The Human Side of Service Engineering, Vol. 108, 2023, 57–66
https://doi.org/10.54941/ahfe1003107

AI-Based Services - Design Principles to Meet the Requirements of a Trustworthy AI

Janika Kutz[1,2], Jens Neuhüttler[1], Jan Spilski[2], and Thomas Lachmann[2,3]

[1] Fraunhofer Institute for Industrial Engineering IAO; Stuttgart, 70569, Germany
[2] Center for Cognitive Science, University of Kaiserslautern-Landau; Kaiserslautern, 67663, Germany
[3] Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija; Madrid, Spain

ABSTRACT

The development of Human-Centered and Trustworthy AI-based services has recently attracted increased attention in politics and science. Even though that technical advances have received many of the attention lately, ethical considerations are becoming more and more important. One of the most valuable publications in this area is the "Ethics Guidelines for Trustworthy AI" of the European Commission (EC). One approach to assist developers in implementing these requirements during the development process is to provide design guidelines. The aim of this paper is to identify which action-oriented design principles can be applied to satisfy the requirements for Trustworthy AI. For this purpose, the design principles published by major providers of commercial AI-based services were contrasted with the seven requirements of the EC. The results indicate that some design principles can be used to meet the requirements of Trustworthy AI. At the same time, however, it becomes clear that work on Ethical AI should be extended by aspects related to Human-AI Interaction and service process quality.

Keywords: AI-based services, Human-centered AI, Trustworthy AI, Design principles

INTRODUCTION

Artificial Intelligence (AI)-based services are increasingly used in both private and professional life. Their use offers many opportunities and benefits, but they can also cause harm (Xu and Dainoff, 2021). Examples of this are stored in the *AI Incident Database* (McGregor, 2021). The reasons for such failures can be biased data as well as complex and non-transparent AI systems (Kaur et al. 2023). As a result, the development and operation of AI-based services are associated with challenges and concerns. Two often mentioned challenges are a lack of technology acceptance and a lack of trust in AI-based services (Kaur et al. 2023; Kutz et al. 2022). Creating Ethically and Trustworthy AI as well as Human-Centered AI (HCAI) has therefore recently received more attention from academia and politics (Xu, 2019). Globally, politicians are considering how to address the challenges caused by the advancement of AI. One of the most valuable publications in this area is the "Ethics Guidelines for Trustworthy AI" of the European Commission (EC), which defines seven requirements for Trustworthy AI (High-Level Expert Group on Artificial Intelligence 2019a). One approach to assist developers in implementing these requirements during the development process is to provide action-oriented design principles. The aim of this paper is to identify which practical design principles can be applied to satisfy the seven requirements for Trustworthy AI.

ETHICAL AND TRUSTWORTHY AI

There are an unmanageable number of guidelines and papers relevant to designing Ethical and Trustworthy AI-based systems, making it difficult for developers and researchers to draw the right conclusions. The best strategy is to limit the focus to review papers that already provide a systematic evaluation and summary of existing work (Hagendorff, 2020; Jobin et al. 2019). Jobin et al., for instance, analyzed 84 ethics guidelines for AI and identified five ethical principles that are globally included (transparency, justice and fairness, non-maleficence, responsibility, and privacy). One of the most comprehensive works on the subject is provided by the EC. In 2019, the High-Level Expert Group on Artificial Intelligence published the "Ethics Guidelines for Trustworthy AI". This guideline's aim is to encourage the development of Trustworthy AI in a human-centered approach. To fulfil the four ethical principles (respect for human autonomy, prevention of harm, fairness, explainability), seven key requirements are defined (see Table 1). Nevertheless, the guidelines are set at a high level and therefore serve more as guidance and less as actual assistance for designing trustworthy AI-based services.

Since the EU is a leading entity in the field of Trustworthy AI and a first regulatory framework is expected with the EU-AI Act (European Commission 2021), we focus in this paper on the requirements formulated by the EC for Trustworthy AI.

Also considering the requirements of the EC is the AI test-guideline of Fraunhofer IAIS (2021). This guideline provides a framework for assessing trustworthiness in a structured way, and at the same time provides guidance for developers to implement these requirements. However, the focus of the guideline is on the verification and not on the provision of design principles for the development.

