Bridging the Knowledge Gap: Understanding User Expectations for Trustworthy LLM Standards

Michaela Benk[1,2], Léane Wettstein[4], Nadine Schlicker[3], Florian von Wangenheim[1], Nicolas Scharowski[4]

[1] ETH Zurich
[2] Mobiliar Lab for Analytics
[3] Philipps-Universität Marburg
[4] University of Basel

Abstract

Researchers, policymakers, and developers of artificial intelligence (AI) are actively collaborating to establish trustworthy AI standards that align with broader societal values, particularly in the context of large language models (LLMs). However, the critical discourse on bridging the vast knowledge gap between experts who shape and implement standards for LLMs and users whose values are at stake remains largely unaddressed. Taking a "bottom-up" perspective and using a mixed-method approach, we first conducted interviews ($N = 12$) to engage with users' perceptions of normative standards in the context of LLMs. We thereby identified 68 specific criteria that users' consider when evaluating whether their values are fulfilled. Second, we conducted an online survey ($N = 379$) to further investigate how users prioritize these standards and the identified criteria in conversational LLM-based applications. Our findings reveal opportunities for strategic communication measures, the importance of transparent governance mechanisms and the necessity of non-technical complements to technical solutions for bridging the knowledge gap. We discuss actionable steps to effectively communicate trustworthy AI standards.

Extended version — https://osf.io/rkb76/?view_only=02f2b6019f114d77b9cc7980246f2163

Introduction

Large language models (LLMs), such as GPT-4 (Open AI) or LLaMA (Meta), have recently gained traction as they are demonstrating unprecedented performance in a wide array of Natural Language Processing (NLP) tasks (Thirunavukarasu et al. 2023; de Winter 2023) and widespread, rapid adoption (Hu 2023). Despite their remarkable capabilities, significant uncertainties remain due to their ethical and social risk, including the spread of misinformation, discrimination, or compromised privacy (Weidinger et al. 2022; Thirunavukarasu et al. 2023; Liao and Vaughan 2023; Bommasani et al. 2021). To mitigate risks, researchers, policymakers, and industry leaders are engaged in a discourse aimed at establishing and implementing value-aligned principles, guidelines and standards for trustworthy AI in the context of LLMs (Hacker, Engel, and Mauer 2023). For example, methods to increase transparency of AI are now extended to LLMs (Liao and Vaughan 2023) and research is exploring how to adequately assess risks along the complex LLM value chain (Sherman and Eisenberg 2024).

While these discussions have contributed to more ethical practices and an increased focus on human-centered approaches, the knowledge gap between experts who define normative standards and users who are impacted by their implementation remains largely unaddressed. For example, users[1] may not have access to relevant information to determine whether certain standards have been met or lack the relevant knowledge to understand what these standards entail (Knowles and Richards 2021; Mokander et al. 2023). Furthermore, misalignments or inconsistencies in communicating standards may arise due to experts' limited understanding of users' interpretations of these standards (Mokander et al. 2023).

Bridging this gap is important for at least three reasons. (I) Trustworthiness is inherently subjective, relying on the alignment between user values and the values implemented within LLMs. In fact, the definition of AI trustworthiness according to the International Standards Organization (ISO) is "the ability to meet stakeholders' expectations in a verifiable way" (International Organization for Standardization 2020). As such, design affordances and cues, such as interface features, documentation, and certifications addressing standards like safety and fairness, are effective only if users can easily identify and understand them (Chiou and Lee 2021; Schlicker et al. 2023; Liao and Sundar 2022). (II) Communicating user-relevant information is an essential mechanism for effective *facework* through which representatives can convey their commitment to ethical standards and build public trust in AI (Knowles and Richards 2021). (III) As research moves toward human-centered AI practices, it is important to clearly define which trustworthiness information can be conveyed through technical mechanisms, such as explainability approaches embedded within the interaction with the LLM, and those that should be communicated through other means, such as certification labels or other forms of governance disclosure (Scharowski et al. 2023;

[1]Following prior research (Scharowski et al. 2023), we define users as layperson or groups who interact with LLMs in their day-to-day activities, for various tasks, without necessarily having expertise in its underlying mechanisms or development process.

Liao and Vaughan 2023). Particularly for standards that are difficult for users to directly observe and that may lack clearly defined evaluation metrics, such as fairness or safety, it is crucial to identify the specific *cues* that effectively and coherently communicate trustworthiness (Liao and Sundar 2022). This entails (a) identifying the specific information that users seek and consider important to make trust judgments, and (b) exploring the various channels through which this information can be effectively communicated.

To address these gaps, our mixed-method study makes three contributions. First, through a set of interviews with users ($N = 12$), we uncover how they define and understand existing normative standards in the context of conversational LLM-based applications. We identify 68 specific criteria, hereafter referred to as *individual standards* (Schlicker et al. 2023), that users employ to evaluate whether the overarching normative standards, henceforth called *high-level standards*, have been met. Second, through an online survey ($N = 348$), we investigate how users evaluate and prioritize existing standards in the context of conversational LLM-based applications. Findings reveal, among others, that users tend to assess the fulfillment of standards - such as security, safety, and fairness - through the lens of transparent information provision and data handling practices. These aspects are more familiar to users compared to other elements. Comparatively less emphasis is placed on transparency and explainability mechanisms within the interaction, highlighting limitations with respect to technical solutions for communicating standards to form trust judgments. Finally, we discuss how our findings can be applied and guide researchers, developers, policymakers in improving the ethical landscape of LLMs through effective communication strategies.

