The Importance of Trust and Acceptance in User-Centred XAI - Practical Implications for a Manufacturing Scenario

ERIKA PUIUTTA, OFFIS Institute of Informatics, Germany
LARBI ABDENEBAOUI, OFFIS Institute of Informatics, Germany
SUSANNE BOLL, OFFIS Institute of Informatics, Germany

Overall, the importance of explainability in AI is increasingly recognised. For instance, the consequences of explainability in safety-critical domains such as healthcare have started to be systematically studied. However, other non-critical everyday AI applications rarely consider the impact of explainability on lay persons. In this position paper, we argue for the potentially positive impact that explainability has on the collaboration between humans and AI systems, especially with regards to trust and acceptance. In our on-going research project 'Digitopias', we focus on the collaboration between workers and AI systems in the manufacturing sector. After presenting common practices and challenges in XAI user studies and studies investigating trust and acceptance, we describe our practical attempt to address these challenges in the developed manufacturing scenario.

CCS Concepts: • **Human-centered computing** → **User studies**; **Field studies**.

Additional Key Words and Phrases: eXplainable AI, Human-Computer Interaction, Trust, Acceptance

ACM Reference Format:
Erika Puiutta, Larbi Abdenebaoui, and Susanne Boll. 2023. The Importance of Trust and Acceptance in User-Centred XAI - Practical Implications for a Manufacturing Scenario. 1, 1 (April 2023), 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the rising ubiquity of AI in nearly all areas of our modern lives, the need for eXplainable AI (XAI) has been steadily recognised and addressed. While, in the past, generated explanations have often failed to consider the ultimate end-user of an AI model [10, 38], the trend of user-centred XAI is slowly gaining more traction. Although, in general, trust and acceptance are identified as crucial factors in Human Machine Interaction [37, 40], most of the existing studies that investigate trust or acceptance do not take XAI into account. This is probably due to the relative recency of XAI technologies. Furthermore, trust and acceptance with regard to XAI are often researched in safety-critical areas like medicine [29], where an erroneous system can lead to loss of life, ignoring areas that could also benefit from increased trust and acceptance, e.g. manufacturing. Thus, to improve human-AI cooperation, this work argues for the importance of trust and acceptance regarding XAI, especially to analyse the influence of user-centred XAI methods on trust and acceptance. We will present common practices in literature and their challenges and limitations. Finally, a scenario in the manufacturing sector illustrates our effort to address these challenges and explains the practical impact on,

Authors' addresses: Erika Puiutta, erika.puiutta@offis.de, OFFIS Institute of Informatics, Escherweg 2, Oldenburg, Niedersachsen, Germany, 26123; Larbi Abdenebaoui, larbi.abdenebaoui@offis.de, OFFIS Institute of Informatics, Escherweg 2, Oldenburg, Niedersachsen, Germany, 26123; Susanne Boll, susanne.boll@offis.de, OFFIS Institute of Informatics, Escherweg 2, Oldenburg, Niedersachsen, Germany, 26123.

e.g., work performance and satisfaction. This scenario will be implemented during the course of our research project 'Digitopias - DIGItal TechnOlogies for Participation and InterAction in Society'.

2 MANUFACTURING SCENARIO

Manufacturing companies already use a variety of automated systems that integrate AI elements. However, the corresponding machines are usually not interactive and can only be used with the support of extensive expert knowledge [22, 49]. Many researchers foresee humans and AI systems working more closely together in factories [1]. Humans can then guide these systems and teach them new skills as needed during the production process.

