REPHRAIN

Protecting citizens online

Social Explainability of AI: The Impact of Non-Technical Explanations on Trust

Frens Kroeger, Coventry University

Bianca Slocombe, Coventry University

Isa Inuwa-Dutse, University of Huddersfield

Baker Kagimu, University College London

Beate Grawemeyer, Coventry University

Umang Bhatt, Cambridge University

June 2023

Social Explainability of AI: The Impact of Non-Technical Explanations on Trust

Frens Kroeger,[1] Bianca Slocombe,[1] Isa Inuwa-Dutse,[2] Baker Kagimu,[3] Beate Grawemeyer,[4] and Umang Bhatt[5]

[1]Centre for Trust, Peace and Social Relations, Coventry University; [2]University of Huddersfield; [3]University College London; [4]Coventry University; [5]Cambridge University

Abstract

In striving for explainable AI, it is not necessarily technical understanding that will maximise perceived transparency and trust. Most of us board planes with little understanding about how the plane works, and without knowing the pilot, because we put trust in the regulatory and authoritative systems that govern the people and processes. By providing knowledge of the governing ecosystem, industries like aviation and engineering have built stable trust with everyday people. This is known as "social explainability." We extend this concept to AI systems using a series of "social" explanations designed with users (based on external certification of the system, data security and privacy). Core research questions are: Do social explanations, purely technical explanations, or a combination of the two, predict greatest trust from users? Does this depend on digital literacy of the user? An interaction between explanation type and digital literacy reveals that more technical information predicts higher trust from those with higher digital literacy, but those of lower digital literacy given purely technical explanations have the worst trust overall. For this group, social explainability works best. Overall, combined socio-technical explanations appear more successful in building trust than purely social explanations. As in other areas, social explainability may be a useful tool for building stable trust for non-experts in AI systems.

1 Introduction

Algorithms have become part of our daily lives, including in high-stakes areas such as criminal justice [Julia et al., 2019; Grgic-Hlaca et al., 2018], and the medical domain [Sturm et al., 2016]. However, they are viewed as black boxes, owing to the opacity of their decision-making processes [Calmon et al., 2017; Deeks, 2019; Dodge et al., 2019] and users are often left with unanswered questions about the process [Gunning, 2017]. Explainable AI (XAI) is about addressing the black-box nature of AI models to be more transparent and trustworthy. Good explanations are expected to be satisfying enough to enable users to develop a meaningful mental model about the functionality of an AI system and its decision process [Hoffman et al., 2018].

Explanation is typically aimed at offering insights into a model's inner workings as a way of improving transparency and trust [Mothilal et al., 2020; Bhatt et al., 2020; Karimi et al., 2021]. However, implementing XAI is difficult to do with consistency as successful explanations vary across users and contexts [Hoffman et al., 2018]. For example, what system developers view as a useful explanation could be confusing to end users or regulators [Society, 2019]. Multiple studies have measured transparency and trust resulting from different types of explanations offered to stakeholders about the inner workings of AI systems or machine learning [Andras et al., 2018; Dodge et al., 2019; Ferreira and Monteiro, 2020]. While a system developer will prefer technical details about AI's innards, regulators require assurance about the data, and end-users might require a better understanding of which factors led to a decision that affects them [Ferreira and Monteiro, 2020; Stuart et al., 2012]. Because explainability needs vary across the general public, policy makers, and expert users, achieving human-centered XAI requires pluralistic explanations [Society, 2019; Ehsan et al., 2021b].

The current study focuses on digital literacy of the user in implementing XAI. The study expands previous research on XAI to incorporate non-technical explanations about the functioning of the AI's ecosystem (rather than the functioning of the system itself). Particularly for non-experts and those with lower digital literacy (in the case of AI), it is not necessarily a technical understanding of a process that will maximise perceived transparency and trust of a system. Prior to the existence of regulatory bodies and external certification, human beings relied for centuries on interpersonal knowledge to decide who to trust in high-stakes situations, for example, who they would allow to perform surgery. Today, we put trust in the regulatory and authoritative systems that govern and provide oversight [Giddens, 1990]. Most of us board planes with little technical knowledge about how the plane works, and we trust certified surgeons to cut us open despite little technical knowledge of the surgical process. Industries like aviation and engineering have built stable trust with people who do not understand the technology, and this is the aim for AI.

