Safety Perceptions of Generative AI Conversational Agents: Uncovering Perceptual Differences in Trust, Risk, and Fairness

Jan Tolsdorf ⋆, Alan F. Luo ◇, Monica Kodwani ⋆, Junho Eum ⋆,
Mahmood Sharif ◁, Michelle L. Mazurek ◇, Adam J. Aviv ⋆

⋆ *The George Washington University,* ◇ *University of Maryland, College Park,* ◁ *Tel Aviv University*

Abstract

Public and academic discourse on the safety of conversational agents using generative AI, particularly chatbots, often centers on fairness, trust, and risk. However, there is limited insight into how users differentiate these perceptions and what factors shape them. To address this gap, we developed a survey instrument based on previous work. We conducted an exploratory study using factor analysis and latent class analysis on survey responses from $n = 123$ participants in the U.S. to offer an initial attempt at measuring and delineating the dimensionality of these safety perceptions. Latent class analysis revealed three distinct user groups with sometimes counterintuitive perception patterns: The Hesitant Skeptics, The Cautious Trusters, and The Confident Adopters. We find that greater usage frequency of AI chatbots is associated with higher trust and fairness perceptions but lower perceived risk. Some demographic traits like sexual orientation, income, and ethnicity also had strong and significant effects on group membership. Our findings highlight the need for more refined measurement approaches and a more nuanced perspective on users' AI safety perceptions regarding trust, fairness, and risk, particularly in capturing the kinds of experiences and interactions that lead users to develop their perceptions.

1 Introduction

Public generative AI chatbots—such as ChatGPT [57], Gemini [21], and DeepSeek [14]—have become the fastest-growing online services in history [30, 59]. Their rapid adoption has sparked debate about their societal and individual impact, particularly around safety risks (e.g., privacy, security, manipulation), fairness, and bias [20, 31, 36–38, 46, 55, 74].

Concerns about privacy, unfair treatment, biased outputs, and other issues can undermine trust in these systems [15, 18, 27]. Yet even skeptical users often continue using them [47, 78], highlighting a complex interplay between perceived utility, risk, fairness, and trust [4, 37]. Despite growing interest in understanding public responses to generative AI chatbots, existing research remains fragmented and largely issue-specific. Many studies focus on particular risks [19, 31], demographic subgroups [6, 22], or specific chatbot features [1]. Moreover, they are often conducted in highly framed or artificial settings that may not reflect how the general population experiences these systems in everyday use [18, 27, 78].

While these studies offer valuable insights, we argue for a broader perspective—one that identifies dominant patterns in how the general population perceives whether AI chatbots produce risky outputs, act unfairly, or can be trusted in everyday use. This perspective moves beyond isolated concerns and draws on behavioral models from privacy, security, and technology adoption research, which show how perceptions shape user concerns, intentions, and behavior [4, 64]. However, advancing this work also requires addressing the prevailing lack of established instruments for measuring safety-related perceptions in the context of generative AI [3, 54, 66].

To address this challenge, we conducted an online survey with 123 U.S.-based users to quantitatively assess safety perceptions of generative AI chatbots. Given that risks are multifaceted [74] and that trust and fairness are multidimensional constructs [3, 54, 66], we combined adapted measurement instruments from prior research with newly developed items tailored to concerns specific to generative AI. Through exploratory factor analysis, we identified core subdimensions of chatbot safety perceptions. We further used latent class analysis [56] to uncover distinct user groups based on their perceptions. To this end, we make the following contributions:

(1) Participants significantly differentiated between discrimination, misinformation, and offensive language as distinct risk factors in AI chatbot interactions. Overall, these

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2025.
August 10–12, 2025, Seattle, WA, United States.

risks are perceived as unlikely, with the highest risk being misinformation and unhelpful outputs.

(2) In terms of trust, opinions varied, and, on average, participants neither fully trusted nor distrusted AI chatbots.

(3) Using an adapted fairness scale from organizational research, we find that participants distinguish between decision-making integrity and politeness when evaluating fairness in AI chatbots. Participants largely agreed that AI chatbots uphold these fairness aspects. However, perceptions of fairness appeared to be strongly affected by trusting chatbots to act in one's personal interests and beliefs about misinformation.

(4) We identified three user groups with distinct perceptions of fairness, trust, and risk: (i) The Hesitant Skeptics (high-risk, low-trust, low-fairness); (ii) The Cautious Trusters (medium-risk, high-trust, high-fairness); and (iii) The Confident Adopters (low-risk, high-trust, high-fairness). These groups exhibit overlapping perceptions on some factors, but diverge on others. AI chatbot usage frequency emerged as the most significant differentiator. Frequent users were more likely to align with the Confident Adopters, whereas infrequent users were more likely to align with the Hesitant Skeptics.

Our findings underscore the value of systematically capturing user perceptions in a generalizable manner, enabling the identification of nuanced differences in how various user populations perceive AI chatbots and their associated safety concerns. Usage frequency of AI chatbots was a key differentiator between user groups, with the most frequent users showing the highest trust and fairness perceptions along with the lowest risk concerns. Our results also highlight the need for future research and the development of new measurement tools for emerging contexts like generative AI conversational agents. In particular, we illustrate the difficulty and importance of developing measurement instruments that take into account sub-constructs of fairness, trust, and risk perceptions.

2 Background

Extensive research in human-computer interaction (HCI) and human-AI interaction (HAI) has explored AI-enabled systems, often defining, conceptualizing, and operationalizing trust, fairness, and risk to align with the specific objectives of individual studies. Systematic literature reviews (SLRs) have highlighted that these definitions are highly context-dependent [3, 54, 66, 74]. In the following, we provide an overview of these aspects regarding the study of trust, fairness, and risk in AI more generally and for generative AI conversational agents in particular.

2.1 User Safety Perceptions of AI

Trust Perceptions Trust in AI has been examined in diverse settings, ranging from video game agents [41] to healthcare [42], recommender systems [65], and algorithmic decision-making more broadly [32, 62, 63]. Trust is shaped by multiple factors, including socioethical considerations, system design, and individual characteristics [3]. Trust is typically conceptualized through dimensions such as perceived competence, reliability, integrity, and benevolence of the AI system, as well as user vulnerability, with user trust being influenced by system transparency, explainability [49, 68], ethical and value alignment, and situational factors [3, 54]. Trust in AI systems is predominantly measured through surveys, with many studies relying on ad hoc questionnaires [3]. Two notable exceptions are the validated 12-item Human-Computer Trust Scale (HCTS) [25] and the 19-item scale developed by Körber [40]. The HCTS offers a theoretically grounded measure for trust in "intelligent systems" more broadly, while Körber's scale places stronger emphasis on trust in automation. Beyond self-report instruments, trust has also been assessed through behavioral indicators, physiological responses (e.g., stress levels), and qualitative methods including interviews and focus groups [3, 54].

Fairness Perceptions Fairness in AI research has predominantly focused on algorithmic fairness, aiming to mitigate unjust, discriminatory, or disparate outcomes in AI-driven decision-making [51, 66]. Technological approaches define fairness through probabilities, logic, and mathematical principles [11, 23, 44, 69, 75], often overlooking human-centered perspectives. Beyond formal definitions, fairness is also studied as a human perception, shaped by individual intuition, affective responses, and social norms [66]. Perceived fairness is typically examined through four dimensions: distributive fairness (fairness of outcomes), procedural fairness (fairness of decision-making processes), interpersonal fairness (respectful and unbiased treatment), and informational fairness (transparency in decision-making) [51, 66]. However, substantial variability exists in the application of these dimensions, with definitions differing across contexts, particularly regarding individual versus group fairness [34, 38, 48].

To measure perceived algorithmic fairness, studies commonly expose participants to an algorithmic decision-making process or outcome and then prompt them to rate its fairness [51, 66]. Another common approach to measuring perceived fairness is based on organizational justice research [10], adapting its original measurement instrument to the AI context [66]. Such scales have been applied across domains, including education and AI-driven hiring [2, 32, 48, 67, 72]. Other studies assess fairness indirectly by examining participant preferences for different AI-mediated processes [66].