In 2022, Kaur et al. published a review about Trustworthy AI. They propose an overview about methods that can be used to address the requirements of the EC. Moreover, they argue that the guidelines should be added by a principle focusing on the acceptance of AI. Furthermore, they mention that "human involvement is essential in this changing era of AI…" (p. 39:28). One approach to address this is Human-Centered AI (HCAI).

Table 1. Ethics guidelines for trustworthy AI - key requirements (high-level expert group on artificial intelligence 2019b).

Requirement	Description
Human agency and oversight	"AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches"[1]
Technical robustness and safety	"AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented." [1]
Privacy and data governance	"Privacy and data governance: besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimised access to data." [1]
Transparency	"Transparency: the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations." [1]
Diversity, non-discrimination and fairness	"Unfair bias must be avoided, as it could could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle." [1]
Environmental and societal well-being	"AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered."[1]
Accountability	"Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate an accessible redress should be ensured." [1]

Note. 1 = (High-Level Expert Group on Artificial Intelligence 2019a)

HUMAN-CENTERED AI

Lately, HCAI get more popular in research. "HCAI focuses on amplifying, augmenting, and enhancing human performance in ways that make systems reliable, safe, and trustworthy." (Shneiderman, 2020, 26:2) Instead of replacing people, HCAI seeks to put people at the center of AI-based services. One way of improving HCAI design is to use design principles, patterns, and guidelines throughout the development process. Of course, policymakers' guidelines provide direction for HCAI design, but only on a high level, as previously stated. Technology companies, for example, published more action-oriented guidelines. The People + AI Guidebook from Google's People + Research Center contains 23 design patterns for developing HCAI. The patterns are sorted along critical questions in the development process and more in-depth information can be found on six thematic categories (User Needs and Defining Success, Data Collection and Evaluation, Mental Models, Explainability and Trust, Feedback and Control, Errors and Graceful Failure; Google PAIR 2019). Another comprehensive work containing action-oriented guidelines is published by Microsoft. Based on a literature review, they identified 18 design guidelines for Human-AI Interaction (Amershi et al. 2019). A web application contains detailed descriptions and practical recommendations for implementation (Microsoft 2021).

PURPOSE OF RESEARCH

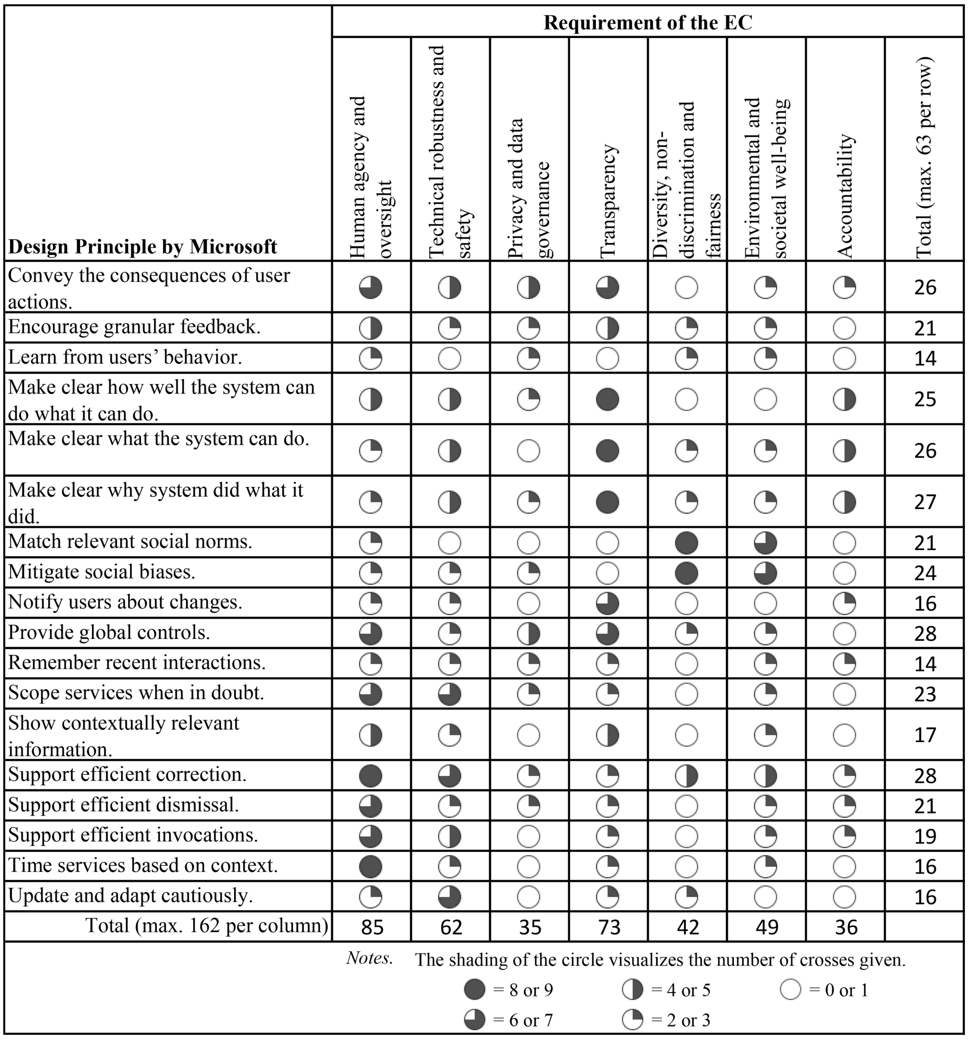
Ethical and Human-Centered AI design go hand in hand, i.e., they serve the same goal of developing AI-based services that are reliable, safe, and trustworthy. While the ethical guidelines published by government organizations are at a high level, the HCAI design principles published by technology companies are more specific and provide guidance for action. To implement the seven requirements in practice, AI developers, management, and other stakeholders need action-oriented design principles or design patterns (Shneiderman, 2020). The aim of this paper is to examine to what extent the application of the design principles published by Microsoft (Microsoft 2021) and Google (Google PAIR 2019) leads to a fulfillment of the requirements of Trustworthy AI set by the EC. For this purpose, the following research question is answered: Which design principles for AI-based services can be identified to fulfil the guidelines of a Trustworthy AI according to the EC? In addition, aspects are to be identified that are considered in the HCAI design but are of minor importance in the debate on ethical design.