Giddens [1990] suggests that this is possible by providing knowledge of the ecosystem in which a system is developed, rather than about how the system itself technically functions. This is referred to as "social explainability." Rather than focusing on technological comprehension, social explanations draw on users' everyday social understandings of institutional systems, and on value judgements that matter to them. This type of explanation is referred to as socio-organisational [Giddens, 1990]. Social explanations may also be considered multi-dimensional. That is, they go beyond purely technical explanations which typically focus on whether a system (in this case, an algorithm) can do what it is supposed

to do (ability dimension of trustworthiness) and give more detail on the purposes of the specific application (benevolence dimension) and/or the values underlying its development and deployment (integrity dimension of trustworthiness; see [Mayer et al., 1995b]).

This approach treats AI as a socio-technical system. Socio-technical systems consider human agents and social institutions as integral components of technical systems. That is, the systems depend not only on technical hardware, but also human behavior and social institutions to function effectively [Kroes et al., 2006]. AI includes each of these elements and so is considered a kind of sociotechnical system [Van De Poel, 2020]. Sociotechnical approaches have been used to inform trust measurement in AI [Benk et al., 2022].

1.1 The Current Study

Inspired by socio-organisational types of explanation which have been proven to be effective bases of trust in complex systems [Giddens, 1990; Kroeger, 2015, 2017] and previous work on expanding explainability [Ehsan et al., 2021a; Ehsan et al., 2021b], we explore how the knowledge of institutional systems pertaining to AI, and the value judgements that matter to users, affect trust in AI for everyday users.

This study aims to develop and test non-technical (social) explanations of an AI system that are comprehensible, relevant and "normal" to users in the context of their everyday life-worlds (compared to more conventional technical explanations). The current drive to develop sophisticated yet approachable technical explanations as a basis for trust may be misguided if social statements are acceptable and preferable for users with less digital expertise. In addition to ensuring this group is not digitally excluded, suitable social explanations would provide alternative (simpler) paths for policymakers and developers in implementing XAI. Of course, the capacity to garner trust in a system is not necessarily a reflection of the *trustworthiness* of the system. In utilising social explainability as a tool, this must be a consideration.

The following research questions are addressed: Do non-technical social explanations, purely technical explanations, or a combination of the two, predict greatest trust from users? Does this depend on digital literacy of the user? The study was also designed to determine any distinction in trust across social variant (integrity, certification, disclosure).

It is hypothesised that more digitally literate participants will have a preference for technical explanations, and less literate participants will have a preference for social explanations. The combined explanations, including the social and a small amount of technical context (about browsing history) are expected to appeal most to those of moderate expertise who are unlikely to be drawn to purely social or purely technical information.

1.1.2 Measuring Trust

Recent studies have investigated how best to evaluate trust [e.g., Jacovi et al., 2021; Poursabzi-Sangdeh et al., 2021; Benk et al., 2022] and multiple metrics have been proposed to evaluate the impact of explanation on the user's perception of trust and reliance on AI [Cahou and Foryz, 2009; Hoffman et al; Benk et al., 2022]. Within XAI, trust is operationalized differently and there is a need for a clear distinction between behavioral measures of reliance and attitudinal (subjective) measures of trust [Scharowski et al., 2022]. Reliance on AI has been attributed to the alignment between the expected and actual output, and the potential consequences if the model is correct or incorrect [Poursabzi-Sangdeh et al., 2021].

A well-known XAI measure by Hoffman et al. [2018] measures participants' trust in a tool (or system) by asking about their confidence in the tool, its predictability, reliability, and efficiency, among other things. In the current study, ad recommendations are used to represent the AI system. Ads shown to participants are not tailored to participant characteristics (which would be necessary to determine efficacy, reliability, predictability etc.), and participants are not given an option to re-engage with the technology. Therefore, most of these items cannot be included in the current measure of trust. The current study uses a simple attitudinal measure of trust made up of two items from the Hoffman scale: whether participants are confident in the system and like using it for decision-making.

