Cognitive Computing and Internet of Things, Vol. 124, 2024, 144–154
https://doi.org/10.54941/ahfe1004712

Explainability as a Means for Transparency? Lay Users' Requirements Towards Transparent AI

Johanna M. Werz, Esther Borowski, and Ingrid Isenhardt

Laboratory for Machine Tools and Production Engineering: Intelligence in Quality Sensing (WZL-IQS) of RWTH Aachen University, 52068 Aachen, Germany

ABSTRACT

With the rise of increasingly complex artificial intelligent systems (AI), their inner processes have become black boxes. The failure of some systems and the largely unregulated market of digital services have prompted governments and organs such as the EU to work on legislation for regulation. Their main requirement is that AI must be transparent for all stakeholders. While AI developers and experts have worked on interpretability and Explainability, social scientists emphasize that explainable AI is hardly understandable for lay users. The question arises as to whether the concept of Explainability can be used to create transparency for laypersons and what (additional) requirements these users might have towards transparent AI. To answer the questions, three fictitious AI apps were discussed in focus groups with n=26 participants. The apps differed in their domain and error significance to be able to identify system dependent requirements. The results indicate that lay users have different expectations and requirements for transparency in AI than technical experts: (a) previous experience with domain and system(s) strongly shape transparency demands, (b) background information beyond Explainability concepts is highly relevant for building trust, and (c) the system factor error-significance acts as a burning glass for transparency requirements. As a summary, the qualitative study shows that Explainability cannot serve as the only means of making systems transparent for lay users. Possible implications for system development are discussed. These implications apply in particular to AI that addresses lay users, i.e. non-computer experts.

Keywords: XAI, Transparent AI, Understandability, User-centered design, Democratic AI

INTRODUCTION

The influence of artificial intelligence (AI) on our work and personal lives has been increasing (Littman et al., 2021). What is currently referred to as artificially intelligent, be it for creating images (Dall-E) or creating text in chat (ChatGPT), is predominantly a black box in terms of how it works, even for the developers of the systems (OpenAI, 2022, 2023).

Under the keywords Explainability or XAI (eXplainable AI) and interpretability, developers and computer science researchers are working on methods to make AI black boxes transparent (Arrieta et al., 2020; Miller, 2019; Mohseni et al., 2021). On the one hand, Explainability is making systems opaque mainly for computer experts. On the other hand, it follows the

assumption "that by building more transparent, interpretable, or explainable systems, users will be better equipped to understand and therefore trust the intelligent agents" (Miller, 2019, p. 3).

However, the research on transparent AI for lay users is inconsistent. Studies find increased usage and trust as well as "worse perceptions of a system, trusting it less because the transparency led [users] to question the system even when it was correct" (Springer, 2019, p. 101).

Therefore, the question has to be answered to what extend the concept of Explainability can be used to create transparency for laypersons and which (additional) requirements lay users might have towards transparent AI. To this end, we conducted three focus groups, in which the participants discussed their requirements for transparent AI based on three fictitious AI apps.

THEORETICAL BACKGROUND

When the term transparency gained popularity in the 1990s, it primarily referred to an understanding of economic policy (Larsson and Heintz, 2020). Later, transparency was often associated with terms such as openness, which in a technical context stands for positive attributes such as "open data" or "open source" (Larsson and Heintz, 2020). At the same time, Stohl et al. (2016) emphasize the "transparency paradox": "when there is an abundance of information available, it is often difficult to obtain useful, relevant information" (p. 134). Therefore, neither visibility nor transparency should be the ultimate goal, but visibility should be managed to improve the effective use of information.

With digitalization and, more recently, the spread of AI systems, the understanding of the concept transparency has expanded to become a prerequisite for the ethical and responsible use of data (Larsson and Heintz, 2020). Transparency has thus become a "modern, surprisingly complex [...] ideal" (Koivisto, 2016, p. 2). That is, transparency in AI has turned into a prerequisite for human decision-making autonomy and thus a target state to be established to meet ethical requirements – and new legal frameworks. EU policy initiatives, such as the Digital Service Act and the work on AI regulations since 2020, show the growing focus on transparency and control in the use of AI (Digital Services Act, 2022). In parallel, a draft law on the regulation of artificial intelligence has been under development since 2021 (AI Act, 2023) requiring systems to be "sufficiently transparent to enable users to interpret the system's output and use it appropriately" (AI Act, 2022).