METHOD

To answer the research question, the action-oriented design principles (18 guidelines for Human-AI Interaction by Microsoft, 23 design patterns of the People + AI Guidebook by Google) are mapped to the requirements of the EC by nine independent raters. For this, each participant got a matrix in which the seven requirements were entered in the columns and the design principles in the rows. A cross was used to indicate whether a design principle contributes to the satisfaction of the requirement. To ensure that all participants have the same understanding of the design principles and requirements, explanations were provided in the matrix. Participants required an average of 90 minutes to complete the matrix. All participants regularly deal with the development and implementation of AI-based services in their daily work or conduct research in the field of data-based services. However, participants from different disciplines were selected to fill out the matrix, e.g., AI engineers, digital developers, psychologists, or researchers in information systems. To evaluate the results, the sum was calculated for each cell. Moreover, each row and each column were summed up. A visualization of the frequencies via pie charts was chosen to present the results.

RESULTS

The evaluation of the matrices shows that the requirements "*Human agency and oversight*", "*Transparency*" and "*Technical robustness and safety*" are most frequently addressed by the application of the action-oriented design principles. The requirements "*Privacy and data governance*", "*Diversity, non-discrimination and fairness*" and "*Accountability*" are less addressed by the guidelines for Human-AI interaction (Microsoft 2022; see Figure 1). For the design patterns from the People + AI Guidebook, this is also evident for the requirements "*Diversity, non-discrimination and fairness*" and