Risk Perceptions Research on AI risk perceptions is closely linked to fairness and trust [38, 50, 53, 60, 61, 70, 71, 76]. Much of this work focuses on bias and its consequences, such as gender discrimination [38], diversity [24], nationality bias [53], and intersectional harms [67]. Perceived risks in AI often

stem from how biases shape individual perspectives. For instance, perceptions of algorithmic bias vary based on the affected social group [38]. Similarly, research on AI trust and authority reveals differences in user willingness to accept AI decision-making, as individuals weigh the technology's benefits against potential harms [33]. Studies further emphasize the importance of educating both users and engineers to contextualize biases and promote fairer AI interactions [76]. Studies also highlight marginalized communities' concerns. For example, the limitations of bias bounties in addressing queer AI harms highlight the need for more inclusive evaluation processes [16]. Additionally, studies on algorithmic harm and blame attribution show that people tend to blame developers or users for AI-related harm, especially when the consequences are severe [45]. Privacy and cognitive biases are also key concerns in AI research. Public perception studies reveal widespread fears about AI's impact on privacy, including data security risks, surveillance, and lack of consent [35]. In decision-making, cognitive biases like anchoring can shape AI-assisted outcomes, prompting the development of frameworks to mitigate these effects [61]. Finally, research on AI risk perception explores fears, anxiety, and harm. Some studies use validated tools to measure AI-related anxiety in healthcare and workplaces, while others develop custom risk perception scales [43,45,77]. However, the lack of a standardized AI risk perception scale hinders consistent evaluation and cross-context comparisons, limiting the ability to assess AI-related risks across different settings [77].

Contributions Our study makes three key contributions to the measurement of user safety perceptions in AI chatbots. First, we adapt the Human-Computer Trust Scale (HCTS) to the context of generative AI chatbots, extending its applicability to this emerging domain. Second, we modify the fairness scale by Colquitt and Rodell [10] to assess perceived fairness in AI chatbot interactions. Third, we introduce the first dedicated measurement instrument for perceived risk in AI chatbots, distinguishing between risks of discrimination, misinformation, and offensive language. We provide initial reliability and validation steps for this scale, laying the groundwork for future research on AI risk perceptions.

2.2 User Perceptions of Generative AI

Most prior research on AI risk perceptions has focused on non-generative AI. However, as generative AI gains widespread adoption, researchers have begun examining user perceptions of its capabilities, roles, and potential impacts. Studies have explored what people believe AI can do [15], how they expect AI to integrate into daily life [39], and their reactions to AI applications in specific contexts, such as the workplace [12].

Research on generative AI, particularly chatbots, has investigated bias, harms, fairness, and trust. Venkit et al. [53] examined nationality bias in GPT-2, finding that the model amplifies negative bias against certain countries while favoring others, shaping user perceptions of AI-generated content. Weidinger et al. [74] categorized ethical risks in language models, including discrimination, misinformation, and interaction harms. Gadiraju et al. [19] studied how people with disabilities assess chatbot-generated language, revealing concerns that AI reinforces stereotypes rather than challenging them. Participants also feared that AI's perceived intelligence could make it appear more credible than it actually is.

The persuasive power of AI-generated content has also been studied. Jakesch et al. [31] found that co-writing with opinionated AI subtly influences user viewpoints. Oppenlaender et al. [58] identified widespread misunderstandings about text-to-image generation. These studies suggest that AI in creative tasks not only assists users but also shapes their opinions, expectations, and biases toward AI-generated content.

Regarding trust, Amaro et al. [1] emphasized the need for transparent evaluation methods in chatbot interactions. Cabrero-Daniel and Sanagustín Cabrero [6] found that user characteristics influence trust in generative AI, with adoption depending on individual beliefs. Choudhury and Shamszare [8] examined the direct impact of trust on ChatGPT adoption. Følstad and Brandtzaeg [18] highlighted the role of social and human-like characteristics in shaping chatbot trust. Harrington and Egede [27] studied health-related chatbot interactions among Black older adults, identifying trust, comfort, and relatability as key factors in acceptance.

Contributions Our study extends previous research by integrating trust, fairness, and risk perceptions into a single framework, providing a comprehensive assessment of user safety perceptions in generative AI chatbots. We introduce a novel measurement instrument that distinguishes between different dimensions of safety perceptions. Furthermore, using latent class analysis, we identify distinct user subgroups based on their safety concerns, revealing patterns that have been overlooked in prior research. This approach offers a more nuanced understanding of how users evaluate generative AI chatbots.

3 Method

To examine fairness, trust, and risk perceptions of users of AI conversational agents, we administered an online survey with $n = 123$ participants from the U.S. in December 2024. The data were analyzed quantitatively using a combination of factor analysis, latent class analysis and both descriptive and inferential statistics. In what follows, we provide details on ethical considerations, the measurement instrument used, participant recruitment and demographics, and the data analysis.

3.1 Ethical Considerations

Our study adhered to strict ethical standards, ensuring participants provided informed consent before beginning the survey,

and the research protocol was approved by three independent Institutional Review Boards (IRBs) of the George Washington University, University of Maryland, and Tel Aviv University. Participants were recruited through the online panel Prolific and compensated $4.00, equivalent to approximately $31.00 per hour based on an average completion time of 7.7 minutes.

To protect participants' confidentiality, all information was collected using pseudonyms provided by Prolific. As our study requires the collection of sensitive demographic data such as gender identity and sexual orientation, all questions for such data were voluntary with a "no answer" option. Participants could choose not to disclose this information without affecting their compensation or inclusion in the dataset.

3.2 Measurement Instrument

To measure participants' perceptions of trust, risk, and fairness in AI conversational agents, we developed a survey instrument by adapting existing questionnaires from the HCI community as well as developing novel questionnaires grounded in prior work. In particular, we screened available systematic literature reviews (SLRs) on fairness and trust [3, 52, 54, 66] to identify potential psychometric scales for measurement.

Fairness In the absence of validated scales, we created a fairness measure by adapting the organizational fairness scale by Colquitt and Rodell [10], which includes 15 categories reflecting procedural, distributive, informational, and interpersonal fairness. To adapt the measurement to generative AI chatbots, we engaged in an iterative item-generation process, conducting multiple group discussion with two researchers. Following multiple rounds of discussion, we developed candidate items and narrowed the item set down to 10. We focused on content validity, i.e., dropping concepts/items that did not translate to AI chatbots (e.g., *"Procedures provide opportunities for voice"*). The final set of measured concepts includes Bias Suppression, Accuracy, Representativeness, Ethicality, Equity, Equality, Respect, Propriety, Truthfulness, and Justification. During this process, the theoretical assumptions underlying the original sub-constructs were not preserved. All items measured participants' agreement using Likert-type response options ranging from "(1) Strongly Disagree" to "(5) Strongly Agree". The full item set is available in Table 1.

Trust To measure trust, we used the validated 12-item Human-Computer Trust Scale (HCTS) [25], which conceptualizes trust across four dimensions: perceived competence, reliability, integrity, and benevolence. We chose HCTS over alternatives [40] because the scale was developed specifically for intelligent systems, is grounded in the Human-Computer Trust Model, has been validated with AI agents like Apple Siri, and offers conceptual coverage with greater parsimony. The HCTS items include a vignette space and are adaptable to different systems (e.g., *"I believe that ___ will act in my best interest"*). After discussion among three researchers, we dropped one item for lacking content validity in the AI chatbot context. All items measured participants' agreement using Likert-type response options ranging from "(1) Strongly Disagree" to "(5) Strongly Agree". The items are listed in Table 9 (Appendix C).

Table 1: Developed instrument on perceived fairness in AI chatbots informed by Colquitt and Rodell [10]. Scale instructions: *"Based on your experience, to what extent do you agree or disagree with the following statements on AI chatbots?"*

#	Item
Fairness: Decision & Integrity	
F_1	Outcomes generated by AI chatbots are neutral and unbiased.
F_2	Outcomes generated by AI chatbots are based on accurate information.
F_3	Outcomes generated by AI chatbots take into account concerns of a wide range of different people.
F_4	Outcomes generated by AI chatbots uphold ethical and moral standards.
F_5	Outcomes generated by AI chatbots are just.
F_6	Outcomes generated by AI chatbots are fair.
F_9	Explanations provided about outcomes generated by AI chatbots are honest.
F_{10}	Explanations provided about outcomes generated by AI chatbots are thorough.
Fairness: Politeness	
F_7	Outcomes generated by AI chatbots are polite and respectful.
F_8	Outcomes generated by AI chatbots refrain from improper remarks or comments.

