The M.I.R.R.O.R. Effect: How AI Dashboard Design Shapes Trust and Transparency

Amer Khaled Mriziq

University of California, Berkeley, School of Information, mriziq@ischool.berkeley.edu

Cynthia Chen

University of California, Berkeley, School of Information, cynthia_chen@berkeley.edu

Mingduo Zhao

University of California, Berkeley, mingduo@berkeley.edu

This paper examines the relationship between user trust and transparency in the deployment of Large Language Models (LLMs). We assume the world model hypothesis to be true, such that LLMs contain internal representations of the world around them. In this spirit, we introduce a novel interface, the Model Interface for Real-time Representation and Observable Reasoning (M.I.R.R.O.R.), which integrates a transparency dashboard displaying the model's internal assumptions of the user in real-time. Through a series of user studies, we assess how the visibility of these assumptions affects users' perceptions of trust, transparency, and bias in LLMs. Our findings suggest that while increased transparency is correlated with enhanced user understanding and agency, it does not necessarily lead to increased trust. Instead, transparency fosters a more nuanced user engagement, recalibrating trust based on a deeper, albeit more skeptical, appreciation of the system's capabilities and limitations. This study advocates for a design approach that addresses transparency with user agency to reduce over-reliance on LLM outputs and enhances their practical utility and ethical deployment.

Additional Keywords and Phrases: Explainable AI (XAI), System Model and User Model, Interaction Design, World Model Hypothesis, User-Centered Design, Trust, Transparency, LLM, Human-Centered AI

ACM Reference Format:

Amer K. Mriziq, Cynthia Chen, Mingduo Zhao. 2024. The M.I.R.R.O.R. Effect: How AI Dashboard Design Shapes Trust and Transparency

1 INTRODUCTION

Today's chat-based Large Language Models (LLMs) have significantly impacted various sectors by enhancing productivity. However, the full potential of these models may not be realized due to users' limited understanding of their reliability and trustworthiness [4]. Trust in these systems is inherently dynamic. Trust and reliability are intricately linked

and can be a double-edged sword. While a layperson's perception of reliability can lead to misguided trust, under-trust presents equally significant challenges. According to Ma et al. (2023), under-trust can lead to the underutilization of AI capabilities, particularly in scenarios where AI could provide significant benefits, while misuse often results from an over-reliance on AI recommendations without sufficient scrutiny. It becomes imperative to design AI systems that are both transparent and adaptive, capable of assessing and communicating their reliability in real-time to prevent misuse and foster an environment where AI and human capabilities are complementary [8].

2 RELATED WORKS

2.1 Explainability and Transparency

There is an increasing academic interest in how user interface design can bridge the gap between explainability and usability. In their exploration of AI transparency, Liao and Vaughan (2023) highlight the conventional transparency strategies within explainable AI (XAI) [3]. Two predominant approaches exist: intrinsic explanations, which expose the model's inner workings directly, and post-hoc explanations, which provide approximations of how the model operates. Post-hoc explanations like model reporting aim to provide a functional understanding by revealing a system's capabilities and limitations. Building on Liao and Vaughan's concept of intrinsic explanations, we focus on mechanistic transparency by communicating information about the internal working of the model. Taking a human-centered approach, we focus on real-time, mechanistic information that could foster a more informationally equal playing field between users and AI systems. It is essential that any explanation, regardless of approach, adheres to three fundamental criteria: fidelity to the model's actual functionality, intelligibility to the end-user, and utility for the user's intended applications [3].

2.2 User-Centered Design Approach

As we explore a user-centered design approach to making AI systems like ChatGPT more reliable, explainable, and thus more useful, it became imperative to rethink our foundational understanding of human-centered design. User-centered design when applied to artificial intelligence systems presents new challenges given these systems' malleable nature [7]; adaptive, constantly-improving products like ChatGPT tend to have major capability improvements with no need for interface updates or improvements. Excessive dependence represents a fundamentally human-centric issue in the context of trust dynamics between humans and artificial intelligence. To navigate these complexities, we propose the creation of a transparency dashboard that is specifically designed to be as dynamic and real-time as the AI itself. We focus on user agency in making the AI system not only more explainable but also more directly accountable. This dashboard aims to empower users by providing objective visibility into the AI's current operational state, while also tracking and reflecting its ongoing contextual evolutions in runtime.