Our findings aim to complement the existing discourse on ethical AI development. In the following, we synthesize prior research on responsible AI guidelines and current approaches to their implementation, as well as the conceptual underpinnings of our study.

Related Work

Principles and Guidelines for Trustworthy LLMs

Recent years have witnessed a proliferation of AI ethics frameworks from both the public and private sectors (Mittelstadt 2019; Floridi 2019; Jobin, Ienca, and Vayena 2019; Toreini et al. 2019; Adler et al. 2022). The EU, for example, has outlined rigorous guidelines for trustworthy AI (European Commission 2019), followed by similar initiatives in other regions (Université de Montréal 2018; Chinese National Governance Committee 2019; UK House of Lords 2017). Tech giants such as Google and Microsoft have also introduced their own responsible AI frameworks, though their scope and focus vary[2]. These initiatives, while critical for the development and deployment of ethical AI, are often articulated at a high-level of abstraction and for a broad set of technological artifacts. Concepts such as "fairness" and "safety" effectively become "empty vessels" (Yeung, Howes, and Pogrebna 2020), which can lead to varied interpretations and implementations (Jakesch et al. 2022; Mittelstadt 2019).

In response, researchers are investigating how to distill high-level AI standards into more meaningful and actionable guidelines (Yurrita et al. 2022; Constantinides et al. 2024; Li et al. 2023). For example, in their value-based framework, Yurrita et al. (2022) identify a set of values and translate them into specific criteria and manifestations that can be used by policymakers and auditors. By employing a circular organization, the authors effectively illustrate the trade-offs between values, such as explainability versus privacy, and identify overlaps among criteria, such as 'data protection', which serves multiple values. Similarly, Li et al. (2023) aim to unify the fragmented approach to translate high-level standards into practice, considering the entire lifecycle of AI and describing how AI systems can be evaluated and assessed on each of these aspects. Moreover, some user studies have focused on the public's attitudes toward high-level standards. For instance, Saxena et al. (2018) tested public perceptions of different definitions of fairness in the context of loan decisions and found a preference for a definition in which individuals are selected or evaluated proportionally to their merit. However, the nature and integration of LLMs into the digital economy presents unique obstacles to the implementation of existing principles, stemming from several key factors: the inherent opacity of LLM architectures, the proprietary nature of the technology (characterized by their initial development and release by a single entity, followed by adaptation across a myriad of applications by numerous other stakeholders), and the complexity of their applications[3] (Liao and Vaughan 2023; Mokander et al. 2023). Given the subjective nature of aligning user values with trustworthy LLMs and the scarcity of actionable guidelines, translating abstract principles into practical implementation remains a significant challenge for AI developers and deployers (Mittelstadt 2019; Jakesch et al. 2022; Mokander et al. 2023).

A bottom-up approach for trustworthy LLMs. In light of this challenge, a key contribution of our research is to synthesize a comprehensive set of criteria that delineate trustworthy LLMs *from (lay) users' point of view*. Specifically, we use a "bottom-up" approach, which focuses on the expectations and information needs of users regarding the implementation of ethical values for LLMs. We argue that this approach can shed light on the specific informational cues that users focus on when assessing the trustworthiness of LLMs, thereby complementing existing top-down approaches led by policymakers, researchers, or industry leaders and providing developers with more actionable information on what information users seek. This differs significantly from the traditional approach in AI research, which typically isolates

[2]These frameworks range from broad, overarching goals, such as avoiding the creation or reinforcement of unfair bias (Google 2023), to more detailed guidelines, including specific actions like "evaluating all datasets to assess inclusiveness" (Microsoft 2022)

[3]We acknowledge the importance of distinguishing between pre-trained LLM models, adapted LLM models, and the application using the model, as suggested by Liao and Vaughan (2023). Although we strive to adhere to these distinctions, for simplicity in our reporting, we at times refer to all collectively as 'LLM'.

the model to elicit user mental models at the system level (Liao and Vaughan 2023; Liao and Sundar 2022).

For this purpose, we adapt the concept of "individual standards," as introduced by Schlicker et al. (2023), which encapsulates the specific criteria used to evaluate the trustworthiness of a given AI system. For example, *user A* may consider an LLM application trustworthy, if it minimizes the potential for stereotyping, while *user B* may consider it trustworthy, if it provides equal quality of service for marginalized groups. Individual standards can be aggregated into high-level standards that reflect the normative principles outlined previously. In summary, high-level standards encapsulate relevant end-user *values*, such as *fairness* or *performance*, while low-level individual standards specify the criteria necessary to determine whether a high-level standard has been met, such as *equality in service quality* or *reduction of stereotypes*.

In the following paragraphs, we describe existing strategies and available channels to communicate to users whether or to what extent LLMs meet a given standard.