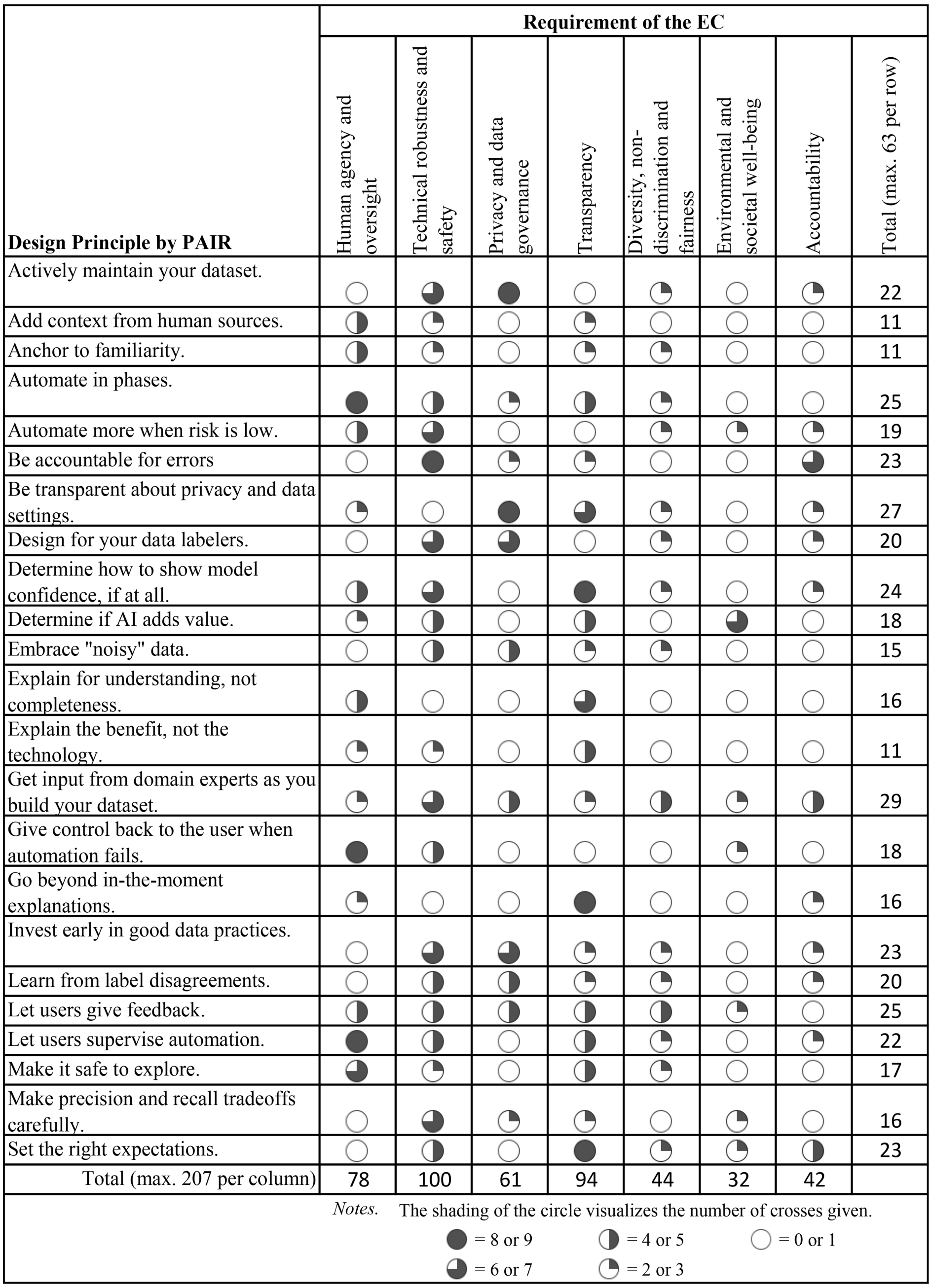
	Requirement of the EC							
Design Principle by Microsoft	Human agency and oversight	Technical robustness and safety	Privacy and data governance	Transparency	Diversity, non-discrimination and fairness	Environmental and societal well-being	Accountability	Total (max. 63 per row)
Convey the consequences of user actions.	◕	◑	◑	◕	○	◔	◔	26
Encourage granular feedback.	◑	◔	◔	◑	◔	◔	○	21
Learn from users' behavior.	◔	○	◔	○	◔	◔	○	14
Make clear how well the system can do what it can do.	◑	◑	◔	●	○	○	◑	25
Make clear what the system can do.	◔	◑	○	●	◔	◔	◑	26
Make clear why system did what it did.	◔	◑	◔	●	◔	◔	◑	27
Match relevant social norms.	◔	○	○	○	●	◕	○	21
Mitigate social biases.	◔	◔	◔	○	●	◕	○	24
Notify users about changes.	◔	◔	○	◕	○	○	◔	16
Provide global controls.	◕	◔	◑	◕	◔	◔	○	28
Remember recent interactions.	◔	◔	◔	◔	○	◔	◔	14
Scope services when in doubt.	◕	◕	◔	◔	○	◔	○	23
Show contextually relevant information.	◑	◔	○	◑	○	◔	○	17
Support efficient correction.	●	◕	◔	◔	◑	◑	◔	28
Support efficient dismissal.	◕	◔	◔	◔	○	◔	◔	21
Support efficient invocations.	◕	◑	○	◔	○	◔	◔	19
Time services based on context.	●	◔	○	◔	○	◔	○	16
Update and adapt cautiously.	◔	◕	○	◔	◔	○	○	16
Total (max. 162 per column)	85	62	35	73	42	49	36	

Notes. The shading of the circle visualizes the number of crosses given.

