10-9-2023

What is the Minimum to Trust AI?—A Requirement Analysis for (Generative) AI-based Texts

Christoph Tomitza
Julius-Maximilians-Universität Würzburg, Germany, christoph.tomitza@uni-wuerzburg.de

Myriam Schaschek
Julius-Maximilians-Universität Würzburg, Germany, myriam.schaschek@uni-wuerzburg.de

Lisa Straub
Julius-Maximilians-Universität Würzburg, Germany, lisa.straub@uni-wuerzburg.de

Axel Winkelmann
Julius-Maximilians-Universität Würzburg, Germany, axel.winkelmann@uni-wuerzburg.de

What is the Minimum to Trust AI?—A Requirement Analysis for (Generative) AI-based Texts

Research Paper

Christoph Tomitza, Myriam Schaschek, Lisa Straub, and Axel Winkelmann

Julius-Maximilians-Universität Würzburg, Würzburg, Germany
{name.surname}@uni-wuerzburg.de

Abstract. The generative Artificial Intelligence (genAI) innovation enables new potentials for end-users, affecting youth and the inexperienced. Nevertheless, as an innovative technology, genAI risks generating misinformation that is not recognizable as such. The extraordinary AI outputs can result in increased trustworthiness. An end-user assessment system is necessary to expose the unfounded reliance on erroneous responses. This paper identifies requirements for an assessment system to prevent end-users from overestimating trust in generated texts. Thus we conducted requirements engineering based on a literature review and two international surveys. The results confirmed the requirements which enable human protection, human support, and content veracity in dealing with genAI. Overestimated trust is rooted in miscalibration; clarity about genAI and its provider is essential to solving this phenomenon, and there is a demand for human verifications. Consequently, our findings provide evidence for the significance of future IS research on human-centered genAI trust solutions.

Keywords: Human-Centered, Generative AI, Trust, Requirements, ChatGPT

1 Introduction

"As an AI language model, I strive to provide accurate and reliable information to the best of my abilities. However, since I am a machine learning algorithm, there is always a possibility that errors or inaccuracies can occur in my responses." (OpenAI 2023*a*). This is a generative Artificial Intelligence (genAI) system's response to the query of whether one can trust its output. Although the genAI's statement is valid, novice genAI end-users are neither enlightened nor confronted with this fact (Ng et al. 2021). Recent advances in genAI technologies present unique value and precede new research directions while posing challenges for society (Kshetri et al. 2023). Its generative nature distinguishes it from other Artificial Intelligence (AI) types. In particular, genAI's ability to produce multiple and divergent outputs for a given input allows them to mimic human behavior in an exploratory way fostering adoption and diffusion (Weisz et al. 2023). Unlike other AI models, genAI technology merges end-user and technology and creates ease of access in everyday life. The acceleration of genAI capabilities accompanied by public access to pre-trained models enable inexperienced users to co-creation with AI (Muller et al. 2022).

18th International Conference on Wirtschaftsinformatik
September 2023, Paderborn, Germany

Alongside the advancement in Natural Language Generation (NLG), the integration of large language models fine-tuned with human feedback unlocked the generation of intelligent chat bots with remarkable abilities (Bai et al. 2022). A recent widespread solution is ChatGPT, provided by Open AI, which produces sophisticated responses to question prompts mostly indistinguishable to responses provided by human experts (Bai et al. 2022). In parallel to its rapid global adoption, attention raised towards its limitations and ethical reservations regarding its outputs misuse or abuse (Kshetri et al. 2023). Concerning novice end-users, AI-generated text poses a potential risk for spreading misinformation, as its complexity challenges the content provenance and veracity (Fielding 2019, Ali et al. 2021).

In this context, researchers discovered that NLG models are prone to produce content unfaithful to its source input, which is referred to as hallucination (JiZiwei et al. 2023). In addition, end-users tend to evaluate sources primarily on superficial criteria, such as the website design or the availability of illustrations, which are elements that can be easily falsified or manipulated (Griesbaum 2022). In the case of AI-generated texts, end-users can neither see the corresponding input source nor verify how credible the information is. Accordingly, well-known methods such as the CRAAP test for verifying information sources are only applicable to a limited extent (Griesbaum 2022, Ali et al. 2021). Therefore, it is even more drastic that NLG models hallucinate content veracity and trustworthiness for the end-user. The following conversation with a ChatGPT bot illustrates a hallucination example and the consequences of misinformation (Figure 1).

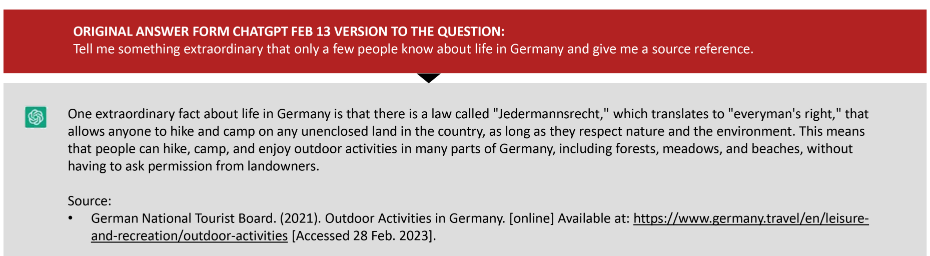

Figure 1. Chat GPT Conversation (OpenAI 2023*b*)

In Germany, it is not allowed to camp wildly - it usually even corresponds to an administrative offense that can be punished with a fine (Hawlitschek et al. 2017). Therefore, the answer given was incorrect. Yet, it is even more problematic that the output is connected to a fake source misleading the end-user's trust. Due to the probabilistic and generative nature of NLG models, outputs are prone to generative variability and inputs to pre-trained models can correspond to multiple outputs (Macneil et al. 2022, Dale 2021). Accordingly, compared to source citation in human-generated texts, it might not display the content provenance. At the current state, the end-user interface of publicly available genAI models do not inform end-users regarding the uncertainty of output. Especially, AI-novices with limited AI literacy might overestimate the trustworthiness of the system and are misled to spread misinformation (Long & Magerko 2020). This backdrop served as motivation to investigate the following research question:

> *What are the requirements to prevent end-users from overestimating trust in AI-generated texts?*

To address our research question, we develop requirements for designing compliant genAI models that anticipate end-users' vulnerability. In doing so, we ground our results in scholarly literature and rely on two complementary survey studies to refine the determined requirements. In light of our findings, we discuss solution proposals to support future end-users in assessing the content veracity and trustworthiness independent of the educational level and AI literacy.