Communicating LLM Trustworthiness to Users

Prior research has explored various strategies to enhance the understanding and communication of normative principles throughout the operational lifecycle of LLMs and LLM-based applications. Building on an existing body of research on trustworthy AI more broadly (Liao and Sundar 2022; Li et al. 2021), several works have focused on the suitability of existing transparency approaches to LLMs. These include documentation for model reporting, such as model cards (Crisan et al. 2022), explanations of a model's internal processes or outputs (Schmude et al. 2023), or uncertainty information (Liao and Vaughan 2023). These methods are designed to help users develop an understanding of a machine learning (ML) model's processes and reasoning, thereby facilitating the evaluation of principles such as robustness, safety and fairness (Mittelstadt 2019).

However, these methods may not suffice or be suitable for users to assess whether individual standards have been met. This is due to several factors: First, these methods are developed according to developers' priorities, and research indicates that users may understand and prioritize standards differently than developers might expect (Jakesch et al. 2022). Second, it may be unreasonable to place the burden of accountability on users, such as expecting them to validate all outputs through the use of explanations, particularly if they lack the necessary knowledge and expertise (Knowles and Richards 2021; Liao and Vaughan 2023). Third, some standards may not be easily observable, such as what types of errors are minimized to guarantee safety. Here, users must rely on other signals or *cues* to make trust judgments (Liao and Sundar 2022; Chiou and Lee 2021). As trustworthiness cues provided by the system are only effective if they are detected and understood by users (Chiou and Lee 2021), research is exploring alternative approaches to effectively communicate trustworthiness to users, including through certification labels (Scharowski et al. 2023). However, determining which information users seek is still an unresolved issue (Liao and Vaughan 2023), and a gap our research seeks to address. Specifically, we formulate the following research questions:

RQ1: What are user-relevant individual standards in the context of LLMs?

RQ2: How do users prioritize standards in the context of LLMs?

Methodology

To answer our research questions, we adopted a mixed-method approach, comprising three phases: (a) a literature search to identify high-level standards, (b) qualitative interviews to identify individual standards for LLMs, and (c) a quantitative user study to examine how users prioritize both the high-level and the identified individual standards for LLMs. We describe each phase below.

Literature Search: Identifying High-Level Standards

In the initial phase of this study, we conducted a narrative literature search to identify principles of AI that are applicable to LLMs. Given the lack of established guidelines in the context of LLMs and varied terminology used in the literature, we identified high-level standards by examining a selection of articles for mentions of standards, frameworks, values, principles, and desiderata for AI. In a second step, four researchers synthesized these findings into a consolidated set of the 13 most frequently cited high-level standards.

Interviews: Identifying Individual Standards

For the second phase of this study, a total of 12 interviews were conducted with users of different backgrounds, ages, and gender that lasted 45 to 60 minutes. The interviews were recorded through field notes and audio recordings.

Participants. Participants were selected based on their familiarity with and prior experience with LLMs. The final sample of 12 participants ($M_{age} = 36.58$, $SD_{age} = 13.34$, $Min_{age} = 21$, $Max_{age} = 59$) was predominantly female (8 women and 4 men) and consisted of students in Business and Economics (P1), Geo-sciences (P8) and Psychology (P2, P3, P4, P6), as well as individuals describing their occupation as teacher (P5), social worker (P9), musician (P10), private banker (P11), television producer (P12) and healthcare worker (P13). Participants were compensated for their participation, with a gift card worth USD 20.

Procedure. All participants provided informed consent, detailing the study objectives, withdrawal rights, and data anonymity. During the interviews, participants initially described their views on the opportunities and challenges associated with using LLMs. After a short briefing on LLM applications and risks, participants were asked to define a set of high-level standards identified in the literature according to their own understanding and in the context of their experiences with LLMs. Participants then received formal definitions derived from the literature and were asked to identify key criteria for their implementation in LLM applications, evaluating four to six standards each.

Survey: Evaluating Individual Standards

Participants. Participants for the online survey were recruited via Prolific[4]. An initial sample of 384 participants were recruited and received £3.00, considered "fair pay" according to the Prolific guidelines, for taking part in the 13-minute online survey. Following recommendations to ensure data quality (Meade and Craig 2012; Brühlmann et al. 2020), a self-reported single item for careless responding was used. After the assessment of data quality, the final sample comprised 379 participants ($M_{age} = 38.41$, $SD_{age} = 11.76$, $Min_{age} = 18$, $Max_{age} = 76$) with a balanced gender distribution (47% female).

Procedure. The third phase of our study was an online survey consisting of two parts. After a brief introduction and giving informed consent, participants were briefed on the applications and potential risks that are associated with the use of LLMs. Next, participants were provided with a formal definition of the high-level standards and asked to rank them by their perceived importance for LLMs. In addition, participants rated the importance of each standard on a scale of 1 to 100.

In the second part of the survey, participants were asked to indicate their preference for the individual standards identified during the interviews, comparing them in pairs within each high-level category (e.g., "system security" vs. "anonymity" for the standard *Security*). They were presented with a balanced and quasi-random pair selection of individual standards and instructed to select the standard they perceived as more important for LLMs. For example, if the participants considered "system security" to be more important than "anonymity" in assessing whether the high-level standard *Security* was met, then they selected this individual standard. They were provided with a definition for each individual standard, which were derived from the interviews. Each participant rated 30 individual pairs, which resulted in an average of 76 ratings for each pair of individual standards. The survey concluded by asking participants for which tasks they mainly use LLMs and requesting feedback concerning the study.