● = 8 or 9 ◑ = 4 or 5 ○ = 0 or 1

◕ = 6 or 7 ◔ = 2 or 3

Figure 1: Mapping of the guidelines for human-AI interaction by Microsoft (Microsoft 2021) and the requirements for trustworthy AI by the EC (high-level expert group on artificial intelligence 2019b).



	Requirement of the EC							
Design Principle by PAIR	Human agency and oversight	Technical robustness and safety	Privacy and data governance	Transparency	Diversity, non-discrimination and fairness	Environmental and societal well-being	Accountability	Total (max. 63 per row)
Actively maintain your dataset.	○	◕	●	○	◔	○	◔	22
Add context from human sources.	◑	◔	○	◔	○	○	○	11
Anchor to familiarity.	◑	◔	○	◔	◔	○	○	11
Automate in phases.	●	◑	◔	◑	◔	○	○	25
Automate more when risk is low.	◑	◕	○	○	◔	◔	◔	19
Be accountable for errors	○	●	◔	◔	○	○	◕	23
Be transparent about privacy and data settings.	◔	○	●	◕	◔	○	◔	27
Design for your data labelers.	○	◕	◕	○	◔	○	◔	20
Determine how to show model confidence, if at all.	◑	◕	○	●	◔	○	◔	24
Determine if AI adds value.	◔	◑	○	◑	○	◕	○	18
Embrace "noisy" data.	○	◑	◑	◔	◔	○	○	15
Explain for understanding, not completeness.	◑	○	○	◕	○	○	○	16
Explain the benefit, not the technology.	◔	◔	○	◑	○	○	○	11
Get input from domain experts as you build your dataset.	◔	◕	◑	◔	◑	◔	◑	29
Give control back to the user when automation fails.	●	◑	○	○	○	◔	○	18
Go beyond in-the-moment explanations.	◔	○	○	●	○	○	◔	16
Invest early in good data practices.	○	◕	◕	◔	◔	○	◔	23
Learn from label disagreements.	○	◑	◑	◔	◔	○	◔	20
Let users give feedback.	◑	◑	◑	◑	◑	◔	○	25
Let users supervise automation.	●	◑	○	◑	◔	○	◔	22
Make it safe to explore.	◕	◔	○	◑	◔	○	○	17
Make precision and recall tradeoffs carefully.	○	◕	◔	◔	○	◔	○	16
Set the right expectations.	○	◑	○	●	◔	◔	◑	23
Total (max. 207 per column)	78	100	61	94	44	32	42	

Notes. The shading of the circle visualizes the number of crosses given.
● = 8 or 9 ◑ = 4 or 5 ○ = 0 or 1
◕ = 6 or 7 ◔ = 2 or 3

Figure 2: Mapping of the design patterns of the people + AI guidebook (Google PAIR 2019) and the requirements for trustworthy AI by the EC (high-level expert group on artificial intelligence 2019b).

"*Accountability*" but also for "*Environmental and social well-being*" (see Figure 2).

The results in Figure 1 also show that principles such as "*Convey the consequences of user actions.*", "*Make clear how well the system can do what it can do.*" and "*Provide global controls.*" address multiple requirements. Principles like "*Learn from users' behavior.*" and "*Remember recent interactions.*" are seldom linked to the requirements of the EC.