2 Methods

The current study takes place in two stages. Firstly, a series of participatory sessions are run with users to engage them in an open discussion about algorithmic trust within the AI ecosystem. This is in line with previous settings where the onus is on involving public actors to shape explainable artificial intelligence (XAI) and related effects [Biran and Cotton, 2017; Deeks, 2019]. Core themes from the sessions inform the creation of social explanations of AI to be used in the online study which seeks to test the social explainability of AI in building trust. Online ad recommendations are utilised to represent the AI system in this study as they are widely familiar and require no technical expertise for participants to assess.

2.1 Participatory Sessions (Pre-study)

A total of 24 participants took part over 4 sessions (58% female; mean age 24). Participants were drawn from the general population and had a range of expertise and experience related to AI. Participants were asked to discuss their use of, or encounters with, AI in everyday life. Groups chose examples of AI systems to focus on and think about throughout the session (e.g., Alexa, ad recommendations, face recognition). They were given ten minutes to write down what information they would like to have, and what they would like to be explained to them (if they could ask anything about these different types of systems). Participants were instructed that information is not limited to technical knowledge.

Participants then read out their key points and discussed them as a group. Sessions were recorded and transcribed. Key themes were determined by discussion points common

across participants and groups. Key themes from the sessions are listed below. These are things the participants were interested in knowing more about:

Disclosure: how much the AI model understands about the user and what data is collected
Confidentiality: how protected the data is, given the system may make recommendations based on sensitive data
Functionality and robustness: how effective a model is, and whether it is prone to mistakes
Understandability: the need to have reliable and understandable explanations
AI development and certification: the laws and regulation governing AI, including who controls the AI

2.2 Main Study

The use case in this study was based on an AI system that recommends specific ads to online users (online platforms enable specific users to be targeted based on certain information about them that is held or accessed by the interested organisations, maximising the utility of the advertisements). The context of advertising was chosen due to its comprehensibility and familiarity to participants on a large scale. Furthermore, because of the privacy concerns involved, targeted advertisements are of interest to policymakers, developers, and users [O'Neil, 2001]. The social explanations used in this study addressed external certification of the system, data security and privacy (informed by key themes from the participatory sessions).

Explanation Types

Four explanation types (variants) were informed by the participatory sessions and tested in the online user study (three types of social explanation, and one technical/control explanation): The technical provides insight into how the AI system works (the algorithm), and the social addresses what might worry people about it (privacy and data protection). Each of the variants has three examples (ads for roofing, project management, and electronics) included in the study, totaling 12 explanations across four variants.

• **Technical variant**. This was adapted from basic technical explanations of XAI. A statistical explanation was given based in the user's interests, location, demographics, and/or history, which inform the probability that the ad will be relevant. An example is as follows:

> *The AI model used the following ranked features about you to decide whether to display or not to display the ad to you: Relevance of the ad to the time of the day is 25%; Interest in similar ads is 20%; Relevance to offline activity is 30%; Relatedness to location-specific ads is 5%; Relevance to search history is 20%. Because the aggregated score of the above features is high, the AI model recommended the ad to you.*

• **Certification variant** (social): This type of explanation is simply a form of reassurance to the participants that the development process of the AI is certified by a regulatory body. The explanation is: *This ad generation AI has been certified by the Information Commissioner's Office.*

• **Disclosure variant** (social): Participants in the participatory session were interested in knowing what data is collected from them and for what use. This variant addresses the purpose of data collection (benevolence dimension of trustworthiness). This variant also includes info on previous browsing history (a form of technical explanation that provides context to the social). An example is:

> *This ad is shown to you because you have searched for carpentry tools earlier.*
>
> *The above online characteristics about you have been used in a specific way and will only be stored for 24 hours; they will not be used for any other purposes.*

• **Integrity variant** (social): This variant focuses on the integrity dimension of trustworthiness. The question here is whether simple reference to values makes a difference (most likely to users with comparatively lower digital literacy). The explanation is: *We value your right to privacy; therefore, we will not pass on your data.*

As outlined, the disclosure variant includes technical context (based on previous browsing history). For one of the three certification and one of the three integrity examples, we also include this technical context (*This ad is shown to you because you have searched for X earlier*), meaning five of the nine non-technical examples reflect a combined explanation (social and technical; or socio-technical). These will be compared to the examples that include social only.