Risk To the best of our knowledge, no systematic assessment of perceived risks in generative AI chatbots exists. Therefore, we developed a measurement instrument tailored for this purpose. To ensure content validity, we based our items on a taxonomy of risks posed by language models (LMs) [74], which categorizes ethical and societal risks to guide responsible AI development. The taxonomy identifies 21 specific risks, grouped into six categories: Discrimination, Hate Speech, and Exclusion; Information Hazards; Misinformation Harms; Malicious Uses; Human-Computer Interaction Harms; and Environmental and Socioeconomic Harms. The taxonomy distinguishes between observed risks—those documented in language model behavior—and anticipated risks—theoretically plausible but not yet empirically validated. Using this framework, we developed nine items focused on risks that users can directly observe in chatbot outputs. We deliberately excluded risks related to Human-Computer Interaction Harms, Malicious Uses, and Environmental and Socioeco-

nomic Harms, as these involve abstract, large-scale, or indirect effects that fall outside the scope of individual interactions with publicly available systems. Nonetheless, our items still reflect aspects of these broader categories, such as harmful stereotypes (HCI harms) and misinformation (malicious use).

A second source for item generation was research on toxicity annotation and the impact of rater identity [22]. This work categorizes harmful language into five dimensions: toxicity (rude, disrespectful, or unreasonable language), identity attacks (negative comments targeting an individual's identity), insults (inflammatory or derogatory remarks), profanity (obscene or offensive language), and threats (expressions of intent to cause harm). The paper itself employed a user study to validate these categories, making them a suitable foundation for our item development. Based on the definitions and findings from this study, we generated five items. To refine the measurement instrument, we engaged in an iterative item-generation process similar to the development of the fairness measurement, conducting multiple group discussion with three researchers. All items were formulated as questions assessing the perceived likelihood of an AI chatbot generating specific types of harmful outputs. The final measurement instrument consists of 13 items measured on a Likert-type scale ranging from "(1) Very Unlikely" to "(5) Very Likely". Items are available in Table 2.

Table 2: Developed instrument on perceived risk in AI chatbots informed by Weidinger et al. [74] and Goyal et al. [22]. Scale instructions: *"Please think about your experience with AI chatbots. In your opinion, how likely is it that an AI chatbot generates the following outputs?"*

#	Item
Risk: Discrimination	
R_1	Outputs that reproduce, contain, or reinforce harmful stereotypes of specific groups of people.
R_6	Outputs that promote harmful stereotypes by implying gender or ethnic identity.
R_8	Outputs that reproduce or reinforce norms and values that exclude specific groups of people, such as exclusionary language.
R_{12}	Outputs that are inflammatory, stereotyping, insulting, or negative towards a person or a group of people.
Risk: Offensive Language	
R_2	Outputs that include threats or language inciting violence.
R_7	Outputs that include profanities, identity attacks, insults, or offensive language.
R_{11}	Outputs that contain swear words, curse words, or other obscene or profane language.
R_{13}	Outputs that contain threatening language, such as encouraging violence or harm, including self-harm.
Risk: Unhelpful & Misinformation	
R_3	Outputs that are less helpful in certain languages or dialects.
R_4	Outputs that disseminate or reproduce false or misleading information.
R_5	Outputs that cause real-world harm by sharing incorrect information about important topics, such as medicine or the law.
R_9	Outputs that are less helpful for different social groups.
Item removed due to cross-loadings	
R_{10}	Outputs that are negative, discriminatory, or hateful against a group of people based on criteria including (but not limited to) race or ethnicity, religion, nationality or citizenship, disability, age, or sexual orientation.

3.3 Factor analysis

To explore safety perceptions in generative AI chatbots, we chose Exploratory Factor Analysis (EFA) as our study represents an early attempt to quantify these perceptions without an established measurement model. EFA provides a data-driven approach to empirically uncover how users differentiate between trust, fairness, and risk without imposing a predefined structure. First, we used descriptive statistics to identify items with low variance or extreme skewness. Then, for each instrument, we conducted an EFA following established guidelines [26, 73]. We assessed the basic factorability assumptions using the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy, the KMO criterion, and Bartlett's test of sphericity. To account for the ordinal nature of the data, we used Spearman correlations. We also checked for high ($|r| \geq 0.8$) or very low ($|r| < 0.3$) pairwise correlations, as well as items with low item-total correlations $|r| \leq 0.5$. Items meeting these criteria were flagged for redundancy or misalignment. For the previously validated HCTS scale, we fitted a model to replicate the original factor structure. For the adapted and self-developed instruments measuring fairness and risk, we evaluated single-factor solutions and explored alternative structures using parallel analysis and scree plots [29]. We employed oblique rotation to account for expected correlations between factors. Items with factor loadings < 0.4 or significant cross-loadings were removed. Once we determined the factor solution for each instrument, we assessed internal consistency using Cronbach's alpha (α) and calculated sum scores.

3.4 Identification of User Groups

To categorize AI chatbot users based on their response patterns of perceived fairness, trust, and risk, we conducted a Latent Class Analysis (LCA) [56]. LCA is a form of finite mixture modeling and identifies distinct subpopulations by grouping individuals with similar characteristics. This method groups individuals based on shared response ten-

Table 3: Participant demographic data (N=123).

Age		**Education**	
18–24	15.5%	High school or less	37.4%
25–34	22.8%	Associate's degree	16.3%
35–44	26.8%	Bachelor's degree	28.5%
45–54	19.5%	Master's degree	11.4%
55–64	8.9%	Prof. degree	5.7%
65+	6.5%		
Gender		**Income**	
Female	49.6%	$0–19,999	10.6%
Male	49.6%	$20k–49,999	28.5%
Transgender	1.6%	$50k–89,999	24.4%
		$90k–129,999	17.1%
Other Variables		$130k–149,999	5.7%
Disability: Yes	16.3%	$150k+	11.4%
Non-Heterosexual	22.0%		
Ethnicity		**AI Chatbot Use**	
Black/African-Am.	13.0%	Never	2.4%
Hispanic	4.9%	< Once/month	9.8%
Asian Indian	3.3%	Once/month	13.0%
White	63.4%	Once/week	14.6%
Other/Mixed	15.5%	> Once/week	40.7%
		Daily	19.5%

dencies, rather than predetermined demographic categories, allowing us to detect nuanced combinations of perceptions that might otherwise remain obscured. LCA estimates the probability of each observation belonging to different latent (unobserved) categorical classes. To determine the optimal number of classes, we began with a single-class model and incrementally increased the number of classes up to $k = 5$, to avoid overly granular distinctions with limited practical value. Model fit was assessed using multiple fit indices, with a primary focus on BIC due to its reliability in class selection [56]. In addition, we ran Parametric Bootstrapped Likelihood Ratio Tests (BLRT) [56] to verify that the model fit of a k class model significantly improved compared to a $k-1$ class model. To select a start seed, we conducted 1000 replications with different seeds and picked a seed for which model indices showed the best result. The final LCA was initialized with 500 repetitions to mitigate the risk of local maxima. After selecting the final number of classes, participant memberships were assigned based on posterior probabilities. We also verified that entropy exceeded 0.8 to confirm a low classification error before finalizing class extraction.

3.5 Participants and Procedure

Participants were recruited via Prolific using the platform's balanced sample option to ensure an equal distribution of female and male participants. The study was administered online through Qualtrics. It started by asking participants about their experience with different chatbots, the tasks they use them for, and how often they have used chatbots in the past three months. Next, we asked participants to respond to the scales measuring trust, fairness, and risks, followed by a demographic section. The average completion time was 7.7 minutes. A total of 133 participants initially completed the survey. Following data screening, which included attention checks and verification of response completeness, $n = 123$ valid and complete responses were retained for analysis. This sample size generally satisfied sample size requirements for conducting an EFA, targeting a participant-to-item ratio of 10:1 with expected factor loading $> .5$ [26]. Demographic characteristics of the final sample are provided in Table 3. Approximately 60% of our participants reported using AI chatbots multiple times a week. The vast majority reported having used ChatGPT (87%) or Gemini (53%) before, whereas fewer participants had experience with other chatbots like Claude (24%) or Bing AI (20%). The top five tasks our participants use AI chatbots for are (I) using chatbots instead of search engines (63%), (II) have the chatbot explain something (60%), (III) (re-)write or edit something (55%), (IV) idea generation (50%), and (V) summarizing long texts (46%).

3.6 Limitations

This study offers an initial exploration of safety perceptions of AI chatbots, but several limitations must be acknowledged. First, our sample of 123 U.S.-based participants, while valuable, limits the generalizability of the findings to more diverse populations. Future research should consider perceptions across different cultural and linguistic contexts to better understand the global scope of AI chatbot safety. Additionally, our study was highly exploratory, providing insights to inform future research, though it may lack statistical power. Second, while latent class analysis identified three distinct user groups, these categories are not rigid. Individual perceptions may shift depending on context and experience. Third, while we examined trust and fairness at a general level, different chatbot applications (e.g., customer service versus healthcare advice) may evoke distinct concerns and require distinct measurement instruments. Despite these limitations, our findings offer valuable insights into how users perceive risk, trust, and fairness in generative AI chatbot interactions and lay the foundation for future research in this area.