Definitions of transparent AI range from AI mechanisms and their underlying logic to the possibility of gaining insight into the black box, improving systems, establishing accountability, and preventing discrimination (Ananny and Crawford, 2018). However, these perspectives do not explain what transparency "means, to whom it is related, and to what extent it is beneficial" (Felzmann et al., 2020, p. 3336). As a result, the concept and understanding of transparency remains "quite malleable and therefore [...] can mean all things to all people" (Fox, 2007, p. 664) – even today, more than 15 years after this quotation.

Two technical approaches to the topic of transparent AI are subsumed under the concepts of interpretability and Explainability (eXplainable AI or XAI). Interpretability is a passive property; it exists in simpler machine learning models whose processes are inherently interpretable, i.e. understandable (Arrieta et al., 2020; Brasse et al., 2023; Mohseni et al., 2021). Explainability, on the other hand, needs to be established, which is usually achieved through subsequent – post-hoc – explanations for a system that would otherwise be incomprehensible and a black box (Ali et al., 2023; Brasse et al., 2023; Herm et al., 2022; Littman et al., 2021; Miller, 2019; Mohseni et al., 2021).

While the research on Explainability is called XAI, the term also refers to the result of this research: the explainable AI itself. Very often, XAI is classified along two categories: on the model validity and on the explanation. The model validity distinguishes between model-specific and model-agnostic methods (Arrieta et al., 2020; Brasse et al., 2023). The distinction between agnostic and specific is relevant to the construction and application of XAI, but the results may appear identical to users. This is different for the explanation-dependent categorization. A very popular distinction is drawn between global and local explanations (Ali et al., 2023; Littman et al., 2021; Mohseni et al., 2021; Molnar, 2019). Global Explainability provides information about the internal processes of a model, how it works in general. It answers the "how" question. Local Explainability refers to individual outcomes or predictions of the AI system: why this result happened and not another. Thus, local explanations answer the "why" question (Herm et al., 2022; Molnar, 2019).

As for many years XAI research has been working on making black box models explainable to computer experts, it followed the idea that comprehensive explanations can help AI experts gain insight into a system (Molnar, 2019; Páez, 2019). However, these requirements would not yet provide transparency or insight for end users. The awareness of this gap grew in the 2020s. With it came the realization that explanations should "depend on the context, the severity of the consequences of their decision [...] and the relevant stakeholders" (Felzmann et al., 2020, p. 3348). Thus, to achieve AI transparency for end users, capabilities from computer science have to be accompanied by insights from social science (Larsson and Heintz, 2020).

However, research on the effect of transparency on end users remains inconclusive. A study by Schmidt et al. (2020) showed negative effects of transparency on information, comprehension, and trust: In a text classification task, participants over-relied on the algorithm for complex texts and followed it even in the presence of errors. In the case of inconsistent explanations, participants incorrectly decided against the algorithm's advice, presumably under the assumption that it was wrong. On the other hand, Shin (2021) showed that transparency in recommending news articles increased the trust of the users and helped them to understand the decision-making process of AI algorithms. Other approaches investigated the communication of accuracy of AI processes as a means to establish transparency and increase advice taking (Werz et al., 2020).

The question of which type of explanation, local or global, is more important to end users is also not consistently assessed in the literature. In a

comparison of local and global explanations, Wanner et al. (2022) showed that end users preferred local explanations regardless of the importance of the task. In contrast, a study of business students who received information from a support system to forecast product demand showed that the provided global explanations had a negative impact on the use of algorithmic advice leading to significantly poorer performance (Lehmann et al., 2020). In part, the inconclusive findings of previous studies might go back to the fact that the term "transparency" itself is a container term meaning different things to different people (Fox, 2007; Miller, 2019). It remains hardly defined what transparency means in a specific situation.

In broad, transparency refers to the ability to understand how a system works and why it produces certain results – in a way that is perceived as understandable and sufficiently informative. However, with respect to end users, who often have little or no prior technical knowledge, may be domain experts, or simply private users of AI, the question has to be answered what transparency means to them – and whether it can be broken down into how and why or into other parts. This question of *what end users consider necessary and sufficiently informative in the context of different AI systems* is the lead question of the current study.