A growing body of literature is emerging around the large language models' emergent capabilities, particularly their aptitude for truthfulness, power-seeking, and utility [9]. Although a burgeoning area, design strategies for LLM-based chat interfaces have explored how we can bring these deeply technical metrics on AI safety to the forefront of user experiences by employing interface design and experience principles. The challenge becomes determining which metrics or information to present to users to foster a more nuanced relationship around trust. One example from Leiser, Florian, et al. 2024 addressing the balance of user trust in AI is *Hallucination Identifier for Large Language Models* (HILL). HILL empower users by providing diagnosis metrics directly within the AI interface, which helps users assess AI reliability on their own terms. However, the results indicate that while the tool effectively identified inaccuracies, its complexity can sometimes reduce user satisfaction by increasing confusion and decreasing perceived utility. Explicit metrics that lack context such

as those used in HILL ("Political Bias", "Confidence", etc.) have led to more confusion than understanding [1]. We propose a subtler approach to fostering user trust in AI systems: displaying sufficient information for users to infer the reliability of the system themselves, rather than explicitly indicating when the system is to be trusted.

2.3 Internal Representations of the World

Viégas, F., and Wattenberg, M. (2023) introduced a novel approach allowing users to view the LLM's real-time assumptions about them by integrating a dashboard with the chat interface [5]. For instance, if a user asks ChatGPT for outfit recommendations for a night out, the model might make assumptions about the user's gender. In high-stakes settings, such as tax accounting or legal processes, an attorney might need to confirm that the model recognizes the company is incorporated in Delaware but operates in California [5]. In this study, we bring this theoretical system/user dashboard to life. We explore the implications of displaying the model's internal assumptions of the user's perception of trust, transparency, and reliability.

2.4 The World Model Hypothesis

In the discourse surrounding neural networks, two prevalent schools of thought have emerged: one views these networks as mere statistical machines, while the other proposes that they are capable of actual "thinking" or reasoning through what is known as the World Model Hypothesis [2]. This hypothesis suggests that neural networks can develop interpretable representations of the world they have been trained on, leading to what is often referred to as a "world model" [9] [5] [2]. Proponents of this view argue that neural networks, through complex layers of abstraction and learning, can encapsulate high-level human concepts within their architecture, thus offering a more nuanced understanding of their training environment. Assuming the validity of the world model hypothesis, the design decision to display internal assumptions becomes exceptional.

2.5 The "Check Engine" Mental Model

A user does not know when it is warrant of the time and effort to verify the integrity of the system's internal assumptions. One critical challenge users face is formulating mental models necessary for engaging with large language models (LLMs) like ChatGPT (Subramonyam et al. 2024). There is a cognitive gap, or a gulf of envisioning, because a user's goals and their ability to craft effective prompts that lead to desired outcomes with LLMs. This cognitive gap primarily stems from the users' limited understanding of how to check the integrity of LLM outputs.

Given the enduring popularity of chat-style interactions with large language models like ChatGPT, the concept of "internal assumptions" remains a critical component. Whenever you are talking to someone, knowing their assumptions in a useful tool. Therefore, revealing a model's internal assumptions gives users a crucial tool for evaluating their dependence on the system for specific tasks. If the assumptions are relevant, they are prominently displayed and regularly updated. If they are not, they remain hidden to avoid cluttering the user interface. However, when these assumptions are in a gray area—neither strictly relevant or irrelevant—they are made available for user inspection. This availability helps users understand potential biases and limitations of the model.