Design patterns of the Google + AI Guidebook that address more than one requirement are, e.g., "*Automate in phases.*" "*Be transparent about privacy and data settings.*" and "*Get input from your domain experts as you build your dataset.*". Few requirements are addressed with patterns such as "*Add context from human sources.*", "*Anchor to familiarity.*" and "*Explain the benefit, not the technology.*"

DISCUSSION AND CONCLUSION

The aim of this paper is to identify action-oriented design principles that can be applied to satisfy the seven requirements for Trustworthy AI by the EC (European Commission 2021). For this purpose, the design principles published by Google (Google PAIR 2019) and Microsoft (2021) were contrasted with the seven requirements of the EC. The application of the design principles by Microsoft, as well as Google, can be used in particular to fulfil the requirements for "*Human agency and oversight*", "*Technical robustness and safety*" and "*Transparency*". The patterns from the People + AI Guidebook also address the requirement after "Privacy and data governance". To conclude, the design principles are partly suitable for designing Trustworthy AI according to the understanding of the EC. However, they are not sufficient to meet all requirements. The following requirements are less addressed by the design principles: "*Diversity, non-discrimination and fairness*", "*Environmental and societal well-being*" and "*Accountability*". Further research should consider this gap and more detailed action-oriented design principles should be formulated. This gap was identified by Kaur et al. (2022) as well: "However, there is still an implementation gap between the research and practice. So, there is a need to establish policies and standards to enforce these guidelines and existing laws into practice." (p. 39:2). With the publication of the EU-AI Act (European Commission 2021), the implementation of the requirements for Trustworthy AI will gain additional relevance and thus also the range of practical recommendations for action.

According to the results, some of the principles can help to meet multiple requirements. Nevertheless, to classify the results, it must be considered that one principle cannot fulfil all requirements at the same time. On the one hand, this would hardly be possible due to the complexity and multifaceted requirements, and on the other hand, the design patterns would lose their specific orientation. It should also be noted that contradictions of objectives can arise when fulfilling the requirements. This is a restriction that was also made clear in the guidelines for testing Trustworthy AI systems by the Fraunhofer IAIS (2021). The same is to be expected when implementing the design principles. Consideration of how to deal with potentially conflicting goals was beyond the scope of this work. Further research needs to be done to define selection criteria, as well as methods for applying these criteria to determine the key requirements and principles for a particular AI-based service. For example, depending on the criticality of the AI-based service, as well as its field of application, the interaction strength between humans and AI, or the selected AI technology itself, differences in the relevance of the implementation of design principles may arise (High-Level Expert Group on Artificial Intelligence 2019a; Kaur et al. 2023).

HCAI aims to put people at the center of AI-based services (Shneiderman, 2022). Looking at the guidelines for Human-AI Interaction formulated by Microsoft as well as the People + AI Guidebook by Google, it becomes clear that some formulated principles do not match the requirements of the EC. This is particularly evident for those principles that focus on Human-AI interaction. As "HCAI focuses on amplifying, augmenting, and enhancing human performance…" (Shneiderman, 2020, 26:2) these principles are not minor important in the discussion about AI-based services that are perceived as reliable, safe, and trustworthy. More research should be conducted to analyze how the two aspects of Human-AI interaction and Ethical AI might be considered together.

In the future, an important addition to the existing design principles could be the combination with insights and approaches from the discipline of service science and engineering. On the one hand, this discipline deals with the design of new services and takes a holistic view of the utilisation process as well as the consideration of contextual factors during utilisation. So far, such factors have only been marginally considered in the existing design principles. On the other hand, new methods are currently being researched on how the perception of quality (including the perceived trustworthiness, safety and usefulness during the usage process) can already be tested during development in order to prevent undesirable developments as early as possible (Neuhüttler et al. 2022). The approaches there do not deal with the objectively assessable technical implementation, but with the perception of users.

Our research has some limitations. The results show differences in the evaluation of matching. One possible reason for this could be that the participants might understand the requirements and design principles differently. It was not possible to verify that everyone understood the requirements and principles equally well, even though detailed descriptions for each requirement and principle attempted to ensure this. To identify reasons for the different matchings, further qualitative studies should be conducted, for example through focus group discussions. In addition, the study should be replicated with a larger sample to validate the results. Another limitation of this study is that it only included the design principles of two sources. As these works are, to our knowledge, the most comprehensive available from a practical point of view, we have nevertheless limited the scope of our study to them. An extension of the literature and desktop research, with particular emphasis on the requirements not covered by the design principles considered, could contribute to a more complete picture. In order to provide actionable principles for each requirement, a comprehensive framework should be established in the future.