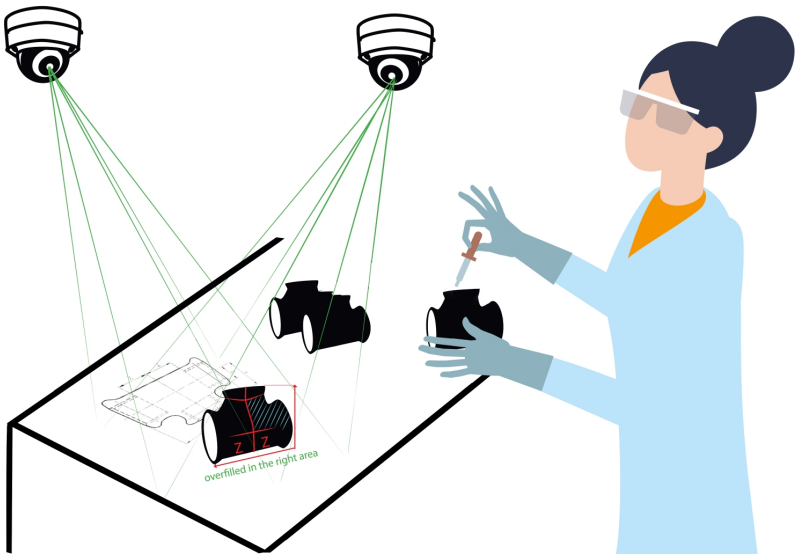

Fig. 1. An illustration of a manufacturing workspace. A worker performs a work step (e.g., filling various holes with glue), and the AI system recognises whether the action was successful. If requested, the AI system may use projected augmented reality to explain the reason for the assessment (in this example, an overfilled hole is shown). The worker can ask for further explanation so that, e.g., examples of images of different overfilled holes from the data set can be projected.

Focusing on the role of XAI in driving acceptance and trust, we are collaborating with two industry partners in our research project to define one use case on an assembly line for each of them that highlights interaction modalities. In the following, these two use cases are abstracted to form a generic scenario: A worker operates at a specific workstation and makes certain changes to given modules. The AI system supports the worker by assessing whether or not the changes made are correct. The worker can interact with the AI system and ask for explanations. As shown in figure 1, the explanation can be provided using projected Augmented Reality (AR). In this case, the exact area of the module responsible for the assessment is highlighted in red. More explanation can also be required; for example, images of similar examples from the data sets. Based on this explanation, the worker can determine if the assessment is reasonable and decide whether to follow the recommendation.

The next section provides an overview of the taxonomy used in related work. After that, we discuss the benefits and challenges of studying trust and acceptance in general and in our scenario in section 5.

3 TAXONOMY OF TRUST AND ACCEPTANCE

A common taxonomy is necessary to increase accuracy and establish the same understanding of the concepts researched. Unfortunately, especially psychological concepts like trust and acceptance can be hard to define and are used in a wide range of publications with slight variations in their definitions - if they are defined at all (for example, in a survey from Vereschak, Bailly and Caramiaux [48], only 26% of the reviewed papers have used definitions of trust). Below, we are broaching the most commonly used definitions and their implications for research in this field.

3.1 Trust

The definitions used are mostly adopted from social sciences and human-human trust and then transferred into human-machine trust [48]. Vereschak, Bailly and Caramiaux [48] have compiled the most frequently occurring aspects of trust:

- it is an *attitude*: An attitude describes 'summary evaluations of objects (e.g. oneself, other people, issues, etc) along a dimension ranging from positive to negative' [4, p. 611]. The important thing to note here is that an attitude is not necessarily followed by the corresponding behaviour [12, 48]; if the appropriate conditions are not met or if there are other interferences, an action might not be performed despite trust being present (e.g., a trusted doctor's treatment plan is not feasible due to economic reasons).
- it is present in situations of *vulnerability*: This describes 'uncertainty of the outcomes of a decision, with potential negative or undesirable consequences' [48, p. 9]. According to Vereschak, Bailly and Caramiaux [48], without vulnerability, there is no need for trust. This means that user studies have to be designed to create situations of vulnerability.
- it is present if the user has *positive expectations* considering the recommendations of an AI agent: As just stated, the user is in a position of vulnerability, i.e., negative consequences are entirely possible. However, as another key element of trust, positive expectations refer to the position that there will likely be no negative consequences.