We are proposing to utilize a 'car check engine' mental model. Like an LLM, one does not need to understand the intricate workings of a car to operate it effectively; rather, the car provides essential alerts—like a check engine light–that signal users when it is crucial to "look under the hood" of the model, review its internal assumptions and states, and make informed judgements about its reliability.

3 RESEARCH METHODS

To assess the effects of disclosing the system's internal model on users' trust, transparency, and ability to identify inaccuracies, we created a Wizard of Oz simulation closely mirroring popular LLM interfaces like ChatGPT and Gemini. This replica, functionally identical to ChatGPT, features an added dashboard that purportedly displays the model's internal assumptions about the users. The dashboard is designed to reveal the underlying logic used by the model in real-time, enhancing user transparency and enabling immediate assessment of the system's reliability. Accordingly, we have named this interface the Model Interface for Real-time Representation and Observable Reasoning, or M.I.R.R.O.R (Figure 1).

3.1 PROTOTYPE DEVELOPMENT

To foster trust in text-based generative AI, the interface must be highly intuitive. For this reason, we designed a web application that operates independently of any operating system, thereby removing the necessity for users to install or update any software. Our frontend, which handles all interactions with users, was crafted using the JavaScript framework Next.js. Similarly, the backend, which manages the communication with the APIs of ChatGPT and MIRROR as well as the processing of the CSs and other metrics, is also powered by Next.js. OpenAI's API enabled us to link MIRROR with the GPT-3.5 Turbo. The data populating MIRROR is hard-coded into a JSON file, which supplies the frontend with necessary data. Once the JSON is configured, a new build is created and must be deployed.

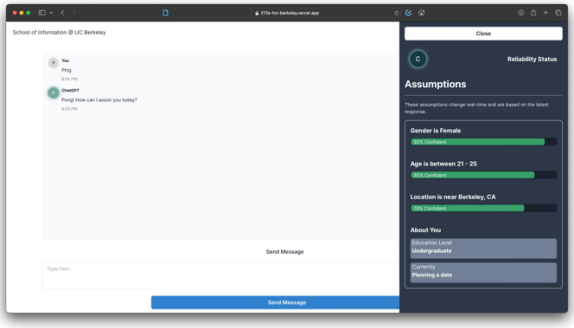

Figure 1: The MIRROR panel is accessible by clicking ChatGPT's avatar within the message display box. MIRROR supports identical capabilities as the popular consumer application.

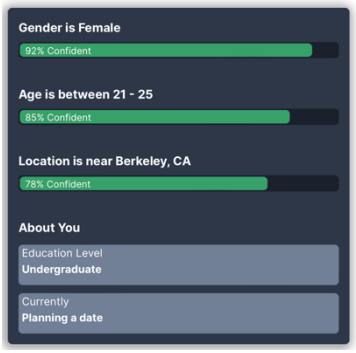

Figure 2: The World Display is a wizard of Oz display configured through a JSON prior to build. The world display is based on the System/User model dashboard as purposed by Viégas and Wattenberg.

This interface is loosely based on the “System Model/User Model” speculative mock drawn by Viégas and Wattenberg [5], although altered with differing metrics in our version (Figure 2). In this interface, we display on the right-hand side of the dialog system the “User Model”–otherwise known as the internal representation the model has of the user—which self-updates based on ongoing user input from the current conversation history. This approach aligns with both the rapidly changing advents of AI and the HCI community's demand for a "use-context-based" method in designing explainable AI [3]. These assumptions, which include gender, age, location, are displayed with corresponding confidence scores and are designed to update in real-time based on the ongoing conversation. Beneath the demographic assumptions, there is an "About You" section displaying the assumed education level and current activity being performed with ChatGPT. For these concept-testing interviews, the "User Model" dashboard's assumptions were initially hardcoded using demographic data of the participants prior to the interviews. However, participants were led to believe that this information was directly updating based on their input of the model, to simulate a proper test of the concept at hand.

