for the human's fairness judgement;
- The challenges and future research directions on the explanability for AI fairness are identified.

2 Fairness

Fairness has become a key element in developing socio-technical AI systems when AI is used in various decision making tasks. In the context of decision-making, fairness is defined as the absence of any prejudice or favoritism towards an individual or a group based on their inherent or acquired characteristics [27, 33]. An unfair algorithm is one whose decisions are skewed toward a particular group. Fairness can be considered from at least four aspects [10]: 1) protected attributes such as race, gender, and their proxies, are not explicitly used to make decisions; 2) common measures of predictive performance (e.g., false positive and false negative rates) are equal across groups defined by the protected attributes; 3) outcomes are independent of protected attributes; and 4) treat similarly risky people similarly.

There are two potential sources of unfairness in machine learning outcomes: those arising from biases in data and those arising from algorithms. Mehrabi et al. [27] summarised 23 types of data biases that may result in fairness issues in machine learning: historical bias, representation bias, measurement bias, evaluation bias, aggregation bias, population bias, Simpson's paradox, longitudinal data fallacy, sampling bias, behavioural bias, content production bias, linking bias, temporal bias, popularity bias, algorithmic bias, user interaction bias, social bias, emergent bias, self-selection bias, omitted variable bias, cause-effect bias, observer bias, and funding bias. Different kinds of discrimination that may occur in algorithmic decision making are also categorised by Mehrabi et al. [27] such as direct discrimination, indirect discrimination, systemic discrimination, statistical discrimination, explainable discrimination, and unexplainable discrimination. Different metrics have been developed to measure AI fairness quantitatively and various approaches have been proposed to mitigate AI biases [6]. For example, statistical parity difference is defined as the difference of the rate of favorable outcomes received by the unprivileged group to the privileged group, and equal opportunity difference is defined as the difference of true positive rates between the unprivileged and the privileged groups. The true positive rate is the ratio of true positives to the total number of actual positives for a given group.

Since the disconnection between the fairness metrics and practical needs of society, politics, and law [21], Lee et al. [24] presented that the relevant contextual information should be considered in an understanding of a model's ethical impact, and fairness metrics should be framed within a broader view of ethical concerns to ensure their adoption for a contextually appropriate assessment of each algorithm.

As AI is often used by humans and/or for human-related decision making, people's perception of fairness is required to be taken into account when designing and implementing AI-informed decision making systems [38]. Following this, people's perception of fairness has been investigated along four dimensions: 1) algorithmic predictors, 2) human predictors, 3) comparative effects (human decision-making vs. algorithmic decision-making), and 4) consequences of AI-informed decision making [38].

3 AI Explanation

The AI explainability has been reviewed thoroughly in recent years [44, 7], which are based on the explanation-generation approaches, the type of explanation, the scope of explanation, the type of model it can explain or combinations of these methods as well as others [1]. For example, explanation methods can be grouped into pre-model, in-model, and post-model methods by considering when explanations are applicable; there are also intrinsic and post-hoc explanation methods by considering whether explainability is achieved through constraints imposed on the AI model directly (intrinsic) or by applying explanation methods that analyse the model after training (post-hoc). Other types of explanations include model-specific and model-agnostic methods, as well as global and local explanation methods.

Miller [28] emphasised the importance of social science in AI explanations and found that 1) Explanations are contrastive and people do not ask why an event happened, but rather why this event happened instead of another event; 2) Explanations are selected in a biased manner. People are adept at selecting one or two causes from an infinite number of causes to be the explanation, which could be influenced by certain cognitive biases; 3) Probabilities probably don't matter. Explanations with statistical generalisations are unsatisfying and the causal explanation for the generalisation itself is usually effective; 4) Explanations are social. They are a transfer of knowledge to people and act as part of a conversation or interaction with people. Therefore, explanations are not just the presentation of associations and causes to predictions, they are contextual.