This paper is structured as follows: Section 2 provides a theoretical background on human-centered genAI and highlights preliminary work. Section 3 outlines the details of our research design. Subsequently, Section 4 and 5 present the derived requirements and supplementary survey results. In Section 6, we discuss our findings and derive implications. Finally, we conclude the paper with an outlook on future research avenues.

2 Theoretical Background

2.1 Human-centered Generative Artificial Intelligence

GenAI refers to AI systems that process current media to generate new and creative output (Sbai et al. 2018, Nigam et al. 2021, Nobari et al. 2021). With the advent of stochastic and generative models, genAI technologies can mimic human behavior to the point where they no longer act as decision-makers but rather as end-user support (Seeber et al. 2020, Zhou et al. 2020). Accordingly, the role of AI is shifting from problem-solving to problem-finding (Seeber et al. 2020). Ignited by the public accessibility of new models, AI-based text generation has reinforced interest in human-centered AI research (Muller et al. 2022). In parallel, research attention shifts toward the method of text generation using intelligent chat bots (Radford et al. 2018) that has the distinct feature of co-creating outputs with end-user interaction (Lund & Ting 2023, King 2023).

Cutting-edge NLG models generate conversational responses to question prompts for everyday scenarios (Bai et al. 2022). While they can enhance productivity by generating outputs almost indistinguishable from human content (Lehmann et al. 2022), it is essential to understand and acknowledge its limitations to prevent end-users from overrated trust in AI-generated texts. Given its probabilistic nature, potential concerns from misuse and abuse can lead to reputation or legal issues, infringement on intellectual property, loss of privacy, or the dissemination of false information (e.g., Kshetri et al. 2023). Beyond those threats, ethical issues, such as bias, arise in the context of genAI as a black-box algorithm (e.g., Jovanović & Campbell 2022, Muller et al. 2022).

Hence, it is essential to explore the end-users' perception of AI-generated text and their trust in the generated outputs. Few studies examined the impact of writing with AI (Lehmann et al. 2022) or empirically investigated the societal implications of AI-generated text and revealed that transparency in the use of AI can backfire (Longoni et al. 2022). Other studies have focused on ethical aspects of ChatGPT from the perspective of how it behaves concerning ethically dubious questions (e.g., Dantas 2023, Hasselbalch 2022). Still, open questions persist on how to prevent end-users from harmful decision-making. In this vein, we explore the end-user requirements of NLG models to enable the creation of human-centered genAI systems that promote more responsible usage.

2.2 Preliminary Work and Research Gap

Human-AI Collaboration. The first and most prominent focal point centers on the intersection between AI and end-users. With the emergence of large language models fine-tuned with human participation (Bai et al. 2022), applications in everyday life range from auto-generating text (Lehmann et al. 2022) to code completion (Weisz et al. 2023). The impact of AI assistance has been extensively researched in recent years, with mixed results about whether AI improves work performance or outcome quality (Weisz et al. 2021). With the rise of genAI as a conversational agent that co-creates solutions with end-users, AI's function evolves from decision support to assistance in activities involving the creation of an artifact (Weisz et al. 2022, Zdanowska & Taylor 2022). Research topics include how and what AI can generate and how can and should end-users contribute to the process and quality of outcomes (Muller et al. 2022).

Transparency and Ethics. As AI applications become more prevalent in everyday life, they raise ethical and societal concerns. In response, governments provided principles and regulations to guide organizations developing AI (Smuha 2019). In this light, several recent publications investigate concepts, such as trustworthiness (e.g., Thiebes et al. 2021), explainability (e.g., Meske et al. 2020, Liao & Varshney 2021), fairness (e.g., Lee et al. 2020, Datta et al. 2021), or responsibility (Blodgett et al. 2022, Arrieta et al. 2020), to develop methods under the umbrella term of AI ethics (Shneiderman 2020). A prevalent theme in the literature investigates system trust and trustworthiness as a contractual phenomenon based on the functionalities an AI system aspires to offer the end-user (Vianello et al. 2022).

Falsehood and Misinformation. Even dedicated researchers have been taken aback by the ability of AI-based chat bots to increase linguistic fidelity (Gašević et al. 2023). With the emergence of genAI models, AI is now being used in schools, universities, and research, and the human role shifts to creators and consumers (Ali et al. 2021). Concurrently, chat bots are criticized for creating "cogent waffle" (Vincent 2022), i.e., producing grammatically correct nonsense, and social media networks even accelerate the spread of misinformation. In light of genAI and media bias, the issue of artificial hallucination, defined as machines producing seemingly realistic output that does not match real-world facts, has emerged as a severe issue (Alkaissi & McFarlane 2023, JiZiwei et al. 2023). In this context, mitigating harm includes understanding the technical systems that create deep fakes, how AI can lead to misinformation, and what trustworthy sources are (Ali et al. 2021). Since young people are sensitive to misleading media, the research efforts on media and AI literacy in education amplify (Laupichler et al. 2022). Multiple studies emphasize the relevance of AI and media literacy in public. However, they also highlight the need for policymakers and AI development organizations to inform end-users to prevent harm (Ali et al. 2021).