While the study focuses less on user factors or environmental factors, it addresses the AI application and its characteristics. As the wide variety of results of usage studies in AI, but also in the technology sector as a whole, show, the context of the application is of fundamental importance for questions of trust and usage (e.g., Ali et al., 2023; Lim and Dey, 2011; Mohseni et al., 2021). Application areas and system characteristics influence users' perceptions of AI systems as well as their requirements for the systems. The aim of the research question was to consider different effects and expectations when investigating AI systems. Accordingly, three exemplary AI applications were created that differ in their task context, usability, and relevance.

METHOD

A discussion with three focus groups was conducted to determine the users' requirements for AI transparency based on different system factors. The three groups took place in September and October 2021. While there were more topics under investigation, for the current analysis we concentrated on *which types of transparency are particularly important for lay users* and whether their requirements go beyond the rather technical concept. What is more, the goal was to find out *which AI system factors lead to the different transparency requirements*. The analysis of the material took place in the project FAIRWork.

For the focus group discussion, the participants were divided into three groups and discussed three fictitious AI applications, each presented by a different moderator. A short description and a screenshot of the fictitious applications served as the basis for the discussion (see Figure 1). Each group consisted of three to four participants. After discussing one app for 15 minutes, the group switched the app and discussed the next question.

This way, all participants discussed all apps but with a different focus. The questions of the three rounds were:

1. What does the app need to explain?
2. Under what conditions would you use the app?
3. How do you react when you realize that the app is wrong?

The Three Apps

The three fictional apps, a financial investment app, a mushroom identification app, and a music selection app will be called Finance App, Mushroom App, and Music App in the following. They were developed to represent different system factors. A pre-test was carried out to ensure that the users perceived the assumed system factors of the three apps accordingly. The apps differed in the form of interaction: written, pictorial, and voice. They differed in their functional requirements: security and accuracy in the Finance App, validity and accuracy in the Mushroom App, and simplicity and accessibility in the Music App. The Finance App and Mushroom App were rated as having a high error-significance, i.e. errors have a high impact and were associated with a high possible loss. The Music App had a low error-significance. The pre-test confirmed that the three apps were equally understandable and aesthetically appealing.

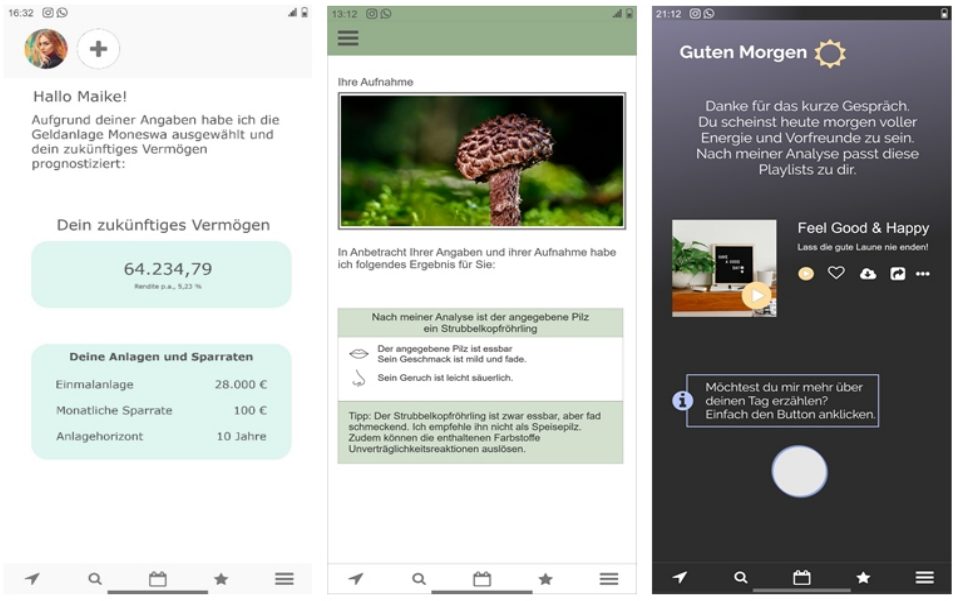

Abbildung 13: Die drei abgebildeten Screenshots der Apps dienten in den Fokusgruppen als Grundlage zur Diskussion

Figure 1: The three screenshots of the fictitious apps were the basis of the discussion.