REFERENCES

Amershi, Saleema/Weld, Dan/Vorvoreanu, Mihaela/Fourney, Adam/Nushi, Besmira-/Collisson, Penny/Suh, Jina/Iqbal, Shamsi/Bennett, Paul N./Inkpen, Kori/Teevan, Jaime/Kikin-Gil, Ruth/Horvitz, Eric (2019). Guidelines for Human-AI Interaction. In: Stephen Brewster/Geraldine Fitzpatrick/Anna Cox et al. (Eds.). Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI'19: CHI Conference on Human Factors in Computing Systems, Glasgow Scotland Uk, 04 05 2019 09 05 2019. New York, NY, USA, ACM, 1–13.

European Commission (2021). Proposal for a Regulation of the European Parliament and the Council. Laying down harmonised rules on artificial intelligence. Brussels. Available online at file:///C:/Users/kutz/Downloads/regulation_ai_875509BF-C386-0D30-2CB7E56A798BA4EA_75788.pdf (accessed 2/9/2022).

Google PAIR (2019). People + AI Guidebook. Designing human-centered AI products. Available online at https://pair.withgoogle.com/guidebook/ (accessed 04.11.22).

Hagendorff, Thilo (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines 30 (1), 99–120. https://doi.org/10.1007/s11023-020-09517-8.

High-Level Expert Group on Artificial Intelligence (2019a). Ethics guidelines for trustworthy AI. Available online at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed 2/9/2023).

High-Level Expert Group on Artificial Intelligence (2019b). Ethics Guidelines for Trustworthy AI. Brussels. Available online at file:///C:/Users/kutz/Downloads/ai_hleg_ethics_guidelines_for_trustworthy_ai-en_87F84A41-A6E8-F38C-BFF661481B40077B_60419.pdf.

Jobin, Anna/Ienca, Marcello/Vayena, Effy (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence 1 (9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Kaur, Davinder/Uslu, Suleyman/Rittichier, Kaley J./Durresi, Arjan (2023). Trustworthy Artificial Intelligence: A Review. ACM Computing Surveys 55 (2), 1–38. https://doi.org/10.1145/3491209.

Kutz, Janika/Neuhüttler, Jens/Spilski, Jan/Lachmann, Thomas (2022). Implementation of AI Technologies in manufacturing - success factors and challenges. In: The Human Side of Service Engineering, 13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022), July 24-28, 2022. AHFE International.

McGregor, Sean (2021). Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. Proceedings of the AAAI Conference on Artificial Intelligence 35 (17), 15458–15463. https://doi.org/10.1609/aaai.v35i17.17817.

Microsoft (2021). Guidelines for Human-AI Interaction. Microsoft HAX Toolkit. Available online at https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/ (accessed 2/8/2023).

Neuhüttler, Jens/Hermann, Sibylle/Ganz, Walter/Spath, Dieter/Mark, Riccarda (2022). Quality Based Testing of AI-based Smart Services: The Example of Stuttgart Airport. In: 2022 Portland International Conference on Management of Engineering and Technology (PICMET), 2022 Portland International Conference on Management of Engineering and Technology (PICMET), Portland, OR, USA, 07.08.2022 - 11.08.2022. IEEE, 1–10.

Poretschkin, Maximilian/Schmitz, Anna/Akila, Maram/Adilova, Linara/Becker, Daniel/Cremers, Armin B./Hecker, Dirk/Houben, Sebastian/Mock, Michael/Rosenzweig, Julia/Sicking, Joachim/Schulz, Elena/Voß, Angelika/Wrobel, Stefan (2021). Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz (KI-Prüfkatalog). Fraunhofer IAIS. Sankt Augustin. https://doi.org/10.24406/PUBLICA-FHG-301361.

Shneiderman, Ben (2020). Bridging the Gap Between Ethics and Practice:. Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. ACM Transactions on Interactive Intelligent Systems 10 (4), 1–31. https://doi.org/10.1145/3419764.

Shneiderman, Ben (2022). Human-centered AI. Oxford, United Kingdom/New York, NY, Oxford University Press.

Xu, Wei (2019). Toward human-centered AI: A Perspective from Human-Computer-Interaction. Interactions 26 (4), 42–46. https://doi.org/10.1145/3328485.

Xu, Wei/Dainoff, Marvin (2021). Enabling human-centered AI: A new junction and shared journey between AI and HCI communities. Available online at https://arxiv.org/pdf/2111.08460v3.