Ethical concerns about AI-based conversational agents have amplified on another level. While high-stakes issues are often relevant in discussions about AI decision-making (Wanner et al. 2020), various genAI models are publicly available and therefore have broader ethical and societal implications (Kshetri et al. 2023, Lund & Ting 2023).

The research on designing compliant genAI models for text generation to mitigate potential harm is still in its infancy (e.g., Sun et al. 2022, Weisz et al. 2023). To date, scant attention has been paid to human-centered requirements to prevent end-users from overestimating trust in AI-generated texts.

3 Research and Survey Design

In order to define requirements that should be considered in society, we followed a four-stage research design and included a double iteration cycle with two surveys (Figure 2).

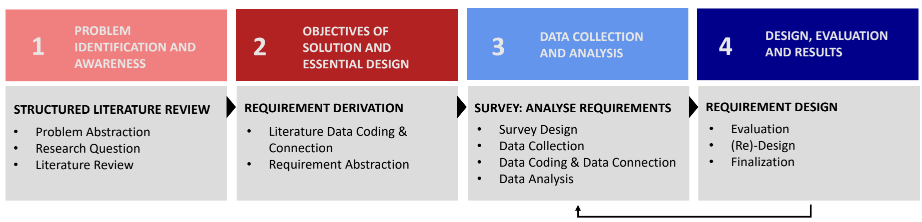

Figure 2. Research Design.

Stage 1: We executed a structured literature review (vom Brocke et al. 2009) to inform our results with prior design knowledge (vom Brocke et al. 2020) on human-centered NLG models regarding end-user trust. For that, we determined the databases relevant to our research problem: AIS eLibrary, ACM, IEEE, Science Direct, and Web of Science. Then, we searched databases using the keywords: (*responsible* OR *ethical* OR *explainable* OR *trust* OR *human-centered*) AND ("*generative AI*" OR "*generative artificial intelligence*"). Subsequently, we limited the identified search results to relevant publications addressing our research question by performing a title and abstract analysis followed by a full-text analysis. Due to the manageability and timeliness of the search results, we did not define further inclusion or exclusion criteria. After a forward and backward search, we finally identified *n*=46 relevant publications that target genAI and related challenges. Within this area of investigation, one can distinguish three general focal points: i) human-AI collaboration (*n*=12), ii) transparency and ethics (*n*=19), and iii) falsehood and misinformation (*n*=9) (see Section 2.2).
Stage 2: After identifying existing concepts, we focused on deriving the requirements. All authors subsequently reread the literature and performed data coding, mainly focusing on the crucial aspects for setting up requirements. We reconciled the coding results in a workshop and cross-checked the findings to define the requirements. Then, we clustered requirements into dimensions for better clarity and structuring. Through multiple iterations, we adjusted the dimensions and requirements to ensure that all key findings were incorporated and appropriately aligned.
Stage 3 and 4: To verify the established requirements, we conducted two surveys in February 2023. The first survey aimed to align the requirements based on the participants' answers, and the second survey evaluated the refined requirements. Additionally, we

conducted a qualitative analysis of the survey results and included the participants' explanations for their decisions and feedback to inform the requirements. In both studies, participants went through a story-based questionnaire about the impact of new text-generation systems (Figure 3).

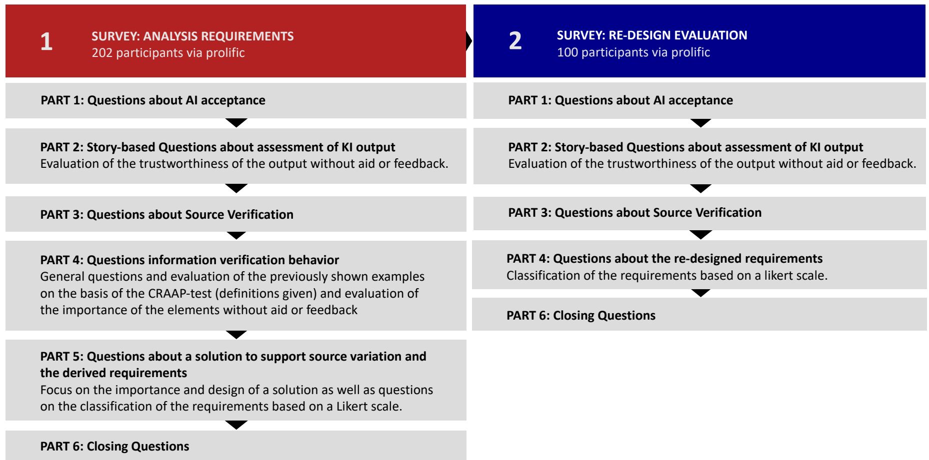

Figure 3. Main sequence and structure of the two survey studies.

We conducted an explanatory cross-sectional survey with quantitative and qualitative elements (Pinsonneault & Kraemer 1993). In detail, the questionnaire included Likert items with five response options and written descriptors accompanied with open question to allow for participant feedback. We implemented both studies using the prolific.co platform[1] with monetary incentive. Regarding the survey participants, we prioritized a divers and transnational main target group with balanced gender and targeted residents of industrialized countries, such as Canada, United States, and European countries. Participants range in age from 18 to 62, and most of our participants are in the 21-30 age group. In general, we targeted potential adult end-users of genAI systems who did not necessarily have prior experience. To ensure a survey that genAI novices could complete, we developed a questionnaire based on a story representing the use of a genAI chat bot. See Tomitza et al. (2023) for more details on the surveys and the corresponding demographics of participants.