Wang et al. [39] highlighted three desirable properties that ideal AI explanations should satisfy: 1) improve people's understanding of the AI model, 2) help people recognize the model uncertainty, and 3) support people's calibrated trust in the model. Therefore, different approaches are investigated to evaluate whether and to what extent the offered explainability achieves the defined objective [44]. Objective and subjective metrics are proposed to evaluate the quality of explanations, such as clarity, broadness, simplicity, completeness, and soundness of explanations, as well as user trust. For example, Schmidt and Biessmann [34] presented a quantitative measure for the quality of explanation methods based on how faster and accurate decisions indicate intuitive understanding, i.e. the information transfer rate which is based on mutual information between human decisions and model predictions. [34] also argued that a trust metric must capture cases in which humans are too biased towards the decisions of an AI system and overly trust the system, and presented a quantitative measure for trust by considering the quality of AI models (see Equ. 1).

$$T = \frac{MI_{\hat{\mathcal{Y}}}}{MI_{\mathcal{Y}}} \tag{1}$$

where T is the trust metric, $MI_{\hat{\mathcal{Y}}}$ is the mutual information between human decisions and model predictions and $MI_{\mathcal{Y}}$ is the mutual information between human decisions and true labels.

Despite the extensive investigations of AI explanations, they still face different challenges [29]. For example, similar to AI models, uncertainty is inherently associated with explanations because they are computed from training data or models. However, many AI explanation methods such as feature importance-based approaches provide explanations without quantifying the uncertainty of the explanation. Furthermore, AI explanations, which should ideally reflect the true causal relations [17], mostly reflect statistical correlation structures between features instead.

4 Explanation for AI Fairness

As discussed previously, fairness and explanation are strongly dependent. Deciding an appropriate notion of fairness to impose on AI models or understanding whether a model is making fair decisions require extensive contextual understanding and domain knowledge. Shin and Park [37] investigated the role of Fairness, Accountability, and Transparency (FAT) in algorithmic affordance. It showed that FAT issues are multi-functionally related, and user attitudes about FAT are highly dependent on the context in which it takes place and the basis who is looking at. It also showed that topics regarding FAT are somehow related and overlapping, making them difficult to distinguish or separate. It demonstrated the heuristic role of FAT regarding their fundamental links to trust.

4.1 Explanation guarantees fairness

The explanation of the decision making is a way to gain insights and guarantee fairness to all groups impacted by AI-related decisions [13]. Lee et al. [24] argued that explanations may help identify potential variables that are driving the unfair outcomes. It is unfair if decisions were made without explanations or with unclear, untrusted, and unverifiable explanations [32]. For example, Begley et al. [5] introduced explainability methods for fairness based on the Shapley value framework for model explainability [25]. The proposed fairness explanations attribute a model's overall unfairness to individual input features, even the model does not operate on protected/sensitive attributes directly.

Warner and Sloan [40] argued that effective regulation to ensure fairness requires that AI systems be transparent. While explainability is one of approaches to acquire transparency. The explainability requires that an AI system provides a human-understandable explanation of why any given decision was reached in terms of the training data used, the kind of decision function, and the particular inputs for that decision. Different proxy variables of fairness are presented for the effective regulation of AI transparency in [40].

4.2 Influence of explanation on perception of fairness

Baleis et al. [3] showed that transparency, trust and individual moral concepts demonstrably have an influence on the individual perception of fairness in AI applications. Dodge et al. [12] investigated the impact of four types of AI explanations on human's fairness judgments of AI systems. The four types of explanations are input influence-based explanation, demographic-based explanation, sensitivity-based explanation, and case-based explanation. It showed that case-based explanation is generally less fair. It was found that local explanations are more effective than global explanations for case-specific fairness issues. Sensitivity-based explanations are the most effective for the fairness issue of disparate impact.