All of these elements are present in the definition from Lee and See [26]:'[Trust is an] attitude that an agent will achieve an individual's goal in a situation characterised by uncertainty and vulnerability'. We will be using the term trust with this definition in mind.

3.2 Acceptance

The mere term acceptance can be defined as 'the behavioural intention or willingness to use, buy, or try a good or service' [19]. There exist several acceptance models that include different factors. The most commonly used one is the *Technology Acceptance Model* [TAM; 7, 8]. The TAM consists of two variables:

- *perceived usefulness*: 'the degree to which an individual believes that using a particular system would enhance his or her job performance' [8, p. 320].
- *perceived ease of use*: 'the degree to which a person believes that using a particular system would be free of effort' [8, p. 320].

The TAM is a flexible instrument and can be extended with additional variables [28]. Nevertheless, it has some limitations: First, it lacks generalisation to different industries [2, 18]. Second, some studies show that perceived ease

of use does not necessarily correlate with behavioural intention [30, 50]. In general, contrasting to what the TAM assumes, a high acceptance or intention to use does not necessarily predict actual behaviour [21] - the same with trust, as mentioned above. Important to note is also that, in their review, Kelly, Kaye, and Oviedo-Trespalacios [19] found few studies that studied actual behaviour. Thus, in the future, a stronger focus on the assessment of actual behaviour is necessary.

There are alternative acceptance models that bear mentioning; the *Unified Theory of Acceptance and Use of Technology* [UTAUT; 47] is a union of eight different acceptance models, including the TAM. It states that individual reactions to technology -based on factors such as performance expectancy and effort expectancy - will influence the intention to use it, and, thus, the actual use of technology. Ineed, it shows high predictive success concerning behavioural intentions [47], but, as already stated above, directly linking intention to use to actual use is problematic.

Another model that was developed not to measure the acceptance of technology, but of AI, specifically, is the *AI device use acceptance model* [AIDUA; 14]. It was developed because factors that drive technology acceptance are not the same as those driving AI acceptance, which diminishes the TAMs predictability [43]. The AIDUA model consists of three stages that measure AI device acceptance in service encounters:

- Primary appraisal: user's evaluation of the importance of AI device use
- secondary appraisal: user's evaluation of the perceived benefits and costs of using the AI device
- outcome: willingness and objection to the use of the AI device

This separation between willingness and objection is a novel take on the potential outcome; users might, in theory, be willing to use an AI device, but might forego the use in preference for actual human service. Thus, we will be using the term acceptance with this model in mind.

4 WHY TRUST AND ACCEPTANCE?

Although trust and acceptance has been the focus of research for many years, there is still inconclusive evidence when it comes to their relationship with XAI. Not many XAI user studies investigate trust AND acceptance together while also ensuring that the explanations are of value to the users. In the following, we argue for the importance of XAI user studies that investigate trust and acceptance together, both from a theoretical and practical perspective.

4.1 In General

It seems to be a common claim that explanations -or at least transparency- lead to user trust which, in turn, increases acceptance [11, 41]. Indeed, there is some evidence for this:

- Druce, Harradon, and Tittle [9] found a significant, albeit small, increase in trust and acceptance in an AI system with explanation vs. an AI system without explanation
- Liu, Chen, Kuo, and Lin [29] found that XAI increases technology trust and perceived value of an AI system
- Shin [41] showed that users trust an algorithm when it fulfils their expected level of FATE (fairness, accountability, transparency, and explainability)

However, there is also contrasting evidence, as in Papenmeier, Englebienne, and Seifert [34], who state that accuracy is more important than explainability and that explanations can even potentially harm trust, e.g., when faulty. Furthermore, they report a mismatch between observed and self-reported values of trust. Kenny, Ford, Quinn, and Keane [20] found that explanations influenced people's perception of error but not trust. Thus, it seems evident that more research has

to be conducted to investigate the influence of XAI on users. But not only that: we also claim that the factors to be investigated should be trust AND acceptance, simultaneously, to answer the following questions and more:

- With regards to human-AI cooperation, are trust and acceptance connected, and if so, how? Is it possible to have trust without acceptance, or vice versa? Is human-AI cooperation only optimal if both are present?
- What other factors do they influence (e.g. intent/willingness to use AI, actual use of AI, satisfaction with an AI system)?
- Are they related to other psychological concepts (e.g. work load) or physiological markers (e.g. heart rate)?