4 Human-Centered Requirements to Prevent End-Users from Overestimating Trust in AI-generated Texts

In the subsequent Section, we present the synthesis of literature and survey data in light of three dimensions established through data coding and analysis. The dimensions classify requirements according to their focus and impact on future envisioned artifacts to mitigate harm. They divide into: i) harm protection, ii) human support, and iii) content veracity. Figure 4 depicts the final requirements and associated dimensions. In the preceding subsections, we present the requirements in more detail.

[1] https://www.prolific.co/

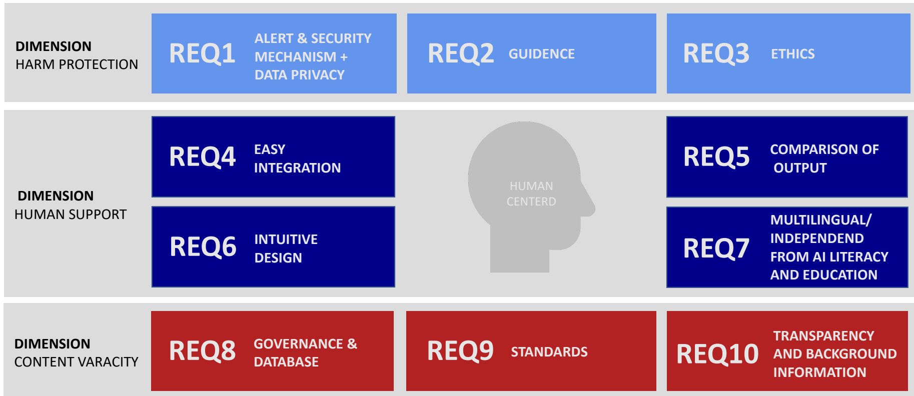

Figure 4. Human-centered requirements to prevent overestimating trust in AI-generated texts.

Harm Protection. Several GenAI studies emphasize the importance of **harm protection** in relation to genAI outputs (According to several (Weisz et al. 2023, Longoni et al. 2022). Accordingly, we see the following requirements from the end-user perspective.

(REQ1) *Integrate alerts and security mechanisms to assess misinformation, detect deep fakes, and maintain personal data protection.* This statement involves data provided by end-users and data they disclose. Therefore, end-users must be aware of erroneous information and insufficient or inappropriate sources (Alkaissi & McFarlane 2023, Ali et al. 2021). In addition, they need privacy protections for their personal data while dealing with genAI systems (Carlini et al. 2021).

(REQ2) *Create helpful guidance for users and educate them about the limitations and potential dangers of using it concerning the veracity and transparency of sources.* To prevent end-users from being misled, it is essential to guide inexperienced or young people in the use of genAI and confront them with the potential risk (Wolf 2019, Laupichler et al. 2022, Ali et al. 2021). So they can learn to evaluate model outputs critically, and fake news should be recognizable easily as such.

(REQ3) *Consider & assure ethics (e.g., discrimination, fairness) and prevent ethically questionable rule-based corrections in favor of freedom of opinion.* The solution should recognize and declare ethically objectionable statements, discrimination, and fairness violations as such (Vianello et al. 2022, Lee et al. 2020, Weisz et al. 2023). However, a solution has not to practice discrimination by certifying sources of certain minorities as untrustworthy. Furthermore, the solution has not to correct automatically unethical statements in favor of freedom of speech.

Human Support. The following requirements regarding **human support** considers solution elements that concern handling. The aim is to avoid creating approaches that overburden the end-user.

(REQ4) *Ensure easy integration with existing information sources and simple use even for the information providers.* For a corresponding solution to gain the most significant possible acceptance by providers, it must be possible to integrate it effortlessly.

(REQ5) *Enable comparability of given information/outputs and ensure transferability /development of the solution to other information sources independent of AI.* The support structure has to enable a comparison of different outputs from one or different models or sources (Dhanorkar et al. 2021). Therefore, when not developed individually for one application (e.g., Chat GPT), the solution provides significant added value.
(REQ6) *Ensure a user-centered, contemporary, intuitive, and inclusive design that enables a ranking of outputs independently.* A high degree of usability enables higher stakes in terms of acceptance. The solution should be intuitive without requiring training time and incorporate desired functionality (Wolf 2019, Zdanowska & Taylor 2022).
(REQ7) *Consider multilingualism and education levels and create comprehensibility of solution regardless of AI competency /AI education level.* The developers of a solution should guarantee ease of use and build a simple and comprehensible support tool that considers sources from different countries and is understandable regardless of age and AI literacy (Laupichler et al. 2022, Ali et al. 2021).

Content Veracity. The last dimension concerns **content veracity**. Negative influences on the output should be made transparent to the user. The synthesis led to the following requirements.
(REQ8) *Establish governance to ensure that the database grounds only on reliable sources or point out if not.* The solution must ensure that the provider only can use reliable databases or sources for their model development (Alkaissi & McFarlane 2023). The database is decisive for what is ultimately output to the end-user (Alkaissi & McFarlane 2023). Model and database have a crucial influence on the output, it is essential to create appropriate indications and clear approaches to devise more clarity of reliability.
(REQ9) *Create a contemporary standard and guarantee the solution's neutrality, comprehensibility, trustworthiness, and explainability.* Achieving a neutral, insusceptible, and contemporary standard that ensures the trustworthiness of genAI models is necessary. Furthermore, especially from an end-user perspective, an independent authority must verify and represent these.
(REQ10) *Demonstrate transparency (e.g., composition, accuracy, AI influence, content quality) by providing background information and support to cross-check.* End-users perceive AI models as black-boxes due to their untraceable complexity (Herm et al. 2023), which influences trustworthiness. Therefore, output accuracy and sources must be indicated to the end-users to extend trust (Dhanorkar et al. 2021). It is essential to display AI involvement and background information (Alkaissi & McFarlane 2023). For reliable results, systems must check responses before displaying them to the end-user, and afterward, guidance for manual evaluation is necessary (Wolf 2019).