Analysis and Coding Procedure

Qualitative interview data and quantitative survey data were used to address **RQ1** and **RQ2**, respectively. The interview data were analyzed using qualitative content analysis, following (Mayring and Fenzl 2019). Qualitative analysis involved the following steps: paraphrasing the relevant content, generalizing it to a higher level of abstraction, and conducting two stages of data reduction to develop a cross-case category system. The coding process was conducted by four researchers and involved several rounds of discussions and iterations. To ensure similar coding strategies, three researchers independently coded four interviews each. To enhance reliability, one interview was coded collaboratively by two researchers to ensure consistency. Discrepancies were discussed and resolved in open sessions with all researchers, and the final cross-case category system was formed during group sessions.

[4]https://www.prolific.com/

Results

Identifying Individual Standards (RQ1)

Our qualitative analysis identified 68 individual standards. The distribution ranges from a minimum of four standards for *Accountability* and *Robustness* to a maximum of eight standards for *User Experience*. For instance, users anticipate that the high-level standard *Safety* be met through unbiased output (S2.1), user-centered design (S2.2) that takes into account their values, or data handling (S.3). Additionally, we uncovered several individual standards that could address multiple high-level standards, though these may be prioritized differently. This is illustrated by the individual standard "transparent information", which may refer to disclosures about the LLMs storage and usage of data (S1.2), about data handling and processing (S2.3), about the provider of the LLM (S3.4), about the responsibilities of the user (S4.4), and information about the training data and further information necessary to understand the LLM (S5.4). Due to space constraints, we refer the interested reader to the extended version of the paper for a detailed description of all standards.

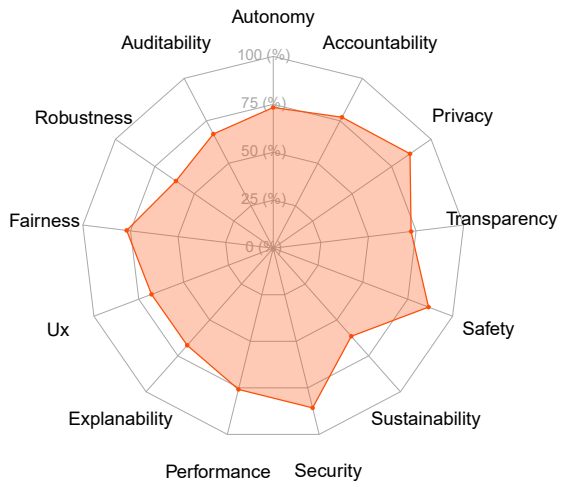

Figure 1: Spider chart illustrating the ranked importance across the 13 different high-level standards, ranging from 1 (="not important at all") to 100 (= "very important").

Evaluating Individual Standards (RQ2)

To contextualize our findings, we initially asked survey participants to indicate for what tasks they use LLMs. The majority of participants (48.3%) use LLMs for Q&A tasks, followed by content generation (24.0%) and proofreading (10.0 %), summarization (9.0 %) and translation (3.7%). Other usages mentioned by participants include administrative tasks and experimentation, with a few participants ($N = 8$) reporting that, while they experiment with LLMs, they do not yet trust them enough to rely on them for actual tasks.

When asked to rate high-level standards, participants assigned the highest importance to privacy ($M = 87.3$, $SD = 16.6$), safety ($M = 86.6$, $SD = 19.0$), security ($M = 86.0$, $SD = 17.1$), accountability ($M = 77.4$, $SD = 20.0$), fairness ($M = 77.0$, $SD = 21.5$) and performance ($M = 75.84$, $SD = 22.42$). In contrast, sustainability received the lowest average rating ($M = 61.4$, $SD = 23.9$). While this more open response format largely confirmed the rankings, we also observed ceiling effects that indicate that some standards are of similar importance to users. The results can be found in Figure 1.

Results from the analysis of the pairwise comparison of individual standards show a number of interesting findings. In the following, we comment on those findings of the six highest-ranked high-level standards, and contextualize them with citations from the interview data. Heatmaps and overall rankings display participants' preferences between pairs of individual standards. Each cell shows the percentage of participants who favored the standard represented by the row over the standard represented by the column. Cell color intensity varies with value deviation from 50%. Values below 50% are increasingly red, while those above 50% become increasingly blue. Heatmaps for all remaining high-level standards can be found in the long version of the paper.