Participants and Procedure

The sample consisted of 350 participants from the UK, 53% female, with an age range of 18 to 77 (mean age 37). Participants' self-categorised digital skill was 17.7% satisfactory, 46.1% good, and 36.2% excellent. The user survey was distributed via Prolific (an online recruitment platform).

1. Explanation of ad recommendations: this includes an explanation of the targeted ads and outlines some of the information used for targeting, including an example.

2. Presentation of three ad examples: each participant is randomly exposed to three of the nine social examples, or three technical examples (the control group). A total of 40 people participated in both the social and technical sections, resulting in these participants having 6 observations in the study (from 3 social and 3 technical variants).

3. Trust measure: Participants respond to the following statements, on a scale of 0 (I disagree strongly) to 5 (I agree strongly): 'I am confident in the system'; 'I like using the system for decision making'. An average score is used to represent trust for each ad seen.

3 Results

The average trust score across all explanation types and variants was 2.32 (SD = 1.18; range 0 to 5). A regression model was run using generalised estimating equations (GEE) to account for the unstructured nature of repeated measures across variants, with trust score as the dependent variable,

	Wald Chi-square	df	p
Digital literacy	9.528	2	.009
Explanation type	3.347	2	.188
Digital literacy * Explanation type	40.573	4	<.001

Table 1. Model effects; predicting trust

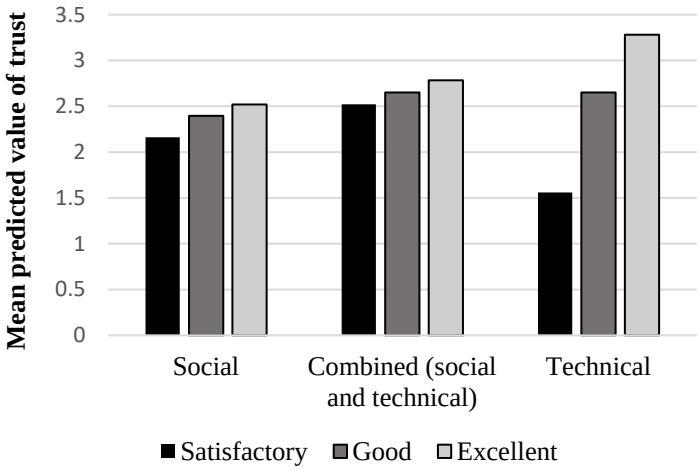

Figure 1. Mean predicted value of trust by explanation type and digital skill

and explanation type (social, technical, combined), variant (integrity, disclosure, certification, technical), age, sex, and digital literacy (satisfactory, good, excellent) as the predictors. This analysis is used to determine which of the predictors have a significant impact on trust scores. The interaction between explanation type and digital literacy is included to examine whether the impact of explanation type differs by category of digital literacy.

In a series of eliminations, non-significant predictors were removed from the model, resulting in the predictors of age, variant type (integrity, disclosure, certification), and sex being removed. The main effect of explanation type is non-significant, but it remains in the model as it is included in the interaction term. The final model is shown in Table 1.

Variant type did not have a significant effect on trust in the model, controlling for other factors. That is, the breakdown of the social variant into categories of integrity, disclosure, and certification did not add explanatory power over and above that provided by broad explanation types (social, combined, technical). In the final model, explanation type was treated simply as these three categories.