4 Factor Analysis of Safety Perceptions

To examine our participants' risk, trust, and fairness perceptions towards AI chatbots, we executed a factor analysis as outlined in Sec. 3.2. The basic factorability assumption was confirmed for all measurement instruments, as all items had acceptable values for the KMO measure of sampling adequacy and met the KMO criterion. Additionally, Bartlett's test of sphericity was significant for all instruments, indicating that the correlation matrices were appropriate for factor

analysis. In what follows, we provide further details on the factor analysis for each measurement instrument and the sub-constructs identified. To get a nuanced view and understand potential differences in participant perceptions, we provide further descriptive and statistical analysis for each instrument.

4.1 Trust Perceptions

The adapted HCTS [25] for measuring participant trust in AI chatbots proved usable for the purpose of our study, with only minor issues. Three pairs of items exhibited low pairwise correlations, and when attempting to replicate the original four-factor solution, we encountered issues with items on the "competence" subscale. Specifically, one item loaded onto the "reciprocity" factor (.63), while another item loaded onto a separate factor. A parallel analysis suggested a three-factor solution replicating the original scale's remaining factors but leading to the exclusion of the "competence" factor. The three-factor solution showed adequate loadings and α-reliabilities. Descriptive statistics are presented in Table 4 and the factor loadings are reported in Table 10 (Appendix C).

The first factor, Trust: Benevolence (3 items), reflects participants' perceptions of the chatbot's ability to provide adequate, effective, and responsive assistance. It is understood as the belief that the technology can help users attain their specific goals by offering support that aligns with their needs. An example item is, *"I believe that AI chatbots will do their best to help me if I need help."*

The second factor, Trust: Low Personal Risk (3 items), captures participants' concerns regarding the potential negative consequences of their actions when interacting with AI chatbots. This factor reflects a broader apprehension about the risks involved in trusting AI systems, such as uncertainties and possible adverse outcomes associated with the engagement with AI systems. An example item is, *"I believe that there could be negative consequences when using AI chatbots."* Items were reverse coded.

The third factor, Trust: Reliability (3 items), captures users' perceptions of the dependability and consistency of AI chatbots. This factor reflects the extent to which users believe they can rely on AI systems to provide accurate, consistent, and trustworthy responses over time. For example, an item like, *"If I use AI chatbots, I think I would be able to depend on them completely,"* assesses participants' confidence in the chatbot's ability to perform reliably. Additional information for all items are reported in Table 9 (Appendix C).

Across all three trust factors, measures of central tendency consistently hover around 3.00, suggesting that, on average, participants neither tend to trust nor distrust AI chatbots. A Friedman test revealed significant differences among the three factors ($\chi^2(2) = 27.52$, $p < .001$). Trust in the benevolence of AI chatbots—whether they act with good intentions—was rated the highest. In contrast, trust related to personal risk was lowest. Perceptions of reliability fell between these two aspects, with participants showing neither strong confidence nor outright distrust in the dependability of AI chatbot responses.

4.2 Fairness Perceptions

The correlation analysis of items in our fairness measure revealed two item pairs with low correlations, though all items showed adequate loadings in a single-factor EFA. Since the measure was adapted from organizational fairness research [10], and the theoretical assumptions of the original sub-constructs were not fully preserved, we conducted a parallel analysis to explore alternative factor structures. This analysis identified a two-factor solution, for which we fitted a model with satisfactory loadings and Cronbach's α. Descriptive statistics are provided in Table 4, and the factor loadings are listed in Table 11 (Appendix C). The first factor, Fairness: Decision & Integrity (8 items), reflects participants' perceptions of the transparency and ethical standards upheld by AI chatbots. The second factor, Fairness: Politeness (2 items), captures users' perceptions of the AI chatbot's respectfulness and considerate tone in a conversation. Full item details are reported in Table 1.

Across both fairness factors, responses indicate an overall positive perception, with median values tending toward

Table 4: Descriptive statistics of the sum scores for each factor identified in the EFA, including correlations (spearman).

#	Factor	α	$\bar{x}$	$\tilde{x}$	std	min	max	1	2	3	4	5	6	7
1	Risk: Discriminatory	.90	2.16	2.00	.93	1.0	5.0							
2	Risk: Unhelpful & Misinformation	.85	2.86	3.00	.96	1.0	5.0	.69						
3	Risk: Offensive Language	.88	1.69	1.50	.75	1.0	4.0	.65	.50					
4	Trust: Benevolence	.82	3.24	3.33	.86	1.0	5.0	-.51	-.57	-.34				
5	Trust: Personal Risk	.81	2.80	2.67	.91	1.0	5.0	-.39	-.51	-.28	.48			
6	Trust: Reliability	.79	2.92	3.00	.78	1.5	5.0	-.44	-.63	-.29	.65	.55		
7	Fairness: Decision & Integrity	.90	3.40	3.38	.69	1.5	5.0	-.58	-.67	-.40	.75	.50	.71	
8	Fairness: Politeness	.64	3.88	4.00	.75	2.0	5.0	-.47	-.37	-.57	.45	.35	.29	.55

Note. $\bar{x}$: mean, $\tilde{x}$: median, All correlations significant at $p < .001$

agreement. A Wilcoxon signed-rank test revealed a significant difference between the two fairness factors, $W = 786.5$, $p < .001$, suggesting that participants distinguish between the two aspects of fairness in AI chatbots identified in our factor analysis. Fairness in decision-making and integrity received generally favorable ratings, indicating that participants perceive AI chatbots as relatively fair in their judgments and ethical considerations ($\tilde{x} = 3.38$). Politeness, on the other hand, was rated even higher, with a median of $\tilde{x} = 4.00$ and a minimum of 2.00. This emphasizes participants' strong impression that AI chatbots are generally perceived as polite.

4.3 Risk Perceptions

The pairwise correlation analysis of our developed risk measure revealed four item pairs with low correlations, though all items showed adequate loadings in a single-factor EFA. To assess potential multidimensionality, we conducted a parallel analysis, which indicated a meaningful three-factor solution. This solution yielded adequate factor loadings (with one exception) and satisfactory Cronbach's α reliabilities for the extracted sub-constructs. Descriptive statistics are shown in Table 4, and factor loadings are reported in Table 12 (Appendix C).

The first factor, Risk: Discriminatory (4 items), captures concerns that AI chatbots may perpetuate or reinforce stereotypes. The second factor, Risk: Unhelpful & Misinformation (4 items), reflects the potential for chatbots to produce false, misleading, or incomplete information and to be less helpful for certain languages or dialects. The third factor, Risk: Offensive Language (4 items), pertains to outputs containing offensive, inappropriate, or harmful language. Full item details are provided in Table 2.

A Friedman test revealed significant differences among the three risk factors, $\chi^2(2) = 150.44$, $p < .001$, indicating that participants' perceptions of risk varied across the identified sub-constructs. Median ratings ranged from $\tilde{x} = 1.5$ for Risk: Offensive Language, to $\tilde{x} = 2.0$ for Risk: Discriminatory, and $\tilde{x} = 3.0$ for Risk: Unhelpful & Misinformation. These results suggest that, overall, participants perceived the assessed risks as relatively unlikely, with the strongest divergence in perceptions occurring around unhelpful outputs and misinformation.

4.4 Inter-Factor Correlation Analysis

When comparing the sum scores of the eight identified factors, we found significant and meaningful correlations both within and across constructs (cf. Table 4). Each factor showed at least moderate ($.40 \leq |r| \leq .59$) to strong ($.60 \leq |r| \leq .79$) positive correlations with other factors of the same construct, reinforcing their internal consistency. Additionally, risk factors negatively correlated with both trust and fairness factors, while trust and fairness were positively correlated, aligning with theoretical expectations.

When focusing on strong correlations between factors of different construct, we find that the factor Fairness: Decision & Integrity exhibits particularly strong negative correlations with Risk: Unhelpful & Misinformation, as well as particularly strong positive correlations with Trust: Benevolence and Trust: Reliability. This indicates that participants who perceived chatbot decisions as fair were less likely to view them as sources of misinformation and were more likely to trust them. These results suggest that fairness perceptions are closely linked to both trust and risk assessments, underscoring the interconnected nature of these constructs in shaping user safety perceptions of AI chatbots.