Furthermore, we focus on the practical implications of findings in this context when investigated in a professional environment, to be discussed in the next section.

4.2 In the Context of the Manufacturing Scenario

As an employee, one might not have much choice in whether or not an AI system is used. If a positive correlation between XAI and trust and acceptance does indeed exist, the question to be investigated is: If employees are required to use a particular AI system, do trust and acceptance matter? I.e., if the actual use of the AI system is not dependent on employees trusting and/or accepting it, what benefit is there when they do, and what drawback when they don't?

On the one hand, one could imagine that distrust leads to time wasted on ignoring or double-checking the recommendations of the AI system [46]. Additionally, the requirement to use an AI system that employees don't trust could lead to friction between employees and supervisors and create an unpleasant working environment. On the other hand, overtrusting - i.e., trusting a system beyond its capabilities [46] - might arguably be even worse; in Robinette, Howard, and Wagner [36], participants followed a robot in an emergency evacuation scenario, even when that robot was shown to have made faulty decisions in the past, and even after noticing the robot going in a wrong direction. Thus, overtrust can lead to risk or loss of life. But even outside of life-threatening situations, individuals relying on AI systems too much could fail to perform their tasks successfully when receiving wrong recommendations or forget to perform the necessary tasks manually ([44], in [46]).

In contrast, treating an AI system as intended, i.e., as a mere means of support, means to realistically assess when to trust - and, thus, follow its recommendations - and when to ignore it. As a consequence, an employee could be more confident in their actions and, consequently, increase their output. Thus, ensuring that appropriate levels of trust (or 'calibrated trust' [33]) are present is the first step to investigating possible beneficial effects on job satisfaction, performance, productivity, or other related concepts. The role of XAI in tackling these issues will be covered in the next section.

5 DESIGN OF XAI USER-STUDIES MEASURING TRUST & ACCEPTANCE

After highlighting the importance of measuring trust and acceptance in the context of XAI user studies, the next section deals with how research in this field is usually executed. The methods' limitations are highlighted, followed by our practical attempt to address them in our manufacturing scenario.

5.1 Common Practices and Challenges

XAI user studies, in general, entail both quantitative methods (e.g., surveys or questionnaires) and qualitative methods (e.g., interviews or field observations). They commonly involve letting users experience an AI model (e.g., one that plays a certain game [3]), followed by presenting one or more types of explanations for the model's behaviour and

then inquiring about the users' preference(s). This is often done with self-generated questions (e.g., 'Which of the rules do you find as more plausible?' [13]) or asking participants to explain or predict the model's actions [3]. Less often, questionnaires are used to evaluate the explanation quality or satisfaction [15, 16].

The same holds true for trust and acceptance: Here, too, psychometrics are still preferred - although the measurement through psychophysiological means (e.g., eye movements or neural measures) is gaining steady traction [24, 45]. Another possibility is the indirect assessment through the individual's behaviour (e.g.: How often do users follow an AI system's recommendation? How often do they intervene and take control?) [24].

Measuring trust and acceptance -especially with regard to XAI - is faced with a variety of challenges:

First, there is the lack of standardisation concerning the definition of terms used and the methods employed; as stated in section 3, some research claims to measure trust or acceptance without even defining what it is measuring. If there are definitions present, they can differ from publication to publication, impeding a proper comparison. The same holds true for the variety of measurements available: how can one compare, e.g., quantitative and qualitative or implicit and explicit measurements, i.e., methods that might differ wildly with regard to their validity or reliability? This is only exacerbated by the tendency to invent new questions instead of using an existing questionnaire.