The significant interaction term for digital literacy by explanation type indicates that explanation type predicts trust differently when participants have different levels of digital literacy. Figure 1 is an illustration of the interaction. The most notable observation is the divisiveness of the technical condition compared to the stability of trust outcomes in the social and combined conditions. For those with 'good' (moderate) digital literacy, trust appears relatively stable across all three conditions. That is, there is little impact of explanation type for this group. For those of low digital literacy, technical explanations result in less trust, and for those of high digital literacy, technical explanations result in greater trust.

Since the model (inclusive of technical explanations) did not demonstrate an effect of social variant type (integrity, disclosure, certification), a second analysis is run to further investigate the impact of social variant type on trust (excluding the technical data). Since variant type is non-significant when controlling for type of explanation (combined or social), five categories are used which represent the interaction between explanation type and variant type: certification (combined explanation), certification (social explanation) etc. A regression model was run using GEE to account for the unstructured nature of repeated measures across variants, with trust score as the dependent variable, and the five categories, age, sex, and digital literacy as the predictors. In a series of eliminations, non-significant predictors were removed from the model, resulting in the five categories being the only remaining predictor (i.e., trust differs across the five categories).

Pairwise comparisons reinforce the significance of explanation type (combined or social). Using disclosure

	Wald Chi-square	df	p
Five categories	32.246	4	<.001

Table 2. Model effects; predicting trust

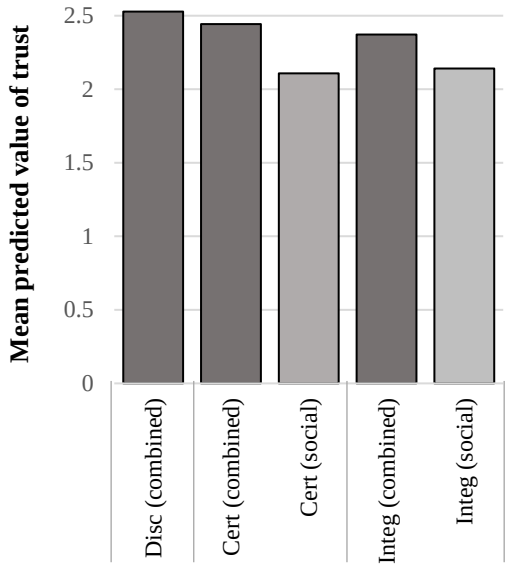

Figure 2. Mean predicted value of trust by category

(the category with the highest trust score) as the reference group, trust score is significantly higher for disclosure than for both social categories (certification and integrity; $p < .001$ in each case), but trust score does not differ significantly between disclosure and the other combined categories ($p = .348$ and .097, respectively). This highlights a distinction between combined explanations (higher trust) and purely social explanations, regardless of variant type. This effect is also illustrated also within variant type (comparing social to combined). For both certification and integrity, combined explanations have significantly higher trust scores than purely social ($p = 0.005$, and $p = .045$, respectively).

4 Discussion

This study extended the concept of social explainability to AI systems using a series of "social" explanations designed with users (based on external certification of the system, data security and privacy). This method is based on providing knowledge of the governing ecosystem, rather than technical information about the system itself, to build trust with those who lack relevant technical expertise [e.g., Giddens, 1990]. Core research questions were: Do social explanations, purely technical explanations, or a combination of the two, predict greatest trust from users? Does this depend on digital literacy of the user?

An interaction between explanation type and digital literacy revealed that more technical information predicts higher trust from those with higher digital literacy, but those of lower digital literacy given purely technical explanations have the worst trust overall (consistent with our hypothesis). These results provide initial insights into social explainability as a useful tool for building stable trust for non-experts in AI systems. However, the results also highlight that social explainability is likely to be problematic if relied upon for building trust with those of greater expertise. This finding makes intuitive sense –an engineer wouldn't be expected to determine trust in a structure based only on knowledge of the ecosystem in which it was constructed.

The combined social/technical explanations used in this study were designed to provide some insight into how trust outcomes change when a small amount of technical information (in this case, about search history) is added to the social explanation. It was expected that combined explanations would appeal most to those of moderate expertise who are unlikely to be drawn to purely social or purely technical information. Instead, adding a small amount of technical context to the social demonstrated small increments in trust scores for all participants (as seen in Figure 1). This was further explored in a follow-up analysis which demonstrated higher trust from combined explanations than purely social explanations, regardless of variant type (Figure 2).