4.3 Fairness and properties of features

Grgic-Hlaca et al. [14] proposed to understand why people perceive certain features as fair or unfair to be used in algorithms based on a case study of a criminal risk estimation tool for the use to help make judicial decisions. Eight properties of features are identified, which are reliability, relevance, volitionality, privacy, causes outcome, causes vicious cycle, causes disparity in outcomes, and caused by sensitive group membership. It was found that people's concerns on the unfairness of an input feature are not only discrimination, but also other consideration of latent properties such as the relevance of the feature to the decision making scenario and the reliability with which the feature can be assessed. In a further study, Grgic-Hlaca et al. [15] proposed measures for procedural fairness (the fairness of the decision making process) that consider the input features used in the decision process in the context of criminal recidivism. The analysis examined to what extent the perceived fairness of a characteristic is influenced by additional knowledge about increasing the accuracy of the prediction. It was found that input features that were classified as fairer were those that improved the accuracy of prediction and those features as more unfair that led to discrimination against certain feature holders of people.

4.4 Fairness and counterfactuals

The use of counterfactuals has become one of popular approaches for AI explanation and making sense of algorithmic fairness [26, 4, 44], which can require an incoherent theory of what social categories are [23].

However, it was argued that the social categories may not admit counterfactual manipulation, and hence may not appropriately satisfy the demands for evaluating the truth or falsity of counterfactuals [23], which can lead to misleading results. Therefore, the approaches used for algorithmic explanations to make sense of fairness also need to consider social science theories to support AI fairness and explanations.

A good example of the use of counterfactuals [18] is algorithmic risk assessment [11]. Algorithmic risk assessments are increasingly being used to help experts make decisions, for example, in medicine, in agriculture or criminal justice. The primary purpose of such AI-based risk assessment tools is to provide decision-relevant information for actions such as medical treatments, irrigation measures or release conditions, with the aim of reducing the likelihood of the respective adverse event such as hospital readmission, crop drying, or criminal recidivism. The advantage of the principle of machine learning, namely learning from large amounts of historical data, is precisely counterproductive, even dangerous [19], here.

Because such algorithms reflect the risk from decision-making policies of the past – but not the current actual conditions. To cope with this problem, [11] presents a new method for estimating the proposed metrics that uses doubly robust estimation and shows that only under strict conditions can fairness be provided simultaneously according to the standard metric and the counterfactual metric. Consequently, fairness-enhancing methods that aim for parity in a standard fairness metric can cause greater imbalance in the counterfactual analogue.

5 Discussion

With the increasing use of AI in people's daily lives for various decision making tasks, the fairness of AI-informed decisions and explanation of AI for decision making are becoming significant concerns for the responsible use of AI and trustworthy decision making. This paper focused on the relations between explanation and AI fairness and especially the roles of explanation on AI fairness. The investigations demonstrated that fair decision making requires extensive contextual understanding. AI explanations help identify potential variables that are driving the unfair outcomes. Different types of AI explanations affect human's fairness judgements differently. Certain properties of features such as the relevance of the feature to the decision making scenario and the reliability with which the feature can be assessed affect human's fairness judgements. In addition, social science theories need to be considered in making sense of fairness with explanations. However, there are still challenges. For example,

- Despite the requirements of the extensive contextual understanding for the fair decision making, it is hard to decide what contextual understanding is the appropriate to boost fair decision making.
- There are various types of explanations. It is significant to decide what explanations that can promote the human's fairness judgement on decision making as expected. While the human's fairness judgement is highly related to users themselves, it is a challenge to justify what explanations are the best for human's fairness judgement.
- Since AI is applied in various sectors and scenarios, it is important to understand whether different application sectors or scenarios affect the effectiveness of explanations on the human's judgement on perception in decision making.

Investigating AI fairness explanations requires a multidisciplinary approach and must include research on machine learning [9], human-computer interaction [31] and social science [30] – regardless of the application domain - because the domain expert must always be involved and can bring valuable knowledge and contextual understanding [16].

All this provides us with clues for developing effective approaches to responsible AI and trustworthy decision-making in all future work processes.

6 Conclusion

The importance of fairness is undisputed. In this paper, we have explored the relationships between explainability, or rather explanation, and AI fairness, and in particular the role of explanation in AI fairness. We first identified the relationships between explanation and AI fairness as one of the most important components for the responsible use of AI and trustworthy decision-making. The systematic analysis of explainability and AI fairness revealed that fair decision-making requires a comprehensive contextual understanding, to which AI explanations can contribute. Based on our investigation, we were able to identify several other challenges regarding the relationships between explainability and AI fairness. We ultimately argue that the study of AI fairness explanations requires an important multidisciplinary approach, which is necessary for a responsible use of AI and for trustworthy decision-making - regardless of the application domain.