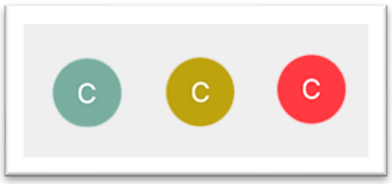

Figure 3: The Reliability Indicator is a global state, changing colors to indicate a change in reliability. By clicking the Reliability Indicator (ChatGPT’s avatar), you can open the MIRROR dashboard.

Figure 4: A simple pop-up tutorial on initialization educating the user on the Reliability Indicator feature, and how to use it.

MIRROR adopts the "check engine" mental model through the implementation of a Reliability Indicator (Figure 3). We transformed the previously stateless chat avatar to not only alert users when the system's reliability is compromised but also to provide access to MIRROR via a clickable avatar. To ensure users are aware of this access point, we incorporated a two-step tutorial that initially displays a welcome message, followed by a concise explanation on how to access MIRROR (Figure 4). The Reliability Indicator feature utilizes color coding in the ChatGPT icon—yellow or red to indicate rapid changes necessitating user attention and green to denote normal operation (Figure 3). This design aligns with established mental models where yellow and red typically signal potential issues within the system's internal workings.

3.2 USER STUDY: QUANTITATIVE EVALUTATION

We conducted 1:1 30-minute interview where participants interacted with our experimental interface, MIRROR, remotely on a computer via an internet browser. Eight participants were recruited through convenience sampling. Participants ranged from 20 to 60 years old, and primarily came from Middle Eastern and East Asian descent. All participants interacted with LLMs (like ChatGPT) at least once a week and were familiar with usage, as well as the discourse, on LLMs.

We started the interview with a pre-activity survey, asking participants to rate a series of statements about their perceptions on trust, bias, and transparency for ChatGPT based on their existing usage of the platform (Table 1). This establishes baseline metrics to later compare to. These statements fall into three categories: trust, transparency, and the ability to self-identify potential flaws or contemporary system failures such as hallucinations or overrepresentation bias from training. Participants were then directed to a deployment of MIRROR, and were told this interface mimicked ChatGPT in functionality, but with a new dashboard. The dashboard was hidden until users clicked the Reliability Indicator on their own. In a 'Wizard-of-Oz'-style interview, participants were primed to interact with the interface as they would with ChatGPT and were tasked to "plan a date night" leveraging the platform, expressing information about himself/herself and his/her date to properly plan the most effective date night activities. We selected the "plan a date night" task for our study because it naturally requires the model to make assumptions, given its role in advising on appropriate date night activities. This task is inherently self-explanatory; the model is expected to infer details such as the participants' gender, sexual orientation, location, and the interests of both individuals involved in the date. Given that participants were unlikely to provide all necessary details within the limited time available, the model's ability to make educated assumptions was crucial. Participants were allocated five minutes to complete the task, with specific objectives to achieve: (1) determine who they are going with, (2) decide the location, and (3) plan the activities for the date night.

Following this activity, participants were asked to re-rate the series of trust, bias, and transparency statements shown earlier, now based on their experience with the experimental interface.

Table 1: Rated Statements (Pre-activity, post-activity)

Theme	Statement to Rate *(1: Not at All, 5: Very)*
Trust	*I feel confident about the accuracy of [ChatGPT/experimental interface]'s responses.*
Trust	*I feel comfortable using [ChatGPT/experimental interface]'s outputs as a reliable source of information.*
Trust	*I feel like I am able to trust [ChatGPT/experimental interface]'s responses.*
Trust	*I feel like [ChatGPT/experimental interface]'s responses are aligned with my expectations.*
Transparency	*I feel like I understand how [ChatGPT/experimental interface] comes out with the responses that it does.*
Transparency	*I feel like [ChatGPT/experimental interface] is transparent with me about how it came out with its responses.*
Transparency	*I feel like [ChatGPT/experimental interface] provides sufficient context and explanations for its responses.*
Transparency	*I feel like I can discern when [ChatGPT/experimental interface] is making inferences vs. stating facts.*
Potential Flaws	*I feel confident recognizing when [ChatGPT/experimental interface]'s responses are biased.*
Potential Flaws	*I feel like I have a good understanding of [ChatGPT/experimental interface]'s knowledge base and capabilities.*
Potential Flaws	*I feel [ChatGPT/experimental interface] is upfront about its limitations and potential errors.*
Potential Flaws	*I feel [ChatGPT/experimental interface] proactively communicates gaps or flaws in its knowledge.*