Appropriate interpretation of the combined category is important. It must be recognised that the combined explanations did not include all technical information from the technical explanation type. Therefore, the difference between combined and technical for those of excellent literacy (illustrated in Figure 1) should not be interpreted to mean that adding social information to technical information diminishes trust for this group. Instead, the small amount of technical information offered appears insufficient to improve trust in the social explanations.

The main takeaways from this study are twofold. Firstly, purely technical explanations are divisive (highest trust from the most digitally literate), whereas social or socio-technical explanations result in relatively stable trust across literacy groups but don't elicit maximum trust from the most digitally literate users. Secondly, combined socio-technical explanations result in greater trust in general than purely social explanations, indicating that a small amount of technical information alongside social statements may be important even for novice users. Taken together, these results indicate that socio-technical explanations may be most useful, particularly if the social elements are common in explanations for all users, and the inclusion of technical information is modifiable (more or less technical information) based on user preferences. Under these conditions, trust from all types of users could be maximized.

The non-significance of variant type (breaking the social categories into constituent explanations of integrity, disclosure, and certification) does not mean that differences between these categories do not exist, but that they are non-significant in the context of the model. Our results provide an initial indication that social explainability is a useful tool for some users, but conclusions are not drawn regarding the best type of social explanations to be used. In reality, these types of social explanation are likely to work together, as well as in combination with technical information. A lack of distinction between trust scores for the three combined categories may reflect that the three types of social explanations in this study are equally useful in trust building (i.e., *any* social info has the same psychological effect), but this requires further investigation.

It should also be noted that average trust across all groups did not approach the maximum of 5 and mean trust overall was less than 2.5. Scores may be skewed partially by a simplistic trust measure, but it is clear that more work is needed to maximise outcomes across all groups. Differences between categories of explanation are significant, but they are relatively small. This is to be expected given the low stakes of the AI system used and the preliminary nature of the study. The current study provides a foundation for future research into social explainability for AI.

Limitations and Directions for Future Research

The participatory sessions produced many valuable insights that could not all be covered under the current scope, including concerns about efficacy and functionality of the system. These ability dimensions are likely to feed into trust [Mayer et al., 1995b]. Future research may manipulate and measure functionality of the system, including speed, and whether recommendations are consistent with individual participant characteristics. Ad recommendations shown in this study were static and not tailored to participants, so we were unable to include the impact of functionality on trust.

Differences in cognitive load are also unaccounted for across explanation types. Technical explanations were longer and more detailed than other types, potentially impacting engagement. Cognitive overload has been shown to reduce over-reliance on AI [Bucinca et al., 2022]. Therefore, we might expect that any impact of cognitive load was the reduction of trust ratings for the technical category.

Digital literacy categorisation was self-reported and included only those who are literate enough to use an online survey platform. Those considered to be highly literate in this sample are also unlikely to truly be experts in the area (and for true experts, a technical explanation would be far more involved). The spread of literacy in this sample is therefore not likely to be representative of the population - effects may be more pronounced in a wider sample. Future research may also tease apart general digital literacy and AI expertise.

Ad recommendations represent low-stakes AI systems, relative to use in areas including criminal justice, medicine, and military [e.g., Julia et al., 2019; Sturm et al., 2016]. It is possible that the suitability of social explanations will differ across domains. In striving for explainable AI and resulting trust, particularly in high-stakes areas, we must continue to reflect on the purpose and potential implications, including the distinction between a trusted system and a trustworthy system. XAI may empower everyday citizens to understand and engage with AI in areas where they would otherwise be restricted by technical knowledge. However, XAI also has the capacity to garner trust from those who would not necessarily trust the system if they were able to fully understand its workings. Trust metrics are limited in conveying whether participants should trust a system [Tim, 2022] and understanding the long-term impact of social explanations may also require longitudinal experience (for users to understand the consequences of such decisions). The reliance on AI for critical decisions necessitates a reliable metric to assess when to trust the AI [Schemmer et al., 2022].