Second, if preexisting questionnaires are used, these might not be appropriate (any more). Many of them are aimed at measuring trust or acceptance in technology [7, 47], automation [25], or human-robot interaction [32]. Few are explicitly aimed at AI (such as the AIDUA [14], see also subsection 3.2), let alone XAI, such as the *Trust Scale Recommended for XAI* [15]. This scale is a five-point Likert scale based on a combination of other scales (mostly the Cahour-Fourzy Scale [6], but also the Schaefer Scale [39], the Madsen-Gregor Scale [31], as well as one item from Jian, Bisantz, and Drury [17]). It asks users 'whether they are confident in the XAI system, whether the XAI system is predictable, reliable, efficient, and believable' [15, p. 49]. However, even if some methods are directly assessing trust in or acceptance of AI, many studies do not even ensure that the participants are aware of the technology being AI [19]. This is especially crucial since, for the general population, the difference between, e.g., autonomous robots and AI seems to be negligible [27].

But before investigating XAI's influence on trust and acceptance, one has to consider the third challenge: to observe an influence, one has to create the appropriate environment for the influence to be possible in the first place. This includes a) a situation of vulnerability for the participant, and b) a true distinction between situations with explanations vs without explanations. As stated in subsection 3.1, trust is only present in vulnerable situations, making reliance on the AI system necessary or at least preferable to the user. In addition, the situation with explanations has to be significantly different to the situation without explanations - which is only the case if the explanation is either useful enough to help, or bad enough to influence the relationship negatively. If the explanation is too simple and does not offer added value, it might just be ignored. This can only be ensured by directly investigating the quality of and satisfaction with explanations.

Not many methods exist that capture these assessments from the user's perspective; the System Causability Scale [SCS; 16] and the Explanation Satisfaction Scale [15] are two examples. The SCS is based on the System Usability Scale [SUS; 5], and, just as the SUS measures the quality of a system's user interface, the SCS measures the quality of explanations and explanation interfaces. Its goal is to 'quickly determine whether and to what extent an explainable user interface (human-AI interface), an explanation, or an explanation process itself is suitable for the intended purpose' [16, p. 4]. Alternatively, the Explanation Satisfaction Scale distills several key attributes of explanations - like understandability, sufficiency of detail, and accuracy - into a Likert scale. Explanation Satisfaction is defined as 'the degree to which users

feel that they understand the AI system or process being explained to them' [15, p. 5]. It is understood as an a posteriori judgement of explanations.

The last challenge is the fact that insights in this field of research are often disjoint from the design process [23]. Participatory design or co-creation are often only considered in late-stage development instead of accompanying the whole process. However, requirements engineering should be present early on in order to cater to end-users without time consuming and economically costly retroactive changes. Undoubtedly, additional challenges are present but not included here due to brevity.