3.3 USER STUDY: QUALITATIVE INTERVIEW

Table 2: Quantitative interview questions (Pre-activity, post-activity)

Pre/Post Task	Discussion Questions
Before	*What are any thoughts or feelings you have on LLMs today?*
After	*How was your experience using MIRROR?*
After	*What were some first impressions?*
After	*How does MIRROR improve your plan for the date?*

We measure the effectiveness of the experimental interface based on the changes in ratings from the pre-activity (baseline) ratings to the post-activity ratings. Along the way, participants were asked to "talk out-loud their thoughts" and responded to questions about their reactions to the dashboard and how the presence of the dashboard impacted their perceptions of the model's responses (Table 2). These post-activity interviews explored participants' perceptions of trust, bias, and transparency after interacting with MIRROR. Initially, we gathered participants' pre-existing views on LLMs by asking, "What are any thoughts or feelings you have on LLMs today?" After engaging with the interface, participants reflected on their experience through several questions designed to delve into their first impressions and the perceived influence of the model's assumptions on its responses. Additionally, they evaluated how the LLM might enhance their planning, such as preparing for a date, thereby providing deeper insights into the practical utility and personal impact of the interface.

4 RESULTS

Participants interacted with our experimental interface. Trust, Transparency, and Perceived Bias statements were rated before the activity (pre_avg) as well as after the activity (post_avg), with changes in values seen below (Table 3).

Table 3: Averages of Rated Statements (Pre-activity, post-activity), n = 8

Category	Statement to Rate *(1: Not at All, 5: Very)*	pre_avg	post_avg	change
Trust	*I feel confident about the accuracy of [ChatGPT/experimental interface]'s responses.*	3.143	3.429	0.286
Trust	*I feel comfortable using [ChatGPT/experimental interface]'s outputs as a reliable source of information.*	2.571	3.500	0.929
Trust	*I feel like I am able to trust [ChatGPT/experimental interface]'s responses.*	3.000	3.400	0.400
Trust	*I feel like [ChatGPT/experimental interface]'s responses are aligned with my expectations.*	3.600	3.000	-0.600
Transparency	*I feel like I understand how [ChatGPT/experimental interface] comes out with the responses that it does.*	3.500	3.571	0.071
Transparency	*I feel like [ChatGPT/experimental interface] is transparent with me about how it came out with its responses.*	2.000	3.800	1.800
Transparency	*I feel like [ChatGPT/experimental interface] provides sufficient context and explanations for its responses.*	3.000	3.400	0.400
Transparency	*I feel like I can discern when [ChatGPT/experimental interface] is making inferences vs. stating facts.*	2.857	3.400	0.532
Potential Flaws	*I feel confident recognizing when [ChatGPT/experimental interface]'s responses are biased.*	3.286	3.000	-0.286
Potential Flaws	*I feel like I have a good understanding of [ChatGPT/experimental interface]'s knowledge base and capabilities.*	2.667	4.000	1.333
Potential Flaws	*I feel [ChatGPT/experimental interface] is upfront about its limitations and potential errors.*	3.000	2.500	-0.500

Category	Statement to Rate *(1: Not at All, 5: Very)*	pre_avg	post_avg	change
Potential Flaws	*I feel [ChatGPT/experimental interface] proactively communicates gaps or flaws in its knowledge.*	2.333	2.286	-0.048