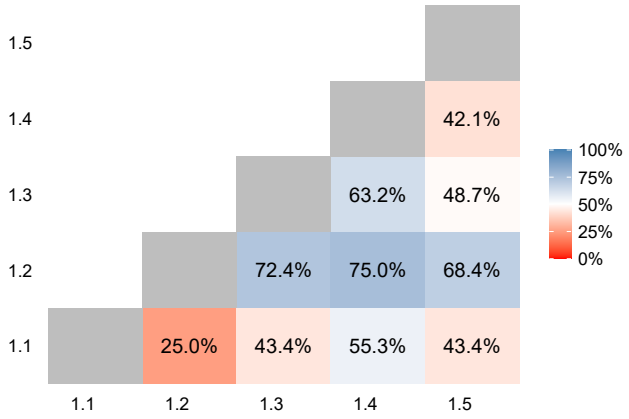

Figure 2: Heatmap displaying **Privacy (S1)**. Individual Standards: 1.1 Data handling; 1.2 Transparent information provision; 1.3 Data Protection; 1.4 Anonymity; 1.5 User Rights.

Privacy (S1) As seen in Figure 2, participants believe that their privacy is better fulfilled with data protection through user login (S1.3), ensuring data are not shared or accessed, than with transparent information provision (S1.2) about *how* their data is used. During the interviews, users frequently mentioned privacy concerns, however, users expressed skepticism whether privacy standards being claimed are indeed being followed. For instance, P10 noted *A log-in can help, or an account, [...] but all of that 'just' increases my trust*, thereby referring to the inability to verify whether the standard is met.

Safety (S2) & Security (S3) Participants considered these standards primarily from a data protection lens. Monitoring (3.6) and governance (3.7) were discussed by several interview participants, however, they lacked information about the regulations that guide these processes.

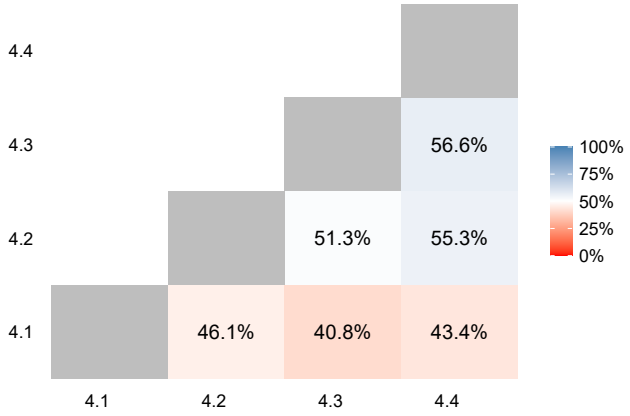

Figure 3: Heatmap displaying **Accountability (S4)**. Individual Standards: 4.1 Clear Responsibilities, 4.2 Responsible Output, 4.3 Contestability, 4.4 Transparent Information Provision.

Accountability (S4) Preferences for Accountability are more evenly distributed within a range of 41-59%, indicating that overall preferences did not lean strongly toward any particular individual standard. Interviews highlighted a general lack of clarity regarding the implementation of this standard, with participant P3 feeling unsure about where to find relevant information on the specific criteria, such as how to contest outputs (4.3) and how to identify who is responsible (4.1).

Transparency (S5) & Fairness (S6) Overall, participants placed more value on transparent information provision with respect to data handling, than other forms of information, such as system or output information. Figure 5 highlights this preference, with values of other exceeding 80% compared to other transparency information. Preferences were more evenly distributed for Fairness (see Figure 4), with the exception of heightened consensus for the importance of equal output (S6.6), i.e., that the quality of the LLM output remains consistent across different users.

Concerns and Wishes

Several participants raised additional concerns and wishes, focusing primarily on the perception of being passive subjects of LLMs rather than empowered consumers with a choice. P11 noted when asked to reflect on their priorities: *Well [...] I have no control anyway.* And further *When you receive information that [my data] will be deleted, no one can guarantee that. You need to have a [certification] label that you can trust this system.* P10 noted: *Ultimately, I am at the mercy of others and have to trust*. P4 noted: *How [will this] affect our lives in the next generations? If we become dependent, if we always ask ChatGPT in every area of life*

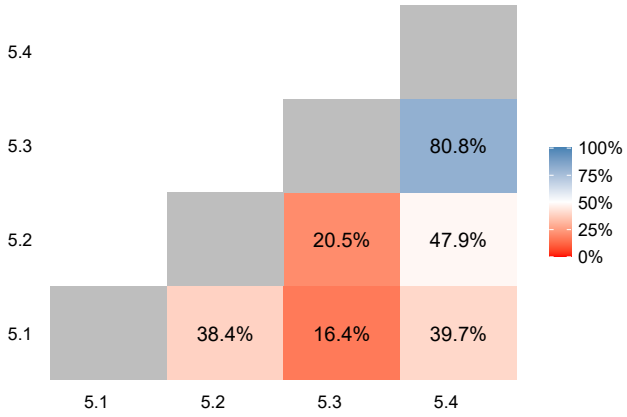

Figure 4: Heatmap displaying **Transparency (S5)**. Individual Standards: 5.1 Developer Information; 5.2 Output Information; 5.3 Data Handling; 5.4 System Information.