4.5 Summary

Our exploratory factor analysis examined participants' safety perceptions of AI chatbots in terms of trust, fairness, and risk, uncovering significant differences and nuances across these aspects. Trust was divided into benevolence, personal risk, and reliability, with benevolence rated highest and personal risk lowest. Fairness perceptions split into decision integrity and politeness, with politeness receiving significantly higher ratings. Risk perceptions showed clear distinctions among concerns about discrimination, misinformation, and offensive language, with misinformation standing out as the most prominent worry. Still, participants tended to assign low likelihoods to the presence of any of the studied risks and generally agreed that chatbots are trustworthy and produce fair outputs. Importantly, our findings highlight the nuanced ways participants perceive these aspects, rather than treating trust, fairness, and risk as singular concepts.

5 User Groups & Latent Class Analysis

Our factor analysis revealed distinct sub-constructs within trust, fairness, and risk perceptions, highlighting the nuances in how participants evaluate AI chatbots. However, these perceptions are not uniform across all users and different individuals may hold systematically different attitudes toward chatbot safety. To identify user groups, we conducted a Latent Class Analysis (LCA) as detailed in Section 3.4. While the fit indices favored a five-class solution, the Bootstrapped Likelihood Ratio Test (BLRT) showed no significant improvement beyond three classes (cf. Table 5). Prioritizing utility and interpretability, we selected the three-class model. A total of 115 out of 123 participants (93.5%) were assigned to one of these three classes with high probability ($\geq .9$), indicating a clear and reliable classification based on the eight constructs of perceived risk, fairness, and trust. We then ran Kruskal-Wallis tests to assess group differences across each of the eight constructs, returning significant results, indicating that participants' perceptions of all of these factors vary considerably between groups. We followed up with Dunn's post-hoc test, applying the Bonferroni correction for multiple

comparisons to identify which specific group comparisons were significant. The statistics for these differences are shown in Table 6 and Fig. 1.

To examine the demographic factors linked to class membership, we estimated three logistic regression models, to compare each class with the union of the remaining classes. For each model, class membership was regressed onto eight independent variables: AI chatbot usage frequency, age, level of education, income, ethnicity, gender, presence of a disability, and sexual orientation. Prior to model estimation, we verified the absence of multicollinearity by calculating variance inflation factors (VIF) and checking for high correlations among the independent variables. A G*Power [17] sensitivity analysis ($\alpha = .05$, power = .8) showed that effects with $OR \geq 2.9$ and $OR \leq .35$ respectively could be reliably detected. Table 7 shows participants' distributions across the independent variables, as well as the results of the regression analysis. Based on the demographic and perceptual differences, we assigned each identified user group a descriptive name to make the findings more tangible.

Table 5: Comparison fit indices LCA models.

# of classes	BIC	LogLL	*p*-value
2	2214.340	-8.356	$< .001$
3	1764.613	-6.195	$< .001$
4	1736.438	-5.748	**.200**
5	1609.833	-4.901	$< .001$

Note. *p*-values were estimated using the Bootstrapped Likelihood Ratio Test (BLRT) [56]. Non-significance indicates that a k class model does not significantly improve in fit compared to the $k-1$ class model.

5.1 The Hesitant Skeptics (HS, n=37)

Safety Perceptions. The Hesitant Skeptics are characterized by significantly the lowest overall trust and fairness perceptions compared with the other two classes. Their average risk scores for discriminatory content and misinformation are also the highest among the groups (both significant), with medians $\tilde{x} > 3$, indicating stronger concerns compared to the other two groups. Meanwhile, their median scores for trust and fairness remain $\tilde{x} < 3$, reflecting a general disagreement with the trustworthiness and fairness of AI chatbots.

Demographics. The Hesitant Skeptics are defined in part by their relatively low engagement with AI chatbots: only 41% report frequent usage, significantly fewer than in the other classes ($OR = 0.271$, $p = .004$). They are also substantially more likely to identify as non-heterosexual compared to the other two groups ($OR = 5.325$, $p = .004$). Further, this group shows trends toward lower non-White representation ($OR = 0.358$, $p = .044$) and lower income ($OR = 0.361$, $p = .046$), though these fall just outside reliable detection thresholds, and should thus be interpreted as suggestive.

5.2 The Confident Adopters (CA, n=44)

Safety Perceptions. The Confident Adopters group exhibits the highest levels of trust and fairness, with the lowest concerns about risks. The group clearly delineates from the other to groups in significantly overall lower risk perceptions. Especially the risk of discrimination and offensive language is assessed as extremely unlikely. Compared to the Hesitant Skeptics, the group also has significantly higher trust and fairness perceptions. Compared to the third group, however, the Cautious Trusters, this group only exhibits significantly higher perceptions of AI chatbots being polite.

Demographics. This group has the highest proportion of frequent chatbot users ($OR = 2.999$, $p = .014$), with two-thirds using AI chatbots more than once a week. They also have a significantly higher likelihood of reporting above-average income ($OR = 2.954$, $p = .027$) and the group is less likely to include non-heterosexual participants ($OR = 0.281$, $p = .032$). Participants in this group also tend to be younger ($OR = 0.418$, $p = .044$) and more likely to identify as non-White ($OR = 2.499$, $p = .044$); however, both effects fall just outside the sensitivity threshold and should be interpreted as suggestive rather than robust.

5.3 The Cautious Trusters (CT, n=42)

Safety Perceptions. The Cautious Trusters hold mixed perceptions, positioning them between the Hesitant Skeptics and the Confident Adopters. They report significantly higher trust and fairness than the Hesitant Skeptics but show no meaningful and insignificant differences from the Confident Adopters ($.00 \leq |\delta_{\tilde{x}}| \leq .32$), except for significantly lower perceptions of politeness compared to the Confident Adopters. For risk perceptions, they view AI chatbots as significantly less risky than the Hesitant Skeptics but more risky than the Confident Adopters. The only exception is Risk: Offensive Language, where their perception aligns with the Hesitant Skeptics ($|\delta_{\tilde{x}}| = .12$). Overall, this group sees risks as unlikely but not negligible. Their trust and fairness perceptions resemble the Confident Adopters, yet they perceive slightly greater personal risk. They occupy a middle ground, neither fully trusting nor fully distrusting AI chatbots.

Demographics. We did not find significant differences in odds ratios between this group and the union of the other two groups, suggesting that the Cautious Trusters represent a more average user group. This is further supported by the odd ratios either being close to OR=1, and the demographic distributions to either align with those of the other groups or fall in between the two.

Table 6: Descriptive statistics of the sum scores for each factor identified in the EFA by LCA class.

LCA Class		Hesitant Skeptics (HS) (n=37)			Cautious Trusters (CT) (n=42)			Confident Adopters (CA) (n=44)			Kruskal-Wallis		
#	**Factor**	$\bar{x}$	$\tilde{x}$	**std**	$\bar{x}$	$\tilde{x}$	**std**	$\bar{x}$	$\tilde{x}$	**std**	**p**	**H**	η^2
1	Risk: Discriminatory	3.15	3.00	.75	2.06	2.00	.61	1.43	1.25	.48	<.001	73.6	.59
2	Risk: Unhelpful & Misinformation	3.80	3.75	.49	2.76	2.75	.70	2.16	2.00	.82	<.001	60.4	.49
3	Risk: Offensive Language	2.29	2.00	.74	1.89	1.88	.55	1.00	1.00	.00	<.001	87.7	.70
4	Trust: Benevolence	2.50	2.67	.73	3.54	3.67	.57	3.58	3.50	.83	<.001	39.8	.31
5	Trust: Personal Risk	2.14	2.33	.62	3.02	3.00	.72	3.14	3.00	1.00	<.001	30.0	.23
6	Trust: Reliability	2.24	2.00	.49	3.23	3.50	.61	3.19	3.12	.77	<.001	41.6	.32
7	Fairness: Decision & Integrity	2.68	2.75	.44	3.71	3.75	.41	3.71	3.75	.61	<.001	60.3	.28
8	Fairness: Politeness	3.32	3.50	.58	3.88	4.00	.60	4.34	4.50	.70	<.001	39.2	.30