7 Acknowledgements

The authors declare that there are no conflict of interests and the work does not raise any ethical issues. Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554, explainable AI.

References

1. Arya, V., Bellamy, R.K.E., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K.R., Wei, D., Zhang, Y.: One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. arXiv:1909.03012 [cs, stat] (2019), http://arxiv.org/abs/1909.03012
2. Asuncion, A., Newman, D.: Uci machine learning repository (2007), https://archive.ics.uci.edu/ml/index.php
3. Baleis, J., Keller, B., Starke, C., Marcinkowski, F.: Cognitive and emotional response to fairness in ai – a systematic review (2019), https://www.semanticscholar.org/paper/Implications-of-AI-(un-)fairness-in-higher-the-of-Marcinkowski-Kieslich/231929b1086badcbd149debb0abefc84cdb85665
4. Barocas, S., Selbst, A.D., Raghavan, M.: The hidden assumptions behind counterfactual explanations and principal reasons. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 80–89. FAT* '20 (2020)
5. Begley, T., Schwedes, T., Frye, C., Feige, I.: Explainability for fair machine learning. CoRR **abs/2010.07389** (2020), https://arxiv.org/abs/2010.07389
6. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J.T., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. CoRR **abs/1810.01943** (2018), http://arxiv.org/abs/1810.01943
7. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)

8. Castelvecchi, D.: Can we open the black box of AI? Nature News **538**(7623), 20 (October 2016)
9. Chouldechova, A., Roth, A.: The frontiers of fairness in machine learning. Communications of the ACM **63**(5), 82–89 (2020). https://doi.org/10.1145/3376898
10. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: A critical review of fair machine learning. CoRR **abs/1808.00023** (2018), `http://arxiv.org/abs/1808.00023`
11. Coston, A., Mishler, A., Kennedy, E.H., Chouldechova, A.: Counterfactual risk assessments, evaluation, and fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT 2020). pp. 582–593 (2020). https://doi.org/10.1145/3351095.3372851
12. Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K.E., Dugan, C.: Explaining models: An empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. p. 275–285. IUI '19 (2019)
13. Ferreira, J.J., de Souza Monteiro, M.: Evidence-based explanation to promote fairness in ai systems. In: CHI2020 Fair and Responsible AI Workshop (2020)
14. Grgic-Hlaca, N., Redmiles, E.M., Gummadi, K.P., Weller, A.: Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In: Proceedings of the 2018 World Wide Web Conference. p. 903–912. WWW '18 (2018)
15. Grgic-Hlaca, N., Zafar, M.B., Gummadi, K.P., Weller, A.: Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In: Proceedings of the Thirty-Second AAAI Conferenceon Artificial Intelligence (AAAI-18). pp. 51–60 (2018)
16. Holzinger, A.: Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Informatics **3**(2), 119–131 (2016). https://doi.org/10.1007/s40708-016-0042-6
17. Holzinger, A., Carrington, A., Mueller, H.: Measuring the quality of explanations: The system causability scale (SCS). KI - Kuenstliche Intelligenz **34**(2), 193–198 (2020)
18. Holzinger, A., Malle, B., Saranti, A., Pfeifer, B.: Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai. Information Fusion **71**(7), 28–37 (2021). https://doi.org/10.1016/j.inffus.2021.01.008
19. Holzinger, A., Weippl, E., Tjoa, A.M., Kieseberg, P.: Digital transformation for sustainable development goals (sdgs) - a security, safety and privacy perspective on ai. In: Springer Lecture Notes in Computer Science, LNCS 12844, pp. 1–20. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-84060-0-1
20. Holzinger, K., Mak, K., Kieseberg, P., Holzinger, A.: Can we trust machine learning results? artificial intelligence in safety-critical decision support. ERCIM News **112**(1), 42–43 (2018)
21. Hutchinson, B., Mitchell, M.: 50 years of test (un)fairness: Lessons for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. p. 49–58. FAT* '19 (2019)
22. Kamiran, F., Žliobaitė, I.: Explainable and Non-explainable Discrimination in Classification, pp. 155–170. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
23. Kasirzadeh, A., Smart, A.: The use and misuse of counterfactuals in ethical machine learning. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021). pp. 228–236 (2021)