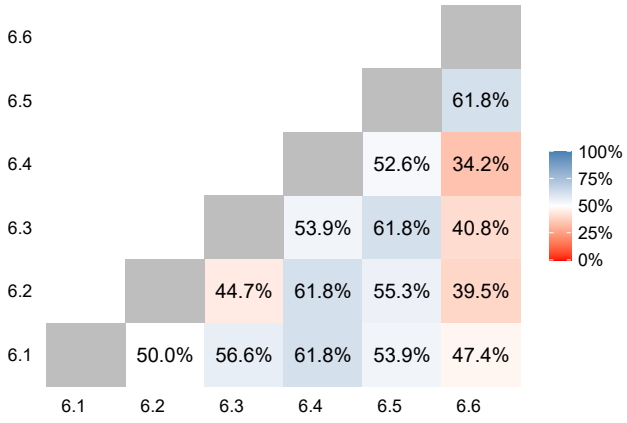

Figure 5: Heatmap displaying **Fairness (S6)**. Individual Standards: 6.1 Equal Access; 6.2 Accessibility; 6.3 Plurality of Opinion; 6.4 Data Curation; 6.5 Content Moderation; 6.6 Equal Output.

and never use our own heads. If future generations only rely on it in school and university. Constant dependency. Finally, P9 raised concerns about agentic LLMs, highlighting scenarios such as the following: *When asked to search for a hotel, the system immediately books one because it has access to my credit card information.*

The interviews further revealed that participants were most familiar with privacy and data standards, but showed knowledge gaps with respect to Accountability and Auditability, which users were largely uncertain about. P11 noted with respect to audits: *I can't really imagine what it's about... [...] Who controls the whole thing?* and further *What [...] are the criteria for checking these systems?* Some users shifted their perspectives after engaging in detailed discussions about the high-level standards. P5 noted: *Before the discussion I would have said explainability and privacy [are most relevant]. After the discussion – accountability.*

Discussion

Our qualitative findings revealing 68 individual standards indicate that users have detailed information requirements and underscore the subjective nature of aligning user values with trustworthy LLMs. For example, to assess whether the value *Accountability* (S4) is met, users may consider whether roles and duties of all parties involved — such as users, developers, and providers of the LLM — are explicitly defined (S4.1), whether the LLM provides truthful outputs (S4.2), whether there is an opportunity to challenge the LLM's output through a human point of contact (S4.3), and where to find information about the responsibilities of both users and LLM providers (S4.4). Additionally, some individual standards occur under several high-level standards and may prompt different questions. For example, the individual standards "data handling" satisfies the value *Privacy* by addressing the question, "Is my data shared or sold to third-party companies?", improving the users' sense of control over their data. However, it may fall short of meeting the value of *Transparency*, which raises the question, "How is my personal data processed and stored?", pertaining to visibility and users' understanding of data handling practices.

These findings collectively suggest that end-user expectations may not align with existing trustworthiness criteria, unless communicated effectively. Moreover, users consider "transparent information" a necessary prerequisite to verify various high-level standards. However, our findings indicate persistent skepticism about privacy and data handling, as users may feel that they lack the option to opt out of using LLMs. Furthermore, knowledge gaps around accountability and auditability contribute to user uncertainty about whom to trust in this landscape, making these areas critical for (regulatory) attention and underscoring the need to prioritize the development and clear communication of governance mechanisms for LLMs to foster a climate of trust.

Finally, our findings suggest that the majority of these standards cannot be effectively communicated through typical technical approaches or methodologies, such as explanations of model outputs, despite their frequent mention as *desiderata* for responsible AI (Langer et al. 2021; International Organization for Standardization 2020). As such, and in line with recent works (Ehsan et al. 2021; Hagendorff 2019), we advocate for a holistic approach that emphasizes not only the technical implementation of standards, but acknowledges the multifaceted evaluation process users face. This is especially important as LLMs obscure the boundaries between base model, adapted model and application itself (Liao and Vaughan 2023). To translate these results into practical guidance, we will briefly explore the cues and channels that can be utilized to bridge the knowledge gap for users when evaluating whether individual standards are met.

Communication Cues and Channels

Our findings reveal significant variability in end-user interpretations of high-level standards. Strategic alignment of user values with trustworthiness signals and cues necessi-

tates a taxonomy that outlines when and how users evaluate LLMs. Such a framework would allow to prioritize and tailor effective communication strategies. To contribute to this endeavour, we draw inspiration from the search, experience, and credence framework from information economics (Nelson 1970; Mitra, Reiss, and Capella 1999; Nelson 1970; Stiglitz 2000). This model categorizes goods based on the degree of information asymmetry between producers and consumers. We propose a similar classification for LLM standards, differentiating between standards that are readily verifiable through documentation (search-based individual standards), those that require user interaction (experience-based individual standards), and those that are challenging to assess even after use (credence-based individual standards) and thereby pose a greater risk for users. By distinguishing search, experience, and credence standards, developers, companies, and policymakers can tailor their communication strategies accordingly. We briefly outline the three types of standards and encourage future research to refine this preliminary taxonomy into a comprehensive classification system.

For *search-based standards*, users can easily assess whether an LLM aligns with their values based on its stated features. For instance, users that value Anonymity (1.4) can access this information through the application's declared policies or certifications provided by the technology provider. Available communication channels include, among others, published model documentation (Crisan et al. 2022) or the provider's terms and policies.