We evaluated the requirements during our survey and the results depicted in Figure 5 show that more than 63% of participants agreed to each requirement. The participants' agreement on requirements is equally distributed, but RQ1 outperforms the others, and RQ6 reached the lowest agreement.

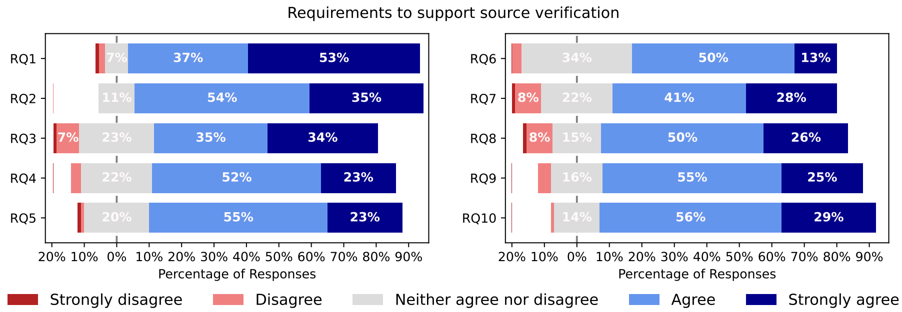

Figure 5. Results of the second study regarding requirements.

5 Supplementary Study Results

The shown requirements provide a first basis for other developments of proper solutions. In our study, it became clear that over-reliance can be triggered simply through grammatically correct statements. To introduce the participants to the study, we asked them to put themselves in the role of planning a trip, showed three statements from ChatGPT, and asked them to rate their confidence in the statements (Figure 6).

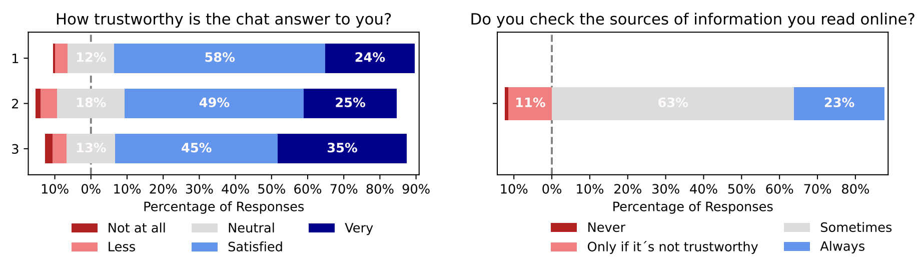

Figure 6. Results of the first survey Part 2 (left) and 3 (right).

The statements shown in Part 2 of the survey were structured as follows: (1) Included possible travel locations, (2) featured a hallucination and incorrect source regarding camping in Germany, and (3) questioned the reliability of the previous statement while providing a detailed answer with additional incorrect source citations. Moreover, we allowed for participant feedback to justify the decision. The most striking finding that we observed reinforced trust by participants when gen AI used (incorrect) source citations. Participants ranked the responses trustworthy if they contained more detail or real-world examples. To increase trust, expanding the number of sources was sufficient. In line with previous studies, artificial hallucinations can cause absolute misdirection by simulating legal content, creating fake documented evidence, and making untrue fake reputations on non-existent websites that imitate credibility (Alkaissi & McFarlane 2023).

Although it is possible to create fraudulent sureness in generated texts by citing sources, these do not have to be correct sources, which we could see from the free text information in which participants indicated that they clicked on links that did not work and still trusted the response. The phenomenon that end-users ignore even aspects that are not true was also evident in the study from Bauer et al. (2023). Another result from the research indicates that most participants only sometimes check every piece of online information, as illustrated in Figure 6. In connection with the aspect that participants were inclined to use AI in everyday life (Figure 7), we see a need for action here.

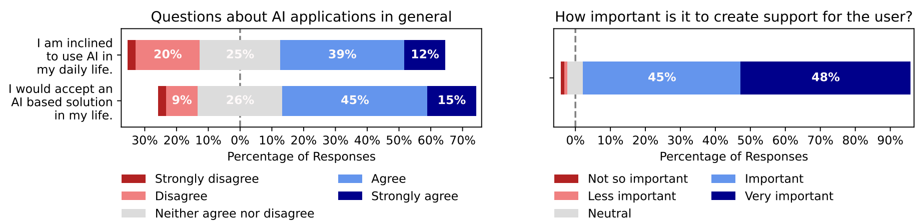

Figure 7. Results of the first survey Part 1 (left) and 5 (right).

While the results on the system´s output trustworthiness unveil an over-reliance bias, a closer look at the participants´ responses reveals a contradiction between trust in using the system and a general distrust of technical systems. After answering the questionnaire on system output and requirements, participants gave us feedback on how they envisage a trustworthy AI system. Despite the propelling results on system trust during usage, the solution suggestions reveal a high preference for human involvement and support. Examples are a real-time chat function, a contact button, expert committees to verify content, CRAAP test guidance, or supervision of AI by human experts. One participant that preferred human verification stated: *"[I want to h]ave a contactable human willing to provide more evidence to prove the AI's authenticity"*. On a higher level of abstraction, we crystallized the paradox of the human need for human verification, namely the desire for human-made content provenance to verify machine output. Exactly this finding is contradictory, as all solutions with human involvement are inferior to machine-based or technical approaches to authenticate content from different perspectives.

6 Discussion

"AI puts everything in front of you. What's there not to believe?". Overall, this excerpt from a participant's response reflects the major insights of our studies. Moreover, it shows above all why harm protection, support for people, and content truthfulness are necessary and why addressing them through our established requirements is important.