5.2 Tackling the Challenges in a Manufacturing Scenario

Following here is our attempt to deal with these challenges in the context of a real-world application, i.e., our aforementioned manufacturing scenario (see section 2). It involves two pilot studies to uncover requirements concerning the shape of explanations, and one main study in which we investigate the influence of XAI on trust, acceptance, and performance. Being in a professional environment and needing to make quick (potentially wrong) decisions, we can assume a situation of vulnerability due to the fact that failure to perform adequately is linked with a risk of reprimand. Then, an AI system that supports the user in a projected AR-environment is developed, and an XAI method (to be determined; likely a saliency map [42] or similar) is applied. Based on users' preferences in past literature, a tentative, small sample of helpful explanations is designed, with different scopes (global vs local; for details on this distinction see Puiutta and Veith [35]) and different formats of explanation (e.g., text vs images vs symbols). In the next step, the first pilot study is conducted to evaluate the preferred explanations' explanation quality and satisfaction with the help of the aforementioned SCS and Explanation Satisfaction Scale (see section subsection 5.1). We will also follow a structured process of requirements engineering to gather free feedback on what changes the users might want. This is done as a field study, i.e., in the actual place of employment. The feedback and results are then applied to the explanations, and the explanation quality and satisfaction are reassessed in the second pilot study. After ensuring the usefulness of the explanations, the main study with at least 30 participants is performed. Here, participants will work with an AI system similar to what can be seen in Figure 1. They will perform the same task in two conditions: with and without explanations for the behaviour of the AI system. The order of conditions is randomised in order to ensure that an increase in trust and/or acceptance is, in fact, due to the explanations and not simply due to higher experience working with the system. In order to encapsulate trust and acceptance in AI (instead of, e.g., technology) with a focus on XAI, trust will be measured through the aforementioned Trust Scale Recommended for XAI, acceptance through the AIDUA (see subsection 5.1). Performance will be measured through key performance indicators, e.g. number of orders processed or mean time spent on one order. Explanation quality and satisfaction will, again, be measured through the SCS and the Explanation Satisfaction Scale.

6 CONCLUSION

In this paper, we have touched on the common practices with regards to XAI user studies and studies that measure trust and acceptance. We highlighted the challenges and limitations and suggested how to tackle them with a practical example in a manufacturing scenario. It has been shown that studies often lack standardised definitions and approaches. Furthermore, existing methods might be(come) obsolete due to the rapid adoption of new technology and AI into our daily lives, making the underlying models and factors inapplicable. Our approach is designed to research, first, which types of explanations are preferred by end-users and, second, what influence these explanations have on trust, acceptance, and performance. This could shed light on a theoretical 'empowerment effect' for which explanations can

play a key role: Explaining the decisions of the system enables the user to decide whether they are reasonable or not. If they are not, the user becomes aware of the system's limitations and can even potentially correct it. This ensures a valuable and successful human-AI relationship that could lead to higher job satisfaction, productivity and performance.

ACKNOWLEDGMENTS

This work was part of the research project Digitopias, funded under the German funding program 'SPRUNG - SPitzenforSchUNG in Niedersachsen'. We thank all our colleagues for their insightful comments and support.

REFERENCES

[1] Amr Adel. 2022. Future of industry 5.0 in society: human-centric solutions, challenges and prospective research areas. *Journal of Cloud Computing* 11, 1 (Sept. 2022). https://doi.org/10.1186/s13677-022-00314-5

[2] Kashan Ali and Kim Freimann. 2021. Applying the Technology Acceptance Model to AI decisions in the Swedish Telecom Industry.

[3] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Alan Fern, and Margaret Burnett. 2019. Explaining Reinforcement Learning to Mere Mortals: An Empirical Study. arXiv:arXiv:1903.09708

[4] Gerd Bohner and Nina Dickel. 2011. Attitudes and Attitude Change. *Annual Review of Psychology* 62, 1 (Jan. 2011), 391–417. https://doi.org/10.1146/annurev.psych.121208.131609

[5] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[6] Béatrice Cahour and Jean-François Forzy. 2009. Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Science* 47, 9 (Nov. 2009), 1260–1270. https://doi.org/10.1016/j.ssci.2009.03.015

[7] Fred D Davis. 1985. *A technology acceptance model for empirically testing new end-user information systems: Theory and results.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[8] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3 (Sept. 1989), 319. https://doi.org/10.2307/249008

[9] Jeff Druce, Michael Harradon, and James Tittle. 2021. Explainable Artificial Intelligence (XAI) for Increasing User Trust in Deep Reinforcement Learning Driven Autonomous Systems. arXiv:arXiv:2106.03775

[10] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (Dec. 2019), 68–77. https://doi.org/10.1145/3359786

[11] Upol Ehsan and Mark Riedl. 2019. On Design and Evaluation of Human-centered Explainable AI systems [position paper].