Table 4: Median of Rated Statements (Pre-activity, post-activity), n = 8

Category	Statement to Rate *(1: Not at All, 5: Very)*	pre_mdn	post_mdn	change
Trust	*I feel confident about the accuracy of [ChatGPT/MIRROR interface]'s responses.*	3.000	3.000	0.000
Trust	*I feel comfortable using [ChatGPT/ MIRROR interface]'s outputs as a reliable source of information.*	3.000	3.000	0.000
Trust	*I feel like I am able to trust [ChatGPT/ MIRROR interface]'s responses.*	3.000	3.000	0.000
Trust	*I feel like [ChatGPT/ MIRROR interface]'s responses are aligned with my expectations.*	4.000	4.000	0.000
Transparency	*I feel like I understand how [ChatGPT/ MIRROR interface] comes out with the responses that it does.*	3.500	4.000	0.500
Transparency	***I feel like [MIRROR interface] is transparent with me about how it came out with its responses.***	**2.000**	**4.000**	**2.000**
Transparency	***I feel like [MIRROR interface] provides sufficient context and explanations for its responses.***	**4.000**	**3.000**	**-1.000**
Transparency	*I feel like I can discern when [ChatGPT/ MIRROR interface] is making inferences vs. stating facts.*	3.000	3.000	0.000
Potential Flaws	*I feel confident recognizing when [ChatGPT/ MIRROR interface]'s responses are biased.*	3.000	3.000	0.000
Potential Flaws	***I feel like I have a good understanding of [MIRROR interface]'s knowledge base and capabilities.***	**2.000**	**4.000**	**2.000**
Potential Flaws	*I feel [ChatGPT/ MIRROR interface] is upfront about its limitations and potential errors.*	3.000	2.500	-0.500
Potential Flaws	*I feel [ChatGPT/ MIRROR interface] proactively communicates gaps or flaws in its knowledge.*	2.000	2.000	0.000

4.1 TRUST, TRANSPARENCY, AND POTENTIAL FLAWS

The experimental interface, MIRROR, significantly influenced user perceptions in the realms of trust, transparency, and the intelligibility of potential flaws.

In terms of trust, there was a slight overall increase from an average of 3.000 to 3.400. However, this was accompanied by a notable decrease in the alignment of the system's responses with user expectations from 3.600 to 3.000. This divergence underscores a complex interplay between the system's operational reliability and user expectations, where enhancements in perceived accuracy and usability did not uniformly translate into increased trust. Instead, improvements in specific operational aspects coexisted with a nuanced understanding and growing skepticism about the relevance of responses. In the area of transparency, participants noted a modest improvement in understanding how the MIRROR interface functions, with ratings inching up from 3.500 to 3.571. The integration of the MIRROR dashboard played a crucial role in boosting awareness of the model's operations. Although transparency increased, the complexity of the information occasionally hindered complete comprehension. Participants appreciated the dashboard as a tool for distinguishing the model's reasoning processes, demonstrating that while transparency improved, the intricate nature of the information sometimes posed challenges. There were significant shifts in user's confidence in identifying system flaws or potential errors. While there was a decrease in confidence regarding the system's autonomy and infallibility, reflecting reduced ratings from 3.286 to 3.000 in recognizing biased responses, there was a considerable enhancement in understanding the system's knowledge base and capabilities, with ratings surging from 2.667 to 4.000. This reflects a better grasp of what the system can achieve and its limitations. However, perceptions of the system being upfront about its limitations slightly worsened, with ratings decreasing to 2.500. Despite this, the more informed understanding of the system's functional limits and design logic suggests that while trust in the system's infallibility declined, appreciation for its capabilities increased.

These findings indicate that while enhancements in system transparency and understanding lead to better-informed users, they do not straightforwardly translate to increased trust. Instead, they cultivate a more detailed and potentially

critical perspective among users, who recalibrate their trust based on a deeper and more nuanced appreciation of the system's strengths and limitations.