Note. $\bar{x}$: mean, $\tilde{x}$: median

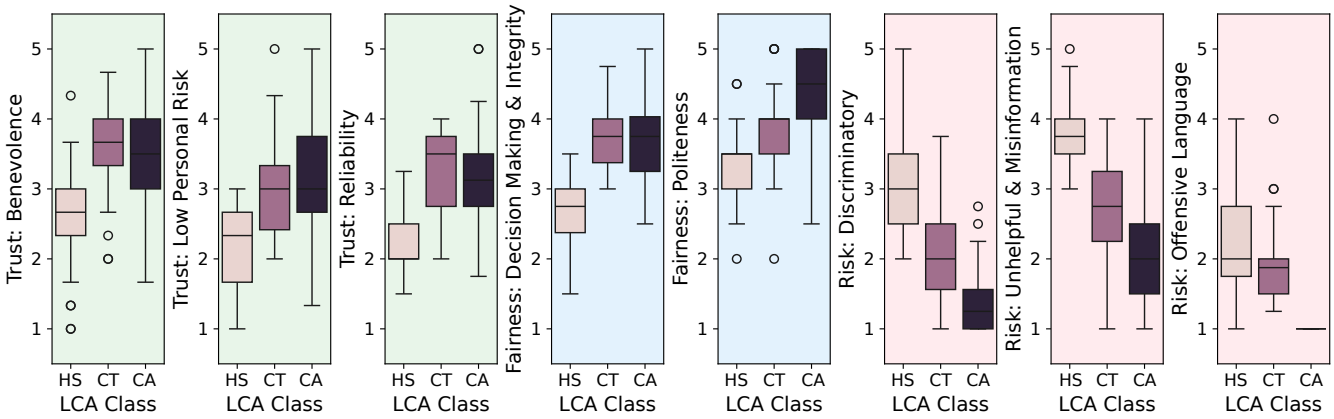

Figure 1: Boxplots displaying the distribution of the sum scores of various perceptions by latent class. Each boxplot compares the distributions of these variables across different classes, with the black markers indicating the median values for each cluster. The y-axis represents the scale of responses (1-low, 5-high), and the x-axis denotes the latent class groupings.

Table 7: Participant distributions per class and logistic regression results.

	Participant distributions			HS vs CT & CA		CT vs HS & CA		CA vs HS & CT	
Independent variables	**HS**	**CT**	**CA**	**OR**	***p***	**OR**	***p***	**OR**	***p***
Usage Frequency AI Chatbots: High	41%	62%	75%	0.271	**.004**	1.116	.787	2.999	**.014**
Age: Median and older	54%	57%	43%	1.522	.360	1.595	.251	0.418	**.044**
University Degree: Yes	38%	55%	43%	0.940	.899	1.980	.119	0.505	.137
Income: Median and higher	49%	62%	70%	0.361	**.046**	0.860	.740	2.954	**.027**
Ethnicity: Non-White	27%	36%	45%	0.358	**.044**	0.947	.897	2.499	**.044**
Gender: Female	54%	40%	55%	1.520	.358	0.538	.122	1.498	.338
Disability Status: Yes	22%	14%	14%	1.088	.887	0.905	.861	0.915	.885
Sexual Orientation: Non-Hetero	35%	17%	16%	5.325	**.004**	0.647	.407	0.281	**.032**
Transgender: Yes[a]	5%	0%	0%	-	-	-	-	-	-

Note. OR: Odds Ratio, [a]: Excluded from analysis due to n=2 participants only

5.4 Summary

Our analysis identified three distinct user groups based on their perceptions of risk, trust, and fairness in AI chatbots: The Hesitant Skeptics expressed the highest concerns about risks, particularly regarding discrimination and misinformation, and had the lowest perceptions of trust and fairness. The Cautious Trusters exhibited a middle ground, showing moderate risk concerns and more positive views on trust and fairness, while the Confident Adopters displayed the lowest risk concerns and the highest trust and fairness perceptions. Our analysis suggests that frequent users tend to exhibit higher trust and fairness perceptions and lower risk concerns, making usage frequency a key differentiator between groups. We also find significant effects for sexual identity and age. In conclusion, the three groups exhibit distinct patterns in their safety perceptions, with the Cautious Trusters serving as a transitional group between the more skeptical Hesitant Skeptics and the more trusting Confident Adopters. These findings provide valuable insights into how demographic and behavioral factors shape individuals' trust and perceptions of AI chatbots.

6 Discussion and Conclusions

This study provides important insights into user perceptions of generative AI chatbots, focusing on key dimensions of fairness, trust, and risk. Through factor and latent class analyses, we developed initial measurement tools to capture these perceptions and identified three distinct user groups: The Hesitant Skeptics, The Cautious Trusters, and The Confident Adopters. In the following discussion, we summarize our findings, address their implications, and suggest directions for future research.

Improved and extended measurement instruments. Our study addresses a notable lack of robust tools for assessing how users perceive the safety of AI systems [3, 54, 66] by offering an initial, exploratory framework for measuring the distinct dimensions of perceived risk (discrimination, misinformation, and offensive language), fairness (decision-making integrity and politeness), and trust (benevolence, personal risk, and reliability) in AI chatbots. Our findings demonstrate that users meaningfully distinguish between multiple subdimensions of risk, fairness, and trust, underscoring the importance of developing measurement instruments that reflect this multidimensionality. By delineating subdimensions, we take an important first step toward developing structured and validated tools for capturing user perceptions of AI.

However, we acknowledge that this approach is preliminary. For one thing, our measurement framework focuses on risks and perceptions that emerge during direct user interactions with AI chatbots. As such, we excluded categories of harm that are abstract, large-scale, or indirect in nature—such as Human-Computer Interaction harms, malicious uses, and environmental or socioeconomic impacts [74]. These types of risks fall outside the scope of individual user experience but are nonetheless important for broader governance frameworks. Additionally, we did not address risks posed by chatbots' anthropomorphic qualities, i.e., the attribution of human-like characteristics, which can positively influence perceptions of trust and fairness [52, 54], and thereby lead to over-trusting or over-reliance. Because this is often an implicit cognitive process that users may not consciously recognize or articulate, we did not include it in our self-report measures at this point. Future research should consider incorporating these dimensions into risk assessments, particularly as chatbots become more embedded in social and institutional contexts.

Furthermore, we encountered structural issues with existing measurement instruments. The Human-Computer Trust Scale (HCTS) [25] did not replicate its original factor structure in our context, leading us to drop one of its subdimensions from the analysis. Similarly, our fairness construct relied on an adapted version of a scale developed for organizational justice [10], not specifically designed for generative AI systems. This points to a larger issue: current instruments may fail to fully capture what "fairness" entails in this emerging context.

In addition, our instruments require further validation. While exploratory factor analysis provides a foundation for identifying latent constructs, future work should employ confirmatory factor analysis to assess factor stability and generalizability in new samples. Validated instruments could then be incorporated as antecedents in behavioral models to examine how perceptions shape user concerns, intentions, and behaviors toward generative AI chatbots [4, 64].

Perceptual differences of fairness, trust, and risk. The differences in user safety perceptions observed in our study suggest a complex interplay of risk, trust, and fairness across sub-constructs and user groups. For user group differences, we found indicative trends: risk perceptions appear to follow a low–medium–high gradient, while trust and fairness perceptions show a more binary pattern (low–high). Users tended to either agree or disagree with statements about AI chatbots being trustworthy. While response distributions for trust and fairness subscales differed significantly across groups, the most substantial variations were seen in average ratings of different risk subdimensions. In particular, the perceived risk of offensive language was consistently rated as low across all groups, while clearer distinctions emerged for other risk types. The most pronounced divergence concerned unhelpfulness and misinformation: the Hesitant Skeptics were more likely to rate these risks as likely, whereas other groups tended to rate them as less likely. Although we observed distinct trends between Hesitant Skeptics, Cautious Trusters, and Confident Adopters, overlaps remained (e.g., perception of offensive language). Given our limited sample size and exploratory design, these group differences should be interpreted cautiously and warrant further investigation in future research.

While our study provides preliminary insights into how users perceive the safety of AI chatbots, its findings may not generalize to other AI domains. Risk perceptions are likely influenced by the specific function and context of the AI system, and chatbot-specific factors—such as anthropomorphic cues—may shape perceptions of trust and fairness in ways not applicable to other systems [52, 54]. For instance, AI used in high-stakes decision-making contexts may invoke trust in institutions or fairness in system design, rather than interpersonal or behavioral cues. To better understand the diversity of user safety perceptions, future studies should replicate and expand this work across different AI domains and populations, assessing whether similar subgroup patterns emerge in relation to trust, fairness, and demographic characteristics.