Trust Miscalibration. Recent research confirms that a solid database and a well-trained model are essential for meaningful output. Furthermore, both factors show improvement through human-assisted uncovering and monitoring hallucinations (Alkaissi & McFarlane 2023), consistent with the participants' feedback. We advocate banning citations based on probability models, and developers must include specific rules for genAI systems in the source codes. Based on the study results, this has a significant and highly positive effect on protecting end-users from overestimating trust. Furthermore, a voluntary continuation of this approach prevents future legal regulations for AI developments, which were otherwise also seen by the survey participants as an instrument for protecting end-users. Instead of relying on legal regulations, many participants propose implementing trust grading systems (e.g., classification labels, trust charts, or raking systems). Further possible solutions included a web widget, a verification platform, color-coding, or the output of keywords for self-research. After all, it is evident that more research is required to develop adequate measures to prevent trust miscalibration.

Human Verification. Studies on end-user trust and trustworthiness of AI systems assume that AI providers and consumers should agree on contracts transparently and show that contracts might differ between the contexts of use and individual preferences (Vianello et al. 2022, Jacovi et al. 2021). In line with this, other studies demonstrate the need for explainable AI solutions to support end-user trust (Mohseni et al. 2021). In the context of our findings, we suggest further investigating the relationship between over-reliance on system output and explainability to mitigate misinformation. However, we want to highlight that this topic is even more complex as the role of the end-user and AI is shifting in the light of generative AI as a conversational agent. Thus, previous findings on AI as decision support might not be directly transferable to the new situation. Furthermore, both factors show improvement through human-assisted evaluation (Alkaissi & McFarlane 2023), which aligns with our participants' feedback.

Clarity. If the derived requirements were implemented in an end-user assessment system, clarity could be achieved on two levels. Firstly, it would foster media or AI literacy, and secondly, it would increase trust in the provider-consumer relationship. To increase media and AI literacy, there is a need for awareness and a better understanding of technical systems based on genAI (Wolf 2019). A solution that supports end-users dealing with chatGPT could increase media and AI literacy by giving the end-user information about algorithms, context, and outputs. A genAI system should not have handwritten rules or algorithms to avoid misinformation and to respect the freedom of opinion. Having unrestricted end-user support is essential during the interpretation phase of AI output. It is crucial for the end-users to realize how to comprehend the output in a supervised environment and evaluate their confidence level (Kshetri et al. 2023). Especially for younger people solutions with protective mechanisms are relevant, as growing up with the new technologies makes them less suspicious. Accordingly, they are more vulnerable to misleading information (Ali et al. 2021).

To increase the trustworthiness in the provider-consumer relationship, the end-user needs background information about the system (Liao et al. 2020). Our study findings display that the background information should contain also insights about the property of AI, creators of AI, sponsors, funding, and the database. The provider can offer the consumer a FAQ solution with the necessary information. It should be standardized for all providers to increase transparency and trust. Providers should be transparent with consumers about the possibility of uncertain outputs and explain potential risks beforehand. This could prevent any disappointment or breach of trust in the future.

Limitations. Our study demonstrates that humans miscalibrate trust in AI-generated text induced by artificial hallucinations and we derive requirements to mitigate the harm caused. However, our findings must be understood relative to their context. The story-based questionnaire with Likert-type and open questions enabled quantitative and qualitative data analysis. Although we tried to mitigate bias in data collection by balancing out gender and sampling nationalities, we observed an uneven distribution in ethnicity. We conducted our survey using prolific.co platform that might not represent a cross-section of the entire society. Thus, we cannot formally rule out the presence of any bias. While our survey could benefit from a larger sample size, due to the received positive feedback and results, we are that the findings are representative and sound.

7 Conclusion and Outlook

GenAI enables new potential for society. Capabilities, such as the support in creating articles or the joint production of music, will increase (Kshetri et al. 2023). Thus, genAI bears risks for end-users, such as misinformation (Rudolph et al. 2023).

Our research identifies requirements to prevent end-users from overestimation trust in AI-generated texts. We conducted a literature review and two studies to develop and evaluate human-centered requirements. Our results contain ten requirements that indicate what a solution needs to support end-users by interpreting and dealing with text genAI. In addition, we abstracted three dimensions for developing such supporting systems. Furthermore, we observed that end-users trust the answers from genAI highly. If the chat bot provides references, the trustworthiness increase, regardless of the source's quality or the answer's content. Another observation is that participants desire to contact human support to verify the statements in case of issues or ambiguities.

With our results, we want to increase the awareness of IS research to investigate the trustworthiness of genAI and develop dedicated support to consider unintended consequences. Future work is required to transform the requirements into design principles and to implement prototypical solutions, such as a source-varifying label or a standardized FAQ. Likewise, there is a need to expand research intensity in trust and genAI and investigate the effects on end-users. In this context, it is interesting to look at the implications for platforms and deduce how the AI trust factor affects customer perception. Overall, long-term studies could be initiated. In addition, we see the current research regarding the influences of genAI on the end-user as expandable to reduce potential dangers and to promote the technology's further development. We want to contribute to the further evolution of technology and consciously point out the risk to enable the development of necessary solutions.

8 Acknowledgement

This work has been developed in the project pipeAI (DIK0143/02) and is partly funded by the Bavarian Ministry of Economic Affairs, Regional Development and Energy within the R&D program "Informations- und Kommunikationstechnologie" managed by the VDI/VDE-IT. The authors are responsible for the content of this publication.

References

Ali , S., DiPaola, D., Lee, I., Sindato, V., Kim, G., Blumofe, R. & Breazeal, C. (2021), 'Children as creators, thinkers and citizens in an ai-driven future', *Computers and Education: Artificial Intelligence* **2**.

Ali, S., DiPaola, D., Lee, I., Hong, J. & Breazeal, C. (2021), Exploring generative models with middle school students, *in* 'Conference on Human Factors in Computing Systems - Proceedings', Association for Computing Machinery.