Sample

In three focus group sessions, a total of $n = 26$ people took part, $n = 15$ identified themselves as women. Half of the participants were students and half were employees. On a scale from 1 (*no knowledge at all*) to 5 (*very much knowledge*), the participants indicated a medium level (median = 3) level of knowledge of AI, with 80% of respondents selecting level 1, 2 or 3.

Focus Group Analysis

To evaluate the focus groups, all video recordings were transcribed and anonymized. The subsequent evaluation of the transcripts was carried out with a summarizing qualitative content analysis according to Mayring (2010). An inductive process was complemented by a deductive phase in which the top-level category "transparency" was added as well as several sub-categories such as "local" and "global". During coding, individual words served as coding units and the participants' entire statements served as context units. In comparing the category systems across the three apps, we were able to identify overlapping, similar, and different categories in order to draw conclusions about the effects of the system factors.

RESULTS AND DISCUSSION

The analysis revealed three pillars of transparency requirements. First, the importance of the domain of the application and prior experiences with domain and system(s), second the importance of background information beyond local and global Explainability, and third the effect of the system factor error-significance.

The influence of prior experience emerged across apps. For instance, in the discussion about the Finance App, participants showed a lot of scepticism linked to previous financial experiences with apps but also with banks or bank accountants: "*I don't trust it. I fell on my nose once, now I leave my fingers off [investments*" (FG2, 126-128). While the Music App faced the highest willingness to be tested, questions about costs reflected prior experiences and comparisons with existing music services such as Spotify. For the Mushroom App, users emphasized the presentation of results. They seemed to compare the app to recommender systems and therefore adapted to experiences with these systems: they wished for certainty scores and alternative results.

As the examples show, prior experience shapes attitudes toward AI systems. These experiences comprise technical systems but also interactions with institutions or humans, and negative sentiments toward whole domains. Additionally, the experience with similar systems influences the expectations towards new ones: their quality and functionalities – and their transparency.

The results illustrate that transparency often concerns specific aspects of a system rather than the entire system. Concerns about background processes in the Finance App led to discussions on banks and investments. As many participants found these topics highly sensitive, they demanded security measures, information about the app's business model, and ideally independent

trust certificates. Music App discussions focused on data privacy due to the perceived sensitivity of voice: "*[…] especially when it comes to voice recognition and mood data and what happened on my day. These are very, very personal things and I would really like to know what is happening with them*" (FG1: 205-209).

When designing AI systems, developers have to consider potential experiences with similar systems and attitudes toward the application domain. Most often, these aspects will not concern entire AI systems but certain aspects of them. User involvement is highly recommended to identify these sensitive aspects, as they might not be obvious to developers and differ between users (i.e., user-centered design).

The second pillar of results concerned background information in addition to local and global Explainability. While the wish for global Explainability remained very broad, local Explainability, explaining how individual results are derived, was of interest in all apps. It did not confine itself to high-risk decisions but comprised a more general need for information:

"*And then […], if they don't want to make the code public, which a non-IT person can't understand anyway, that you really tell the user, ok, based on the keywords, based on your voice colour, on your tone of voice, we found out this and that*" (FG 3, 146-152).

However, an even more important transparency aspect that came up frequently in the two high-risk apps was information about the creators or the data basis of the apps. Participants seemed to hope that knowing the apps' developers could offer insights into the systems' quality:

"*[The app] says 'I don't recommend it as an edible mushroom'. That raises the question: who is this 'I'? Is it Mr. Muller from next door who has just developed an app and scanned some mushrooms from the encyclopaedia? Or is it perhaps the German Society for Mushroom Research? […] Which makes it rather more realistic that what it says could be true*" (FG2_2, 42-47).

In the case of the Finance App, the desire for background information manifested in one for transparency regarding the underlying business model and potential profit interests. Participants believed they could gain trust in the systems through well-known institutions: established names of banks, institutions, or certifying entities provided the opportunity to vouch for a new, less-known system.