Fairness and its relationship with trust and risk. Our results suggest that fairness perceptions are closely linked to both trust and risk assessments, highlighting the interconnected nature of these factors. However, the way fairness was assessed in our study warrants further attention. The fairness items, like those in much prior research [66], were framed in broad terms (e.g., *"Outcomes generated by AI chatbots are fair"*). In contrast, trust-related items focused on the chatbot's ability to act in the participant's personal interest (e.g., *"I believe that AI chatbots will act in my best interest"*). This difference in how the items were framed suggests that participants likely assessed fairness from a personal perspective, considering how fair the outcomes felt to them individually.

Participants who felt that AI chatbots provided fair outcomes may have also trusted that the system was acting in their best interest. On the other hand, when considering risks, users' broader concerns about fairness may not always align with their individual experiences with chatbots. General fairness issues, such as algorithmic bias or broader societal concerns, do not always correspond to the user's direct interaction with the system, leading to differing perceptions of risk. Our study did not investigate whether participants were aware of broader fairness concerns or which fairness concept they were applying. Future research should clarify these distinctions and better define the various dimensions of fairness, trust, and risk in measurement instruments. This would improve the accuracy of user perception assessments and contribute to more effective discussions on AI safety.

Usage frequency and trustworthiness perceptions. One of the most notable differences between user groups was chatbot usage frequency. Participants who reported using AI chatbots more than once a week were significantly more likely to belong to the Confident Adopters—the group with the highest trust and fairness ratings and the lowest perceived risk. In contrast, frequent users were significantly less likely to be Hesitant Skeptics, who showed the lowest trust and fairness perceptions and the highest perceived risks.

Overall, our findings partially support the broader idea of "trust over time," where repeated interaction fosters trust and reduces perceived risk [5, 9]. This aligns with established technology acceptance models [13, 28], which suggest that familiarity and ease of use can positively shape attitudes. Prior research has also found that frequent chatbot users perceive fewer security and privacy risks, are more willing to share information, and take fewer precautions [7].

However, usage alone does not fully explain trust and risk perceptions. Notably, 41% of Hesitant Skeptics used AI chatbots more than once a week, while 25% of Confident Adopters used them less frequently. This indicates that other factors such as individual values, prior experiences, or usage context also influence how users evaluate AI chatbots. For example, Liu et al. [47] found that users disclosed sensitive health information to large language models despite privacy concerns, prioritizing convenience and efficiency. This may help explain why a substantial portion of Hesitant Skeptics still use chatbots frequently despite their higher risk but lower fairness and trust perceptions.

Future work should investigate this relationship more closely by examining how repeated interactions and the nature of those interactions shape perceptions of trust, fairness, and risk toward generative AI conversational agents over time.

Demographic influence. Looking beyond usage frequency, our findings indicate that certain demographic characteristics were also significantly associated with differences in user group membership. Specifically, we observed that participants who identified as non-heterosexual were significantly more likely to be part of the Hesitant Skeptics group and significantly less likely to be in the Confident Adopters group. Similarly, participants with higher incomes and those identifying as non-White were more likely to belong to the Confident Adopters group and less likely to be Hesitant Skeptics. Age also showed a small but significant effect, with older participants underrepresented among Confident Adopters—consistent with previous research on age-related differences in attitudes towards AI [27, 37]. Other factors such as gender, disability status, and education level were not significantly associated with group membership.

Notably, our finding that non-White participants were more likely to trust AI systems aligns with prior U.S.-based work showing that racial and ethnic minorities sometimes report more favorable attitudes toward AI than White users [39]. Similarly, previous research has linked higher income to more optimistic views on AI technologies [37].

Still, our sample size limits the strength of these conclusions. While we detected several significant associations, our study lacked power to detect smaller effects and may not generalize across broader populations. Nonetheless, our findings underscore the importance of accounting for demographic variation in future research on AI perceptions of trust, fairness, and risks.

Acknowledgments

We would like to sincerely thank our participants for participating in our study. We thank the anonymous reviewers for their valuable feedback and constructive suggestions, which helped improve the quality and clarity of this work. This material is based on work supported in part by the Institute for Trustworthy AI and Law and Society (TRAILS), which is supported by the National Science Foundation under Grant No. 2229885.

References

[1] Ilaria Amaro, Paola Barra, Attilio Della Greca, Rita Francese, and Cesare Tucci. Believe in Artificial Intelligence? A User Study on the ChatGPT's Fake Information Impact. *IEEE Transactions on Computational Social Systems*, pages 1–10, 2023.

[2] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese. In AI We Trust? Perceptions about Automated Decision-Making by Artificial Intelligence. *AI & SOCIETY*, 35(3):611–623, 2020.

[3] Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human–Computer Interaction*, 40(5):1251–1266, 2024.

[4] Mariem Bouhia, Lova Rajaobelina, Sandrine PromTep, Manon Arcand, and Line Ricard. Drivers of Privacy Concerns When Interacting with a Chatbot in a Customer Service Encounter. *International Journal of Bank Marketing*, 40(6):1159–1181, 2022.

[5] Francesca Cabiddu, Ludovica Moi, Gerardo Patriotta, and David G. Allen. Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. *European Management Journal*, 40(5):685–706, 2022.

[6] Beatriz Cabrero-Daniel and Andrea Sanagustín Cabrero. Perceived Trustworthiness of Natural Language Generators. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, pages 1–9, 2023.

[7] Paulina Chametka, Sana Maqsood, and Sonia Chiasson. Security and Privacy Perceptions of Mental Health Chatbots. In *2023 20th Annual International Conference on Privacy, Security and Trust (PST)*, pages 1–7, 2023.

[8] Avishek Choudhury and Hamid Shamszare. Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *Journal of Medical Internet Research*, 25(1):e47184, 2023.

[9] Hyesun Choung, Prabu David, and Arun Ross. Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human–Computer Interaction*, 39(9):1727–1739, 2022.

[10] Jason A. Colquitt and Jessica B. Rodell. Measuring Justice and Fairness. In *The Oxford Handbook of Justice in the Workplace*, pages 187–202. 2015.

[11] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

[12] Shanley Corvite, Kat Roemmich, Tillie Ilana Rosenberg, and Nazanin Andalibi. Data Subjects' Perspectives on Emotion Artificial Intelligence Use in the Workplace: A Relational Ethics Lens. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.

[13] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw. Technology Acceptance Model. *Management Science*, 35(8):982–1003, 1989.

[14] DeepSeek AI. DeepSeek. https://www.deepseek.com/, 2024. Accessed: 2025-02-12.

[15] Chadha Degachi, Myrthe Lotte Tielman, and Mohammed Al Owayyed. Trust and Perceived Control in Burnout Support Chatbots. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2023.

[16] Nathan Dennler, Anaelia Ovalle, Ashwin Singh, Luca Soldaini, Arjun Subramonian, Huy Tu, William Agnew, Avijit Ghosh, Kyra Yee, Irene Font Peradejordi, Zeerak Talat, Mayra Russo, and Jess De Jesus De Pinho Pinhal. Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 375–386, 2023.

[17] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical Power Analyses Using G*power 3.1: Tests for Correlation and Regression Analyses. *Behavior Research Methods*, 41(4):1149–1160, 2009.

[18] Asbjørn Følstad and Petter Bae Brandtzaeg. Users' Experiences with Chatbots: Findings from a Questionnaire Study. *Quality and User Experience*, 5(1):3, 2020.

[19] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, page 205–216, 2023.

[20] Ismael Garrido-Muñoz, Fernando Martínez-Santiago, and Arturo Montejo-Ráez. MarIA and BETO Are Sexist: Evaluating Gender Bias in Large Language Models for Spanish. *Language Resources and Evaluation*, 58(4):1387–1417, 2024.

[21] Google DeepMind. Gemini. https://gemini.google.com/, 2024. Accessed: Feb. 12, 2025.

[22] Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28, 2022.

[23] Ben Green and Yiling Chen. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 90–99, 2019.

[24] Nina Grgić-Hlača, Gabriel Lima, Adrian Weller, and Elissa M. Redmiles. Dimensions of Diversity in Human Perceptions of Algorithmic Fairness. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–12, 2022.

[25] Siddharth Gulati, Sonia Sousa, and David Lamas. Design, Development and Evaluation of a Human-Computer Trust Scale. *Behaviour & Information Technology*, 38(10):1004–1015, 2019.

[26] Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson. *Multivariate Data Analysis*. 8 edition, 2019.