[12] Rino Falcone and Cristiano Castelfranchi. 2009. Socio-cognitive model of trust. In *Human Computer Interaction: Concepts, Methodologies, Tools, and Applications.* IGI Global, 2316–2323.

[13] Johannes Fürnkranz, Tomáš Kliegr, and Heiko Paulheim. 2019. On cognitive preferences and the plausibility of rule-based models. *Machine Learning* 109, 4 (Dec. 2019), 853–898. https://doi.org/10.1007/s10994-019-05856-5

[14] Dogan Gursoy, Oscar Hengxuan Chi, Lu Lu, and Robin Nunkoo. 2019. Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management* 49 (Dec. 2019), 157–169. https://doi.org/10.1016/j.ijinfomgt.2019.03.008

[15] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. arXiv:arXiv:1812.04608

[16] Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS). *KI - Künstliche Intelligenz* 34, 2 (Jan. 2020), 193–198. https://doi.org/10.1007/s13218-020-00636-z

[17] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (March 2000), 53–71. https://doi.org/10.1207/s15327566ijce0401_04

[18] Sage Kelly, Sherrie-Anne Kaye, and Oscar Oviedo-Trespalacios. 2022. A Multi-Industry Analysis of the Future Use of AI Chatbots. *Human Behavior and Emerging Technologies* 2022 (Nov. 2022), 1–14. https://doi.org/10.1155/2022/2552099

[19] Sage Kelly, Sherrie-Anne Kaye, and Oscar Oviedo-Trespalacios. 2023. What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics* 77 (Feb. 2023), 101925. https://doi.org/10.1016/j.tele.2022.101925

[20] Eoin M. Kenny, Courtney Ford, Molly Quinn, and Mark T. Keane. 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence* 294 (May 2021), 103459. https://doi.org/10.1016/j.artint.2021.103459

[21] J. Keung, R. Jeffery, and B. Kitchenham. 2004. The challenge of introducing a new software cost estimation technology into a small software organisation. In *2004 Australian Software Engineering Conference. Proceedings.* IEEE. https://doi.org/10.1109/aswec.2004.1290457

[22] Steffen Kinkel, Marco Baumgartner, and Enrica Cherubini. 2022. Prerequisites for the adoption of AI technologies in manufacturing – Evidence from a worldwide sample of manufacturing companies. *Technovation* 110 (Feb. 2022), 102375. https://doi.org/10.1016/j.technovation.2021.102375

[23] Marion Koelle, Swamy Ananthanarayan, and Susanne Boll. 2020. Social Acceptability in HCI: A Survey of Methods, Measures, and Design Strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* ACM. https://doi.org/10.1145/3313831.3376162

[24] Spencer C. Kohn, Ewart J. de Visser, Eva Wiese, Yi-Ching Lee, and Tyler H. Shaw. 2021. Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology* 12 (Oct. 2021). https://doi.org/10.3389/fpsyg.2021.604977

[25] Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation.

[26] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

[27] Yuhua Liang and Seungcheol Austin Lee. 2017. Fear of Autonomous Robots and Artificial Intelligence: Evidence from National Representative Data with Probability Sampling. *International Journal of Social Robotics* 9, 3 (March 2017), 379–384. https://doi.org/10.1007/s12369-017-0401-3

[28] Chen-Yang Lin and Ni Xu. 2021. Extended TAM model to explore the factors that affect intention to use AI robotic architects for architectural design. *Technology Analysis &amp Strategic Management* 34, 3 (March 2021), 349–362. https://doi.org/10.1080/09537325.2021.1900808

[29] Chung-Feng Liu, Zhih-Cherng Chen, Szu-Chen Kuo, and Tzu-Chi Lin. 2022. Does AI explainability affect physicians' intention to use AI? *International Journal of Medical Informatics* 168 (Dec. 2022), 104884. https://doi.org/10.1016/j.ijmedinf.2022.104884

[30] Zhan Liu, Jialu Shan, and Yves Pigneur. 2016. The role of personalized services and control: An empirical evaluation of privacy calculus and technology acceptance model in the mobile context. *Journal of Information Privacy and Security* 12, 3 (July 2016), 123–144. https://doi.org/10.1080/15536548.2016.1206757

[31] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.