24. Lee, M.S.A., Floridi, L., Singh, J.: Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. SSRN Scholarly Paper ID 3679975, Social Science Research Network (July 2020), https://papers.ssrn.com/abstract=3679975
25. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4768–4777. NIPS'17 (2017)
26. McGrath, R., Costabello, L., Van, C.L., Sweeney, P., Kamiab, F., Shen, Z., Lécué, F.: Interpretable credit application predictions with counterfactual explanations. CoRR **abs/1811.05245** (2018), http://arxiv.org/abs/1811.05245
27. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. CoRR **abs/1908.09635** (2019), http://arxiv.org/abs/1908.09635
28. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
29. Molnar, C., Casalicchio, G., Bischl, B.: Interpretable machine learning – a brief history, state-of-the-art and challenges. arXiv:2010.09337 [cs, stat] (October 2020), http://arxiv.org/abs/2010.09337
30. Piano, S.L.: Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Humanities and Social Sciences Communications **7**(1), 1–7 (2020). https://doi.org/10.1057/s41599-020-0501-9
31. Robert Jr, L.P., Bansal, G., Melville, N., Stafford, T.: Introduction to the special issue on ai fairness, trust, and ethics. AIS Transactions on Human-Computer Interaction **12**(4), 172–178 (2020). https://doi.org/10.17705/1thci.00134
32. Rudin, C., Wang, C., Coker, B.: The age of secrecy and unfairness in recidivism prediction. Harvard Data Science Review **2**(1) (3 2020). https://doi.org/10.1162/99608f92.6ed64b30, https://hdsr.mitpress.mit.edu/pub/7z10o269, https://hdsr.mitpress.mit.edu/pub/7z10o269
33. Saxena, N.A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D.C., Liu, Y.: How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. p. 99–106. AIES '19 (2019)
34. Schmidt, P., Biessmann, F.: Quantifying interpretability and trust in machine learning systems. In: Proceedings of AAAI Workshop on Network Interpretability for Deep Learning 2019 (2019)
35. Schneeberger, D., Stöger, K., Holzinger, A.: The european legal framework for medical AI. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) Machine Learning and Knowledge Extraction. pp. 209–226. Lecture Notes in Computer Science, Springer International Publishing (2020)
36. Schumann, C., Foster, J.S., Mattei, N., Dickerson, J.P.: We need fairness and explainability in algorithmic hiring. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. p. 1716–1720. AAMAS '20 (2020)
37. Shin, D., Park, Y.J.: Role of fairness, accountability, and transparency in algorithmic affordance. Computers in Human Behavior **98**, 277–284 (2019)
38. Starke, C., Baleis, J., Keller, B., Marcinkowski, F.: Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature (2021)
39. Wang, X., Yin, M.: Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making, p. 318–328. ACM (2021)
40. Warner, R., Sloan, R.H.: Making artificial intelligence transparent: Fairness and the problem of proxy variables. Criminal Justice Ethics **40**(1), 23–39 (2021)

41. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2979–2989. Copenhagen, Denmark (Sep 2017)
42. Zhou, J., Chen, F.: 2d transparency space—bring domain users and machine learning experts together. In: Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent, pp. 3–19. Human–Computer Interaction Series, Springer International Publishing (2018)
43. Zhou, J., Chen, F. (eds.): Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent. Human–Computer Interaction Series, Springer International Publishing (2018)
44. Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics **10**(5) (2021)
45. Zhou, J., Khawaja, M.A., Li, Z., Sun, J., Wang, Y., Chen, F.: Making machine learning useable by revealing internal states update — a transparent approach. International Journal of Computational Science and Engineering **13**(4), 378–389 (2016)