Experience-based standards emerge from real-time use and can be observed while users are engaging with the system. This category is of particular relevance to developers. For example, at the time of writing, ChatGPT's interface includes the disclosure 'ChatGPT can make mistakes. Check important info.', which addresses transparent information provision (S3.4) to fulfill *Security*, signaling users to verify the correctness of important information. Here, future research could compare the effectiveness of additional or alternative cues, such as communicating uncertainty within the interaction (Liao and Vaughan 2023).

For *credence-based standards*, information can be obtained neither before, nor during the interaction. They are therefore characterized by high uncertainty and should be the focus of strategies that aim to bridge knowledge gaps. For example, users who value fairness may not easily verify whether they are presented with unbiased and balanced viewpoints (S6.3). Here, communicating standards through other means, such as certification labels can shift users' trust from an unknown actor to a recognized third-party entity, such as a regulatory body (Scharowski et al. 2023). Developers could implement measures to shift safety standards from credence-based to experience-based standards, for example, by clearly specifying that LLM outputs may exhibit bias under certain conditions.

To conclude, the knowledge gaps arising from blurred boundaries between LLM models and downstream application are barriers that may foster a climate of mistrust – also described by Knowles and Richards (2021) as *negative facework*. This climate is already evident in users' perceptions of the widespread lack of transparency regarding their privacy. Understanding how effectively different channels help to communicate individual standards is a crucial step towards aligning user expectations with trustworthy LLM standards, thus contributing to the creation of *positive facework* and public trust.

Limitations and Future Work

This study has some limitations that we wish to address. While aiming for a representative sample, our focus on depth over breadth inherent in qualitative research resulted in a potentially non-representative interview group. In particular, we did not examine how various demographic groups might prioritize different standards. As our study is exploratory in nature, we aim to provide an initial understanding of end-user values and perspectives.

Additionally, the survey design only allowed for the comparison of individual standards within a single high-level standard. Future work could compare individual standards across different high-level standards and investigate how users respond in certain "trade-off scenarios" with conflicting standards. Furthermore, our study was designed to identify and prioritize user-relevant standards, and as such, it was confined to qualitative and survey-based methods. Future research could benefit from incorporating experimental methodologies to assess how various communication channels and cues affect users' perceptions of risk and ability to evaluate individual standards, for various tasks. Lastly, our primary focus was on conversational applications of LLMs, which are widely accessible. Future research should test the generalizability of our findings across a broader range of applications not specifically covered in this study.

Conclusion

As public concerns about the trustworthiness of LLMs grow, it is crucial to recognize that users may have diverse expectations or concerns about what makes an LLM application 'trustworthy', and that these concerns may not always align with formal criteria that have been adapted for value-aligned AI. Understanding and engaging with the public's interpretation of established standards may help bridge the knowledge gap between experts and users, and shape a narrative that reflects the diverse perspectives and values of society. This work aimed to contribute to the discourse on ethical AI development, by empirically investigating how users understand and prioritize high-level standards in the context of LLMs. Our findings, including strong preferences for governance, monitoring, and transparent information provision highlight the need for more research on strategic communication channels and cues that effectively communicate user-relevant standards for trustworthy LLMs.

References

Adler, R.; Alcala, A.; Anton, M.; Armbruster, T.; Arntzen, S.; Aschenbrenner, D.; Axt, K.-D.; Baumgartner, R.; Becker, N.; Beiter, R.; Bendig, T.; Benecke, R.; Benner, P.; Bernhardt, B.; Beyer, P.; Bich, K.; Biehler, J.; Bieringer,

L.; Binder, A.; and Zucker, B. 2022. Deutsche Normungsroadmap Künstliche Intelligenz Ausgabe 2.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N. S.; Chen, A. S.; Creel, K. A.; Davis, J.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N. D.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T. F.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M. S.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J. F.; Ogut, G.; Orr, L. J.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y. H.; Ruiz, C.; Ryan, J.; R'e, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K. P.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M. A.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; and Liang, P. 2021. On the Opportunities and Risks of Foundation Models. *ArXiv*, abs/2108.07258.

Brühlmann, F.; Petralito, S.; Aeschbach, L. F.; and Opwis, K. 2020. The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, 2: 100022.

Chinese National Governance Committee. 2019. Governance Principles for the New Generation Artificial Intelligence – Developing Responsible Artificial Intelligence. https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html2019. [Accessed 13-05-2024].

Chiou, E. K.; and Lee, J. D. 2021. Trusting Automation: Designing for Responsivity and Resilience. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 65: 137 – 165.

Constantinides, M.; Bogucka, E.; Quercia, D.; Kallio, S.; and Tahaei, M. 2024. RAI Guidelines: Method for Generating Responsible AI Guidelines Grounded in Regulations and Usable by (Non-)Technical Roles. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).

Crisan, A.; Drouhard, M.; Vig, J.; and Rajani, N. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 427–439. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

de Winter, J. C. F. 2023. Can ChatGPT Pass High School Exams on English Language Comprehension? *International Journal of Artificial Intelligence in Education*.

Ehsan, U.; Liao, Q. V.; Muller, M. J.; Riedl, M. O.; and Weisz, J. D. 2021. Expanding Explainability: Towards Social Transparency in AI systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

European Commission. 2019. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed: 2024-04-10.