With these considerations in mind, this study provides a valuable foundation for the social explainability of AI, pointing towards the potential futility of developing sophisticated technical explanations to suit a general population if social statements (or socio-technical explanations) are acceptable and appropriate to gain user trust.

Acknowledgments

This research was funded by the National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN).

References

[Andras et al., 2018] P. Andras, L. Esterle, M. Guckert, T. A. Han, P. R. Lewis, K. Milanovic, T. Payne, C. Perret, J. Pitt, S. T. Powers. Trusting intelligent machines: Deepening trust within socio-technical systems. IEEE Technology and Society Magazine, 37(4):76–83, 2018.

[Bansal et al., 2020] G. Bansal, T. Wu, J. Zhu, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. S. Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. arXiv preprint arXiv:2006.14779, 2020.

[Bhatt et al., 2020] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley. Explainable machine learning in deployment. In Proceedings of the 2020 Conference on Fairness, Ac-countability, and Transparency, pages 648–657, 2020.

[Benk et al., 2022] M. Benk, S. Tolmeijer, F. von Wangenheim, and A. Ferrario. The value of measuring trust in ai-a socio-technical system perspective. arXiv preprint arXiv:2204.13480, 2022.

[Biran and Cotton, 2017] O. Biran and C. Cotton. Explanation and justification in machine learning: A survey. In IJCAI-17 workshop on explainable AI (XAI), volume 8, 2017.

[Bucinca et al., 2022] Z. Bцcinca, M. B. Malaya, and K. Z. Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1):1–21, 2021.

[Calmon et al., 2017] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 3995–4004, 2017.

[Carhour and Forzy, 2009] B. Cahour and J.-F. Forzy. Does projection into use improve trust and exploration? an example with a cruise control system. Safety science, 47(9):1260–1270, 2009.

[Carton et al., 2016] S. Carton, J. Helsby, K. Joseph, A. Mahmud, Y. Park, J. Walsh, C. Cody, C. E. Pat- terson, L. Haynes, and R. Ghani. Identifying police officers at risk of adverse events. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 67–76, 2016.

[Deeks, 2019] A. Deeks The judicial demand for explainable artificial intelligence. Columbia Law Review, 119(7):1829–1850, 2019.

[Dodge et al., 2019] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan. Explaining models: an empirical study of how explanations impact fairness judgment. In Proceedings of the 24th International Conference on Intel-ligent User Interfaces, pages 275–285, 2019.

[Ehsan et al., 2021a] U. Ehsan, Q. V. Liao, M. Muller, M. Riedl, and J. Weisz. Expanding explainability: towards social transparency in ai systems. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 41:1–19, 2021.

[Ehsan et al., 2021b] U. Ehsan, S. Passi, Q. V. Liao, L. Chan, I. Lee, M. Muller, M. O. Riedl. The who in explainable ai: How ai background shapes perceptions of ai explanations. arXiv preprint arXiv:2107.13509, 2021.

[Ferreira and Monteiro, 2020] J. J. Ferreira and M. S. Monteiro. What are people doing about xai user experience? a survey on ai explainability research and practice. In Inter-national Conference on Human-Computer Interaction, pages 56–73. Springer, 2020.

[Giddens, 1990] A. Giddens. The consequences of modernity. Oxford, Polity Press, 1:1–19, 1990.

[Grgic-Hlaca et al., 2018] N. Grgic-Hlaca, E. M. Redmiles, K. P. Gummadi, and A. Weller. Human perceptions of fair-ness in algorithmic decision making: A case study of criminal risk prediction. In Proceedings of the 2018 World Wide Web Conference, pages 903–912, 2018.

[Gunning, 2017] D. Gunning. Explainable Artificial Intelligence (XAI), DARPA/I2O, 2017.

[Hargittai, 2009] E. Hargittai. An update on survey measures of web-oriented digital literacy. Social science computer review, 27(1):130–137, 2009.