Alkaissi, H. & McFarlane, S. I. (2023), 'Artificial hallucinations in chatgpt: Implications in scientific writing.', *Cureus* **15**.

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. & Herrera, F. (2020), 'Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai', *Information Fusion* **58**, 82–115.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T. et al. (2022), 'Training a helpful and harmless assistant with reinforcement learning from human feedback', *aXiv preprint arXiv:2204.05862* .

Bauer, K., von Zahn, M. & Hinz, O. (2023), 'Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing', *Inf. Syst. Res.* .

Blodgett, S. L., Liao, Q. V., Olteanu, A., Mihalcea, R., Muller, M., Scheuerman, M. K., Tan, C. & Yang, Q. (2022), 'Responsible Language Technologies: Foreseeing and Mitigating Harms', *Conference Human Factors in Computing Systems - Proceedings* .

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., Erlingsson, U. et al. (2021), Extracting training data from large language models., *in* 'USENIX Security Symposium', Vol. 6.

Dale, R. (2021), 'Gpt-3: What's it good for?', *Natural Language Engineering* **27**(1), 113–118.

Dantas, A. C. (2023), 'Ethics in chatgpt and other ai's'.
URL: *https://medium.com/@adilmarcoelhodantas/ethics-in-chatgpt-and-other-ais-ee31ce8e9f09*

Datta, P., Whitmore, M. & Nwankpa, J. K. (2021), 'A Perfect Storm', *Digital Threats: Research and Practice* **2**(2).

Dhanorkar, S., Wolf, C. T., Qian, K., Xu, A., Popa, L. & Li, Y. (2021), 'Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations across the AI Lifecycle', *DIS 2021 - Proceedings of the 2021 ACM Designing Interactive Systems Conference: Nowhere and Everywhere* pp. 1591–1602.

Fielding, J. A. (2019), 'Rethinking craap: Getting students thinking like fact-checkers in evaluating web sources', *College & Research Libraries News* **80**(11), 620.

Gašević, D., Siemens, G. & Sadiq, S. (2023), 'Empowering learners for the age of artificial intelligence', *Computers and Education: Artificial Intelligence* p. 100130.

Griesbaum, J. (2022), *Informationskompetenz*, Springer Berlin Heidelberg, pp. 67–98.

Hasselbalch, G. (2022), 'Testing chatgpt's ethical readiness'.
URL: *https://dataethics.eu/testing-chatgpts-ethical-readiness/*

Hawlitschek, F., Kranz, T. T., Elsner, D., Fritz, F., Mense, C., Müller, M. B. & Straub, T. (2017), 'Sharewood-forest–a peer-to-peer sharing economy platform for wild camping sites in germany', *HOHENHEIM DISCUSSION PAPERS IN BUSINESS, ECONOMICS AND SOCIAL SCIENCES* p. 265.

Herm, L.-V., Heinrich, K., Wanner, J. & Janiesch, C. (2023), 'Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability', *International Journal of Information Management* **69**, 102538.

Jacovi, A., Marasović, A., Miller, T. & Goldberg, Y. (2021), 'Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai', *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* pp. 624–635.

JiZiwei, LeeNayeon, FrieskeRita, YuTiezheng, SuDan, XuYan, IshiiEtsuko, Jin, B., MadottoAndrea & FungPascale (2023), 'Survey of hallucination in natural language generation', *ACM Computing Surveys* **55**.

Jovanović, M. & Campbell, M. (2022), 'Generative Artificial Intelligence: Trends and Prospects', *Computer* **55**(10), 107–112.

King, M. R. (2023), 'The Future of AI in Medicine: A Perspective from a Chatbot', *Annals of Biomedical Engineering* **51**(2), 291–295.

Kshetri, N., Hughes, L., Jeyaraj, A., Kar, A. K., Dwivedi, Y. K., Slade, E. L. & Wright, R. (2023), 'Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy', *International Journal of Information Management* **forthcoming**.

Laupichler, M. C., Aster, A., Schirch, J. & Raupach, T. (2022), 'Artificial intelligence literacy in higher and adult education: A scoping literature review', *Computers and Education: Artificial Intelligence* **3**, 100101.

Lee, M. K., Grgić-Hlača, N., Tschantz, M. C., Binns, R., Weller, A., Carney, M. & Inkpen, K. (2020), 'Human-centered approaches to fair and responsible AI', *Conference on Human Factors in Computing Systems - Proceedings* .

Lehmann, F., Markert, N., Dang, H. & Buschek, D. (2022), 'Suggestion Lists vs. Continuous Generation: Interaction Design for Writing with Generative Models on Mobile Devices Affect Text Length, Wording and Perceived Authorship', *ACM International Conference Proceeding Series* pp. 192–208.

Liao, Q. V., Gruen, D. & Miller, S. (2020), 'Questioning the AI: Informing Design Practices for Explainable AI User Experiences', *Conference on Human Factors in Computing Systems - Proceedings* pp. 1–15.

Liao, Q. V. & Varshney, K. R. (2021), 'Human-Centered Explainable AI (XAI): From Algorithms to User Experiences'.

Long, D. & Magerko, B. (2020), What is ai literacy? competencies and design considerations, *in* 'Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems', CHI '20, Association for Computing Machinery, New York, NY, USA, p. 1–16.

Longoni, C., Fradkin, A., Cian, L. & Pennycook, G. (2022), 'News from Generative Artificial Intelligence Is Believed Less', *ACM International Conference Proceeding Series* **22**, 97–106.

Lund, B. & Ting, W. (2023), 'Chatting about ChatGPT: How May AI and GPT Impact Academia and Libraries?', *SSRN Electronic Journal* .

Macneil, S., Tran, A., Mogil, D., Bernstein, S., Ross, E. & Huang, Z. (2022), 'Generating diverse code explanations using the gpt-3 large language model', *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 2* **2**, 3.