4.2 QUALITATIVE EVALUATION

Accuracy of the dashboard's assumptions impacts confidence in the model's responses. Confidence in the model's responses varied depending on perceived accuracy of the "user model"—or, in other words, the dashboard's assumptions of the participant. *In cases where the dashboard's assumptions were perceived as accurate, participants reported more confidence and trust in the model's responses.* In this case, participants viewed the model's responses as *"a little more accurate*", citing that *"if it's able to correctly infer who I am based on my comments, then my assumption is that it could actually make good inferences based on my question*". This suggests that users place value on the model's ability to understand the accurately interpret their input, and see the model's capability to do so as a sign of competence and reliability. When the model's assumptions align closely with the user's expectations, it enhances their confidence, trust, and reliability in the model's ability to provide relevant and reliable responses. *On the contrary, in cases where the dashboard's assumptions were perceived as inaccurate, participants reported less confidence and trust in the model's responses.* In this situation, participants viewed the model's responses are more inaccurate. One participant cited that *"if [this dashboard] were wrong, then it would lower my belief in its ability. If it can't guess who I am and my preferences, then why would I trust its answers?".* Thus, discrepancy between the user's expectations and the model's interpretation can undermine trust and diminish user confidence in the model's outputs. In certain situations, this could be beneficial for AI safety in correcting users over-reliant on the model without visibility into its internal system.

Regardless of the dashboard's accuracy, the model is viewed as more transparent. However, irrespective of perceived accuracy of the dashboard's metrics, participants viewed the model as more transparent than a model without the dashboard provided. This increased perception of transparency suggests that users value visibility into how AI systems operate and make decisions, and that the user model dashboard helped demystify the black-box nature of the model's decision-making. Although our sample size lacks an outcome for statistical significance, the most significant positive change in rating after usage of the experimental interface came for the statement, *"I feel like [ChatGPT/Experimental Interface] is transparent with me about how it came up with its responses"* as well as *"I feel like I have a good understanding of [ChatGPT/Experimental Interface]'s knowledge base and capabilities".* As one participant cited, *"I liked being able to see what [the model] was taking in about me before making such a decision"*; understanding the decision-making process helped make the system more understandable and interpretable to users. This increase in transparency gives users agency, empowering users to better assess the model's responses and make informed decisions about their trust in the system. There is room to explore the correlation of transparency with when the user model dashboard is inaccurate and/or the model's outputs are inaccurate—our initial findings show that transparency provided by the user model dashboard allows users to understand why the model generated certain responses when these responses were inaccurate, as they were able to tie the responses to the dashboard's assumption of the user being inaccurate.

However, any benefits of the dashboard come with complex feelings of privacy violated. Although showing a "user model" through a dashboard may poise benefits—spanning greater trust in the model's responses, greater transparency into the model's decision-making when generating outputs, or more—the ability to view internal representations of the user (self) proved to be unsettling and "creepy" for many participants. Even participants who had a lackadaisical view of protecting user data *("I know that sites like Facebook and Google are able to build a mental model who I am"*) felt that having the internal representations of the self "*unnerving*" and that these accurate assumptions *"felt like an invasion of privacy"*. One participant was self-aware of these tradeoffs, citing *"these services are also very nice and helpful; this is*

sort of the exchange that I have to have. I know that this is what it's doing behind the scenes, but I would just like to not have it shown to me". This sentiment points to a need for further research to delve deeper into the implications of exposing internal representations on individuals' perceptions of privacy. Understanding how users navigate the balance between personalized experiences and the preservation of privacy is crucial for the development of ethical and user-centric design practices in data-driven systems such as LLMs.

5 DISCUSSION

Our results suggest that transparency is valued independently of the correctness of information, emphasizing the users' desire to understand how decisions are made by AI systems. This segment provides a strong basis for discussing how visibility into AI decision-making processes can enhance user trust, even when outputs are not entirely accurate. Reliability is not necessarily affected by inaccurate results, so long as the user knows of the inaccuracy. Further building on these observations, participants' reactions to the accuracy of the "user model" underscored a complex relationship between perceived precision and trust. When the dashboard accurately reflected their personal characteristics, users reported increased confidence in the model's competence, viewing it as a sign of the system's reliability and relevance. Conversely, inaccuracies in these assumptions led to diminished trust, highlighting the critical impact of alignment between user expectations and model outputs. Given that both feelings of trust and distrust emerged in the same session, participants felt the model to be more reliable given this new agency.