[Hoffman et al., 2018] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable ai: Challenges and prospects. arXiv preprint arXiv:1812.04608, 2018.

[Jacovi et al., 2021] A. Jacovi, A. Marasovi´c, T. Miller, and Y. Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 624–635, 2021.

[Julia et al., 2019] A. Julia, L. Jeff, M. Surya, and K. Lauren. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks, 2019.

[Karimi et al., 2021] A.-H. Karimi, B. Sch¨olkopf, and I. Valera. Algorithmic recourse: from counterfactual expla-nations to interventions. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparen-cy, pages 353–362, 2021.

[Kim, 2003] D. Kim. The effects of trust-assuring arguments on consumer trust in internet stores. 2003.

[Kroeger, 2015] F. Kroeger. The development, escalation and collapse of system trust: From the financial crisis to socie-ty at large. European Management Journal, 33(6):431–437, 2015.

[Kroeger, 2017] F. Kroeger. Facework: creating trust in systems, institutions and organisations. Cambridge Jour-nal of Economics, 41:487–514, 2017.

[Kroeger, 2020] F. Kroeger. What is trust in technology? conceptual bases, common pitfalls and the contribution of trust research. 2020.

[Kroes et al., 2006] P. Kroes, M. Franssen, I. v. d. Poel, and M. Ottens. Treating socio-technical systems as engineering systems: some conceptual problems. Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research, 23(6):803–814, 2006

[Mayer et al., 1995a] R. Mayer, J. Henry, and D Schoorman. An integrative model of organizational trust. The Academy of Management Review, 20:709–734, 1995.

[Mayer et al., 1995b] R. C. Mayer, J. H. Davis, and F. D Schoorman. An integrative model of organizational trust. Academy of management review, 20(3):709–734, 1995.

[Mothilal et al., 2020] R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparen-cy, pages 607–617, 2020.

[O'Neil, 2001] D. O'Neil. Analysis of Internet Users' Level of Online Privacy Concerns. Social Science Computer Review. 2001;19(1):17-31. doi:10.1177/089443930101900103

[Poursabzi-Sangdeh et al., 2021] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach. Manipulating and measuring model interpreta-bility. In Proceedings of the 2021 CHI conference on hu-man factors in computing systems, pages 1–52, 2021.

[Scharowski et al., 2022] N. Scharowski, S. A. Perrig, N. von Felten, and F. Br¨uhlmann. Trust and reliance in xai–distinguishing between attitudinal and behavioral measures. arXiv preprint arXiv:2203.12318, 2022.

[Schemmer et al., 2022] M. Schemmer, P. Hemmer, N. K¨uhl, C. Benz, and G. Satzger. Should i follow ai-based advice? measuring appropriate reliance in human-ai decision-making. arXiv preprint arXiv:2204.06916, 2022.

[Society, 2019] T. R. Society. Explainable ai: the basics, policy briefing, 2019.

[Stuart et al., 2012] H. C. Stuart, L. Dabbish, S. Kiesler, P. Kinnaird, and R. Kang. Social transparency in networked information exchange: a theoretical framework. In Proceedings of the ACM 2012 conference on Computer Sup-ported Cooperative Work, pages 451–460, 2012.

[Sturm et al., 2016] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller. Interpretable deep neural networks for single-trial eeg classification. Journal of neuroscience methods, 274:141– 145, 2016.

[Tim 2022] M. Tim. Are we measuring trust correctly in explainability, interpretability, and transparency research? TRAIT: Trust and Reliance in AI-Human Teams, 2022.

[Tintarev and Masthoff, 2007] N. Tintarev and J. Masthoff. Effective explanations of recommendations: user-centered design. In Proceedings of the 2007 ACM conference on Recommender systems, pages 153–156, 2007.

[Van Deursen et al., 2014] A. J. Van Deursen, E. J. Helsper, and R. Eynon. Measuring digital skills. from digital skills to tangible outcomes project report. 2014.

[Van De Poel, 2020] I. van de Poel. Embedding values in artificial intelligence (AI) systems. Minds and Machines, 30(3):385–409, 2020.