Meske, C., Bunde, E., Schneider, J. & Gersch, M. (2020), 'Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities', *https://doi.org/10.1080/10580530.2020.1849465* **39**, 53–63.

Mohseni, S., Zarei, N. & Ragan, E. D. (2021), 'A multidisciplinary survey and framework for design and evaluation of explainable ai systems', *ACM Transactions on Interactive Intelligent Systems (TiiS)* **11**.

Muller, M., Chilton, L. B., Kantosalo, A., Maher, M. L., Martin, C. P. & Walsh, G. (2022), 'GenAICHI: Generative AI and HCI', *Conference on Human Factors in Computing Systems - Proceedings* .

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W. & Qiao, M. S. (2021), 'Conceptualizing AI literacy: An exploratory review', *Computers and Education: Artificial Intelligence* **2**, 100041.

Nigam, A., Pollice, R., Krenn, M., Gomes, G. D. P. & Aspuru-Guzik, A. (2021), 'Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (stoned) algorithm for molecules using selfies', *Chemical Science* **12**, 7079–7090.

Nobari, A. H., Rashad, M. F. & Ahmed, F. (2021), 'Creativegan: Editing generative adversarial networks for creative design synthesis', *Proceedings of the ASME Design Engineering Technical Conference* **3A-2021**.

OpenAI (2023*a*), 'Chatgpt feb 13 version'.

OpenAI (2023*b*), 'Chatgpt feb 13 version'.

Pinsonneault, A. & Kraemer, K. (1993), 'Survey research methodology in management information systems: an assessment', *Journal of management information systems* **10**(2), 75–105.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al. (2018), 'Improving language understanding by generative pre-training'.

Rudolph, J., Tan, S. & Tan, S. (2023), 'Chatgpt: Bullshit spewer or the end of traditional assessments in higher education?', *Journal of Applied Learning and Teaching* **6**(1).

Sbai, O., Elhoseiny, M., Bordes, A., LeCun, Y. & Couprie, C. (2018), 'Design: Design inspiration from generative networks', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11131 LNCS**, 37–44.

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G. J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G. & Söllner, M. (2020), 'Machines as teammates: A research agenda on ai in team collaboration', *Information & Management* **57**, 103174.

Shneiderman, B. (2020), 'Bridging the Gap Between Ethics and Practice', *ACM Transactions on Interactive Intelligent Systems (TiiS)* **10**(4).

Smuha, N. A. (2019), 'The eu approach to ethics guidelines for trustworthy artificial intelligence', *Computer Law Review International* **20**(4), 97–106.

Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K. & Weisz, J. D. (2022), 'Investigating Explainability of Generative AI for Code through Scenario-based Design', *International Conference on Intelligent User Interfaces, Proceedings IUI* **22**, 212–228.

Thiebes, S., Lins, S. & Sunyaev, A. (2021), 'Trustworthy artificial intelligence', *Electronic Markets* **31**, 447–464.

Tomitza, C., Schaschek, M., Straub, L. & Winkelmann, A. (2023), 'Supplementary material for "what is the minimum to trust ai?—a requirement analysis for (generative) ai-based texts"'.
URL: *https://b2share.eudat.eu/records/5109b4ee67894844a419c20e6522dbbe*

Vianello, A., Laine, S. & Tuomi, E. (2022), 'Improving Trustworthiness of AI Solutions: A Qualitative Approach to Support Ethically-Grounded AI Design', *https://doi.org/10.1080/10447318.2022.2095478* .

Vincent, J. (2022), 'Openai's new chatbot can explain code and write sitcom scripts but is still easily tricked'.
URL: *https://www.theverge.com/23488017/openai-chatbot-chatgpt-ai-examples-web-demo*

vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R. & Cleven, A. (2009), Reconstructing the giant: On the importance of rigour in documenting the literature search process, *in* 'European Conference on Information Systems', pp. 2206–2217.

vom Brocke, J., Winter, R., Hevner, A. & Maedche, A. (2020), 'Special issue editorial–accumulation and evolution of design knowledge in design science research: a journey through time and space', *Journal of the Association for Information Systems* **21**(3), 9.

Wanner, J., Herm, L.-V., Heinrich, K., Janiesch, C. & Zschech, P. (2020), White, grey, black: Effects of xai augmentation on the confidence in ai-based decision support systems., *in* 'Proceedings of Forty-First International Conference on Information Systems'.

Weisz, J. D., Maher, M. L., Strobelt, H., Chilton, L. B., Bau, D. & Geyer, W. (2022), 'HAI-GEN 2022: 3rd Workshop on Human-AI Co-Creation with Generative Models', *International Conference on Intelligent User Interfaces, Proceedings IUI* pp. 4–6.

Weisz, J. D., Muller, M., He, J. & Houde, S. (2023), 'Toward General Design Principles for Generative AI Applications'.

Weisz, J. D., Muller, M., Houde, S., Richards, J., Ross, S. I., Martinez, F., Agarwal, M. & Talamadupula, K. (2021), 'Perfection Not Required? Human-AI Partnerships in Code Translation', *International Conference on Intelligent User Interfaces, Proceedings IUI* pp. 402–412.

Wolf, C. T. (2019), 'Explainability scenarios: Towards scenario-based XAI design', *International Conference on Intelligent User Interfaces, Proceedings IUI* **Part F1476**, 252–257.

Zdanowska, S. & Taylor, A. S. (2022), 'A study of UX practitioners roles in designing real-world, enterprise ML systems', *Conference on Human Factors in Computing Systems - Proceedings* .

Zhou, Y., Koyama, Y., Goto, M. & Igarashi, T. (2020), 'Generative Melody Composition with Human-in-the-Loop Bayesian Optimization'.