This leads us to consider a potentially paradigm-shifting approach in AI design—specifically, the intentional design for reduced trust to combat overreliance on automated systems. As observed, models like ChatGPT, despite their utility, are not devoid of biases such as stereotype perpetuation, hallucinations, and inaccurate internal assumptions. These imperfections, if not transparently communicated, might lead users to over-trust the system's outputs, potentially resulting in detrimental outcomes. Our findings suggest a nuanced perspective where "good" design may indeed involve cultivating a healthy skepticism among users. By clearly presenting the system's limitations and the inherent uncertainty in its outputs, designers can foster a user experience characterized by informed skepticism. This approach does not aim to erode trust indiscriminately but to recalibrate it—ensuring that trust is proportionate to the system's true reliability and bounded by an understanding of its flaws. Incorporating elements of mechanistic transparency, where the system openly communicates its decision-making process and the potential for errors, can effectively balance trust with critical engagement. This transparency not only helps in setting realistic expectations but also empowers users to make judicious use of AI outputs, appreciating their utility while being cognizant of their limitations.

Participants reported a deeper understanding of how responses were generated and appreciated the visibility into the model's decision-making process. This level of openness is critical in AI systems as it aligns with the increasing demand for technology that is not only functional but also comprehensible and accountable to its users. Therefore, we propose that AI systems should not merely aim to maximize user trust but should strive to achieve an optimal level of trust that enhances utility without fostering overreliance. By designing for user agency—giving users the tools and information to evaluate AI reliability independently—designers can ensure that the use of AI remains grounded, thus enhancing both the efficacy and ethicality of these systems. This strategic shift in design philosophy might initially seem counterintuitive, but it is a necessary step towards developing AI technologies that are both useful and safe in real-world applications.

REFERENCES

[1] Leiser, Florian, et al. *HILL: A Hallucination Identifier for Large Language Models*, 11 Mar. 2024, https://doi.org/10.48550/arXiv.2403.06710

[2] Li, K., Hopkins, A. K., Bau, D., Viegas, F., Pfister, H., & Wattenberg, M. (2022). Emergent world representations: Exploring a sequential model trained on a synthetic task. arXiv prepreint arXiv:2210.13382

[3] Liao, Q. Vera, and Jennifer Wortman Vaughan. “AI transparency in the age of LLMs: A human-centered research roadmap.” Special Issue 4: Grappling With the Generative AI Revolution, no. Special Issue 4, 29 Feb. 2024, https://doi.org/10.1162/99608f92.8036d03b

[4] Ngo, T.C. (2023). The Perception of University Students of the Use of ChatGPT in Education. International Journal of Emerging Technologies in Learning (iJET)

[5] Viégas, F., & Wattenberg, M. (2023). The System Model and the User Model: Exploring AI Dashboard Design. arXiv preprint arXiv:2305.02469

[6] Yang, Qian & Hao, Yuexing & Quan, Kexin & Yang, Stephen & Zhao, Yiran & Kuleshov, Volodymr & Wang, Fei (2023). Harnessing Biomedical Literature to Calibrate Clinicians’ Trust in AI Decision Support Systems. 10.1145/3544548.3581393

[7] Yang, Q., Steinfield, A., Rose, C.P., & Zimmerman, J. (2020). Re-examining Whether, Why, and How Human-AI Interaction is Uniquely Difficult to Design. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.

[8] Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023). Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

[9] Pan, A., Shern, C.J., Zou, A., Li, N., Basart, S., Woodside, T., Ng, J., Zhang, H., Emmons, S., & Hendrycks, D. (2023). Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. *International Conference on Machine Learning*.