For these aspects, the influence of novelty was evident, particularly in the Music App. Here, questions went beyond the established feature of music selection but concerned language analysis, mood identification, and their connection to the results. The desire for transparency seems to show a dependence on novelty that might even diminish when processes or systems become familiar. What is more, the boundaries between global and local transparency blurred. It appeared that laypersons do not make this technical, theoretical distinction. Instead, in addition to Explainability, other aspects contribute to a lay understanding of transparency.

The third pillar concerns the system factor error significance. This factor differentiated the Music App (low error significance) from Mushroom and Finance App (high error significance) as rated in a pre-test and stated by participants during the discussions. Higher error significance seemed to

effect a greater demand for background information. As mentioned above, participants sought details that could provide trust or promise assurances: Risk management requirements, security certificates or audits by independent institutions were raised almost solely for the Mushroom and the Finance App:

"*Of course, it helps if the whole thing is based on a company that has been around for a while, for example a bank that has been around for a long time. [...]*" (FG1, 723-731).

In contrast, the Music app did not prompt requests for information on authorship. Trust in this context was established more through participants testing and evaluating the app themselves or by ratings of other users. The system factor error significance manifested in a higher sensibility, especially towards the background institution: qualified developers, trustworthy providers, and an independent institute validating the app. In contrast, for the less risky Music App assurances were relevant only regarding specific aspects, in this case data privacy.

Despite users demanding transparency independently of system factors, the AI type, domain of the application and previous experience with domain and system(s) shape the kind of transparency they require. On the one hand, the requirements towards and concepts of transparency change as system features change. On the other hand, the systems' parts of which users demand transparency change as well. Finally, the factor error significance acts like a burning glass, intensifying all concerns and requirements.

LIMITATIONS

Several limitations of the study have to be considered when interpreting the results. For one, the very broad concept of transparency might go back to the way of questioning in the focus groups. As the moderators did not ask for transparency definitions directly, the discussed topics were broad and touched on many concepts. A next study could add direct questions for transparency, e.g., at the end of the groups, to gather users' insights directly. It could also test whether such directly inquired concepts of transparency differ from the comprehensive ones identified in this study.

What is more, qualitative studies like interviews or focus groups can only conduct what participants think and say explicitly. However, a lot of research has worked on the gap between intention and action. The privacy paradox, for instance, is the effect that users of online services state to care a lot about privacy issues. However, when usage is under investigation, they hardly show any concerns (Barth and de Jong, 2017). A qualitative study can serve to explore people's concepts and reasoning. However, quantitative studies investigating the usage of transparent AI should validate qualitative results with experimental designs.

Lastly, previous studies on transparency and Explainability – just as the current one – show the high dependence of results on context, domain, and user group. Future studies should validate the findings for other contexts, e.g., the working world where time pressure, accountability and (in)voluntariness come into play.

IMPLICATIONS AND CONCLUSION

In summary, the analysis of the concept of transparency for lay users reveals the importance of the domain of application, prior experiences, and the system factor error significance. Additionally, laypersons possess a comprehension of transparency extending beyond technical aspects. This comprehensive perspective is evident in their preferences regarding local versus global transparency: they expect local transparency, partially global transparency, with a pronounced emphasis on supplementary background information comprising authorship, third-party evaluations, data protection, as well as mechanisms for control and security.

With these implications for AI development, the current study adds to the recent discussions on transparent AI. While for a long time technical solutions have been in the focus, the study sheds light on lay users' requirements towards transparent AI. Adding to the concept of human-centered AI, transparency concepts require the users' views to be able to build reliable and trustworthy systems.

Despite its limitations, the study provides a perspective on what AI developers should consider when developing for lay usage. Quantitative studies adding to the current qualitative results as well as further perspectives on usage, e.g. in the working context, would extend the picture of how to establish AI transparency for lay end users.

ACKNOWLEDGEMENT

This work has been partially supported by the FAIRWork project (www.fairwork-project.eu) and has been funded within the European Commission's Horizon Europe Programme under contract number 101069499. This paper expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this paper.

REFERENCES

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., et al. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence, Information Fusion Volume 99, p. 101805, doi: 10.1016/j.inffus.2023.101805.

Ananny, M. and Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability, New Media & Society Volume 20 No. 3, pp. 973–989, doi: 10.1177/1461444816676645.

Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion Volume 58, pp. 82–115, doi: 10.1016/j.inffus.2019.12.012.

Barth, S. and de Jong, M. D. T. (2017). The privacy paradox – Investigating discrepancies between expressed privacy concerns and actual online behavior – A systematic literature review, Telematics and Informatics Volume 34 No. 7, pp. 1038–1058, doi: 10.1016/j.tele.2017.04.013.

Brasse, J., Broder, H. R., Förster, M., Klier, M. and Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions, Electronic Markets Volume 33 No. 1, pp. 1–30, doi: 10.1007/s12525-023-00644-5.

Felzmann, H., Fosch-Villaronga, E., Lutz, C. and Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence, Science and Engineering Ethics Volume 26 No. 6, pp. 3333–3361, doi: 10.1007/s11948-020-00276-4.

Fox, J. (2007). The uncertain relationship between transparency and accountability, Development in Practice Volume 17 No. 4–5, pp. 663–671, doi: 10.1080/09614520701469955.

Herm, L.-V., Heinrich, K., Wanner, J. and Janiesch, C. (2022). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability, International Journal of Information Management, p. 102538, doi: 10.1016/j.ijinfomgt.2022.102538.

Koivisto, I. (2016). The anatomy of transparency: the concept and its multifarious implications, EUI Working Paper MWP Volume 19.

Larsson, S. and Heintz, F. (2020). Transparency in artificial intelligence, Internet Policy Review Volume, 9 No. 2.

Lehmann, C. A., Haubitz, C. B., Fügener, A. and Thonemann, U. W. (2020). Keep It Mystic? – The Effects of Algorithm Transparency on the Use of Advice.

Lim, B. Y. and Dey, A. K. (2011) "Investigating intelligibility for uncertain context-aware applications", proceedings of the 13th International Conference on Ubiquitous Computing, Association for Computing Machinery, New York, NY, USA, pp. 415–424, doi: 10.1145/2030112.2030168.

Littman, M. L., Ajunwa, I., Berger, G., Boutilier, C., Currie, M., Doshi-Velez, F., Hadfield, G., et al. (2021). Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report, Stanford University, Stanford, CA.

Mayring, P. (2010). Qualitative Inhaltsanalyse: Grundlagen Und Techniken (12. Überarbeitete Auflage), (12th ed.), Beltz, Weinheim.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence Volume 267, pp. 1–38, doi: 10.1016/j.artint.2018.07.007.

Mohseni, S., Zarei, N. and Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems, ACM Transactions on Interactive Intelligent Systems Volume, 11 No. 3–4, p. 24:1-24:45, doi: 10.1145/3387166.

Molnar, C. (2019). Interpretable Machine Learning, (1st ed.), Christoph Molnar (CC Attribution 2.0).

OpenAI. (2022). Introducing ChatGPT. OpenAI, 30 November, available at: https://openai.com/blog/chatgpt (accessed 3 March 2023).

OpenAI. (2023). DALL·E 2. OpenAI, available at: https://openai.com/product/dall-e-2 (accessed 3 March 2023).

Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI), Minds and Machines Volume 29 No. 3, pp. 441–459, doi: 10.1007/s11023-019-09502-w.

Springer, A. (2019). Accurate, Fair, and Explainable: Building Human-Centered AI, UC Santa Cruz.

Stohl, C., Stohl, M. and Leonardi, P. M. (2016). Managing opacity: Information visibility and the paradox of transparency in the digital age, International Journal of Communication Volume 10, pp. 123–137.

Wanner, J., Herm, L.-V., Heinrich, K. and Janiesch, C. (2022). A social evaluation of the perceived goodness of explainability in machine learning, Journal of Business Analytics, Taylor & Francis Volume 5 No. 1, pp. 29–50, doi: 10.1080/2573234X.2021.1952913.

Werz, J. M., Borowski, E. and Isenhardt, I. (2020) "When imprecision improves advice: Disclosing algorithmic error probability to increase advice taking from algorithms", in: Stephanidis, C. and Antona, M. (Eds.), HCI International 2020 - Posters, Springer International Publishing, Cham, pp. 504–511, doi: 10.1007/978-3-030-50726-8_66.