Floridi, L. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, 1: 261–262.

Google. 2023. AI Principles Progress Update 2023. https://ai.google/static/documents/ai-principles-2023-progress-update.pdf. [Accessed 13-05-2024].

Hacker, P.; Engel, A.; and Mauer, M. 2023. Regulating ChatGPT and other Large Generative AI Models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.

Hagendorff, T. 2019. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30: 99 – 120.

Hu, K. 2023. ChatGPT sets record for fastest-growing user base - analyst note.

International Organization for Standardization. 2020. ISO/IEC TR 24028:2020 Information technology, Artificial intelligence, Overview of trustworthiness in artificial intelligence. Technical report, International Organization for Standardization. Available from https://www.iso.org/standard/77608.html.

Jakesch, M.; Buçinca, Z.; Amershi, S.; and Olteanu, A. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1–11.

Knowles, B.; and Richards, J. T. 2021. The Sanction of Authority: Promoting Public Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 262–271. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kästner, L.; Schmidt, E.; Sesing, A.; and Baum, K. 2021. What Do We Want From Explainable Artificial Intelligence (XAI)? - A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artif. Intell.*, 296: 103473.

Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; and Zhou, B. 2021. Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, 55: 1 – 46.

Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; and Zhou, B. 2023. Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, 55: 1 – 46.

Liao, Q.; and Sundar, S. S. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1257–1268. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Liao, Q. V.; and Vaughan, J. W. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *ArXiv*, abs/2306.01941.

Mayring, P.; and Fenzl, T. 2019. Qualitative Inhaltsanalyse. In Baur, N.; and Blasius, J., eds., *Handbuch Methoden der empirischen Sozialforschung*, (pp. 633–648). Springer VS, Wiesbaden.

Meade, A. W.; and Craig, S. B. 2012. Identifying careless responses in survey data. *Psychological methods*, 17(3): 437.

Microsoft. 2022. Microsoft Responsible AI Standard, v2. https://www.microsoft.com/en-us/ai/principles-and-approach. [Accessed 13-05-2024].

Mitra, K.; Reiss, M.; and Capella, L. M. 1999. An examination of perceived risk, information search and behavioral intentions in search, experience and credence services. *Journal of Services Marketing*, 13: 208–228.

Mittelstadt, B. D. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1–7.

Mokander, J.; Schuett, J.; Kirk, H. R.; and Floridi, L. 2023. Auditing large language models: a three-layered approach. *ArXiv*, abs/2302.08500.

Nelson, P. 1970. Information and Consumer Behavior. *Journal of Political Economy*, 78: 311 – 329.

Saxena, N. A.; Huang, K.; DeFilippis, E.; Radanovic, G.; Parkes, D. C.; and Liu, Y. 2018. How Do Fairness Definitions Fare?: Examining Public Attitudes Towards Algorithmic Definitions of Fairness. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.

Scharowski, N.; Benk, M.; Kühne, S. J.; Wettstein, L.; and Brühlmann, F. 2023. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.

Schlicker, N.; Baum, K.; Uhde, A.; and Langer, M. 2023. A Micro and Macro Perspective on Trustworthiness: Theoretical Underpinnings of the Trustworthiness Assessment Model (TrAM).

Schmude, T.; Koesten, L.; Möller, T.; and Tschiatschek, S. 2023. On the Impact of Explanations on Understanding of Algorithmic Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 959–970. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Sherman, E.; and Eisenberg, I. 2024. AI Risk Profiles: A Standards Proposal for Pre-deployment AI Risk Disclosures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23047–23052.

Stiglitz, J. E. 2000. The Contributions of the Economics of Information to Twentieth Century Economics. *Quarterly Journal of Economics*, 115: 1441–1478.

Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; and Ting, D. S. W. 2023. Large language models in medicine. *Nature Medicine*, 29: 1930 – 1940.

Toreini, E.; Aitken, M.; Coopamootoo, K. P. L.; Elliott, K.; Zelaya, C. V. G.; and van Moorsel, A. 2019. The relationship between trust in AI and trustworthy machine learning technologies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

UK House of Lords. 2017. Artificial Intelligence Committee AI in the UK: ready, willing and able? https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.html. [Accessed 13-05-2024].

Université de Montréal. 2018. The Montréal Declaration for a Responsible Development of Artificial Intelligence. https://montrealdeclaration-responsibleai.com/the-declaration/. [Accessed 13-05-2024].

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J. F. J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C.; Brown, S. M.; Kenton, Z.; Hawkins, W. T.; Stepleton, T.; Birhane, A.; Hendricks, L. A.; Rimell, L.; Isaac, W. S.; Haas, J.; Legassick, S.; Irving, G.; and Gabriel, I. 2022. Taxonomy of Risks posed by Language Models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Yeung, K.; Howes, A.; and Pogrebna, G. 2020. Governance by Human Rights–Centered Design, Deliberation, and Oversight: An End to Ethics Washing. In *The Oxford Handbook of Ethics of AI*. Oxford University Press. ISBN 9780190067397.

Yurrita, M.; Murray-Rust, D.; Balayn, A.; and Bozzon, A. 2022. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.