[32] Bertram F Malle and Daniel Ullman. 2021. A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction*. Elsevier, 3–25.

[33] Patricia L. McDermott and Ronna N. ten Brink. 2019. Practical Guidance for Evaluating Calibrated Trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 63, 1 (Nov. 2019), 362–366. https://doi.org/10.1177/1071181319631379

[34] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. https://doi.org/10.48550/ARXIV.1907.12652

[35] Erika Puiutta and Eric M. S. P. Veith. 2020. Explainable Reinforcement Learning: A Survey. In *Lecture Notes in Computer Science*. Springer International Publishing, 77–95. https://doi.org/10.1007/978-3-030-57321-8_5

[36] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. 2017. Effect of Robot Performance on Human–Robot Trust in Time-Critical Situations. *IEEE Transactions on Human-Machine Systems* 47, 4 (Aug. 2017), 425–436. https://doi.org/10.1109/thms.2017.2648849

[37] Alessandra Rossi, Patrick Holthaus, Giulia Perugia, Sílvia Moros, and Marcus Scheunemann. 2021. Trust, Acceptance and Social Cues in Human–Robot Interaction (SCRITA). *International Journal of Social Robotics* 13, 8 (Dec. 2021), 1833–1834. https://doi.org/10.1007/s12369-021-00844-z

[38] Wojciech Samek and Klaus-Robert Müller. 2019. Towards Explainable Artificial Intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer International Publishing, 5–22. https://doi.org/10.1007/978-3-030-28954-6_1

[39] Kristin Schaefer. 2013. The perception and measurement of human-robot trust. (2013).

[40] Felix Schoeller, Mark Miller, Roy Salomon, and Karl J. Friston. 2021. Trust as Extended Control: Human-Machine Interactions as Active Inference. *Frontiers in Systems Neuroscience* 15 (Oct. 2021). https://doi.org/10.3389/fnsys.2021.669810

[41] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (Feb. 2021), 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

[42] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:arXiv:1312.6034

[43] Kwonsang Sohn and Ohbyung Kwon. 2020. Technology acceptance theories and factors influencing artificial Intelligence-based intelligent products. *Telematics and Informatics* 47 (April 2020), 101324. https://doi.org/10.1016/j.tele.2019.101324

[44] P Sparaco. 1995. Airbus seeks to keep pilot, new technology in harmony. *Aviation Week & Space Technology* 142, 5 (1995), 62–63.

[45] Katerina Tzafilkou and Nicolaos Protogeros. 2017. Diagnosing user perception and acceptance using eye tracking in web-based end-user development. *Computers in Human Behavior* 72 (July 2017), 23–37. https://doi.org/10.1016/j.chb.2017.02.035

[46] Daniel Ullrich, Andreas Butz, and Sarah Diefenbach. 2021. The Development of Overtrust: An Empirical Simulation and Psychological Analysis in the Context of Human–Robot Interaction. *Frontiers in Robotics and AI* 8 (April 2021). https://doi.org/10.3389/frobt.2021.554578

[47] Venkatesh, Morris, Davis, and Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* 27, 3 (2003), 425. https://doi.org/10.2307/30036540

[48] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–39. https://doi.org/10.1145/3476068

[49] YuanBin Wang, Pai Zheng, Tao Peng, HuaYong Yang, and Jun Zou. 2020. Smart additive manufacturing: Current artificial intelligence-enabled methods and future perspectives. *Science China Technological Sciences* 63, 9 (May 2020), 1600–1611. https://doi.org/10.1007/s11431-020-1581-2

[50] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Does stated accuracy affect trust in machine learning algorithms.

Received 23 February 2023; revised 6 April 2023; accepted 10 March 2023