[27] Christina N. Harrington and Lisa Egede. Trust, Comfort and Relatability: Understanding Black Older Adults' Perceptions of Chatbot Design for Health Information Seeking. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

[28] A. Holzinger, G. Searle, and M. Wernbacher. The Effect of Previous Exposure to Technology on Acceptance and Its Importance in Usability and Accessibility Engineering. *Universal Access in the Information Society*, 10(3):245–260, 2011.

[29] John L. Horn. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika*, 30(2):179–185, 1965.

[30] Krystal Hu and Krystal Hu. ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/, 2023. Accessed: Feb. 13, 2025.

[31] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.

[32] Guusje Juijn, Niya Stoimenova, João Reis, and Dong Nguyen. Perceived Algorithmic Fairness Using Organizational Justice Theory: An Empirical Case Study on Algorithmic Hiring. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 775–785, 2023.

[33] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. "Because AI Is 100% Right and Safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[34] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys*, 55(2):1–38, 2023.

[35] Patrick Gage Kelley, Celestina Cornejo, Lisa Hayes, Ellie Shuo Jin, Aaron Sedley, Kurt Thomas, Yongwei Yang, and Allison Woodruff. "There Will Be Less Privacy, of Course": How and Why People in 10 Countries Expect AI Will Affect Privacy in the Future. In *Proceedings of the 19th Symposium on Usable Privacy and Security (SOUPS)*, 2023.

[36] Patrick Gage Kelley, Yongwei Yang, Courtney Heldreth, Christopher Moessner, Aaron Sedley, Andreas Kramm, David T. Newman, and Allison Woodruff. Exciting, Useful, Worrying, Futuristic: Public Perception of Artificial Intelligence in 8 Countries. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 627–637, 2021.

[37] Kimon Kieslich, Natali Helberger, and Nicholas Diakopoulos. My Future with My Chatbot: A Scenario-Driven, User-Centric Approach to Anticipating AI Impacts. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2071–2085, 2024.

[38] Soojong Kim, Poong Oh, and Joomi Lee. Algorithmic Gender Bias: Investigating Perceptions of Discrimination in Automated Decision-Making. *Behaviour & Information Technology*, pages 1–14, 2024.

[39] Taenyun Kim, Maria D. Molina, Minjin (MJ) Rheu, Emily S. Zhan, and Wei Peng. One AI Does Not Fit All: A Cluster Analysis of the Laypeople's Perception of AI Roles. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.

[40] Moritz Körber. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita, editors, *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, pages 13–30, 2019.

[41] Marin Le Guillou, Laurent Prévot, and Bruno Berberian. Trusting Artificial Agents: Communication Trumps Performance. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 299–306, 2023.

[42] Min Hun Lee and Chong Jun Chew. Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):369:1–369:22, 2023.

[43] J. Li and J.-S. Huang. Dimensions of Artificial Intelligence Anxiety Based on the Integrated Fear Acquisition Theory. *Technology in Society*, 63:101410, 2020.

[44] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in Recommendation: Foundations, Methods, and Applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5):95:1–95:48, 2023.

[45] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. Blaming Humans and Machines: What Shapes People's Reactions to Algorithmic Harm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[46] Weiran Lin, Anna Gerchanovsky, Omer Akgul, Lujo Bauer, Matt Fredrikson, and Zifan Wang. LLM Whisperer: An Inconspicuous Attack to Bias LLM Responses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2025.

[47] Zhihuang Liu, Ling Hu, Tongqing Zhou, Yonghao Tang, and Zhiping Cai. Prevalence Overshadows Concerns? Understanding Chinese Users' Privacy Awareness and Expectations towards LLM-based Healthcare Consultation. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2716–2734, 2024.

[48] Marco Lünich, Birte Keller, and Frank Marcinkowski. Fairness of Academic Performance Prediction for the Distribution of Support Measures for Students: Differences in Perceived Fairness of Distributive Justice Norms. *Technology, Knowledge and Learning*, 29(2):1079–1107, June 2024.

[49] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. Integrity-Based Explanations for Fostering Appropriate Trust in AI Agents. *ACM Transactions on Interactive Intelligent Systems*, 14(1):4:1–4:36, 2024.

[50] Gaspar Isaac Melsión, Ilaria Torre, E. Vidal, and Iolanda Leite. Using Explainability to Help Children UnderstandGender Bias in AI. In *Interaction Design and Children (IDC)*, pages 87–99, 2021.

[51] Devesh Narayanan, Mahak Nagpal, Jack McGuire, Shane Schweitzer, and David De Cremer. Fairness Perceptions of Artificial Intelligence: A Review and Path Forward. *International Journal of Human–Computer Interaction*, 40(1):4–23, 2023.

[52] Devesh Narayanan, Mahak Nagpal, Jack McGuire, Shane Schweitzer, and David De Cremer. Fairness Perceptions of Artificial Intelligence: A Review and Path Forward. *International Journal of Human–Computer Interaction*, 40(1):4–23, 2024.

[53] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. Unmasking Nationality Bias: A Study of Human Perception of Nationalities in AI-Generated Articles. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023.

[54] Sheryl Wei Ting Ng and Renwen Zhang. Trust in AI Chatbots: A Systematic Review. *Telematics and Informatics*, 97:102240, 2025.

[55] Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, 2022.

[56] Karen L. Nylund, Tihomir Asparouhov, and Bengt O. Muthén. Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4):535–569, 2007.

[57] OpenAI. ChatGPT. https://chatgpt.com/, 2024. Accessed: Feb. 12, 2025.

[58] Jonas Oppenlaender, Johanna Silvennoinen, Ville Paananen, and Aku Visuri. Perceptions and Realities of Text-to-Image Generation. In *Proceedings of the 26th International Academic Mindtrek Conference*, pages 279–288, 2023.

[59] Luke Hughes published. Only Two Weeks in and AI Phenomenon DeepSeek Is Officially Growing Faster than

ChatGPT. https://www.techradar.com/pro/security/only-two-weeks-in-and-ai-phenomenon-deepseek-is-officially-growing-faster-than-chatgpt, 2025. Accessed: Feb. 13, 2025.

[60] Martin Ragot, Nicolas Martin, and Salomé Cojean. AI-generated vs. Human Artworks. A Perception Bias Towards Artificial Intelligence? In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2020.

[61] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):83:1–83:22, 2022.

[62] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Re-Imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, page 328, 2021.

[63] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 1616–1628, 2022.

[64] H. Jeff Smith, Tamara Dinev, and Heng Xu. Information Privacy Research: An Interdisciplinary Review. *MIS Quarterly*, 35(4):989–1016, 2011.

[65] Nasim Sonboli, Jessie J. Smith, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. Fairness and Transparency in Recommendation: The Users' Perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 274–279, 2021.

[66] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. *Big Data & Society*, 9(2):205395172211151, 2022.

[67] Tom van Nuenen, Jose Such, and Mark Cote. Intersectional Experiences of Unfair Treatment Caused by Automated Computational Systems. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–30, 2022.

[68] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):327:1–327:39, 2021.

[69] Sahil Verma and Julia Rubin. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*, pages 1–7, 2018.

[70] Sarah E. Walsh and Karen M. Feigh. Mental Models of AI Performance and Bias of Nontechnical Users. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4116–4121, 2023.

[71] Clarice Wang, Kathryn Wang, Andrew Y. Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan. When Biased Humans Meet Debiased AI: A Case Study in College Major Recommendation. *ACM Transactions on Interactive Intelligent Systems*, 13(3):17:1–17:28, 2023.

[72] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[73] Marley W. Watkins. Exploratory Factor Analysis: A Guide to Best Practice. *Journal of Black Psychology*, 44(3):219–246, 2018.

[74] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of Risks Posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 214–229, 2022.

[75] Mohammad Yaghini, Andreas Krause, and Hoda Heidari. A Human-in-the-loop Framework to Construct Context-aware Mathematical Notions of Outcome Fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 1023–1033, 2021.

[76] Chien Wen (Tina) Yuan, Nanyi Bi, Ya-Fang Lin, and Yuen-Hsien Tseng. Contextualizing User Perceptions about Biases for Human-Centered Explainable Artificial Intelligence. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.

[77] Emily S. Zhan, María D. Molina, Minjin Rheu, and Wei Peng. What Is There to Fear? Understanding Multi-Dimensional Fear of AI from a Technological Affordance Perspective. *International Journal of Human–Computer Interaction*, pages 1–18, 2023.

[78] Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–26, 2024.

A Survey instrument

Consent and Prolific ID
Welcome to our survey on AI Chatbots and AI assistants! Your experiences and opinions will help us understand how we can make AI systems better for everyone. In the following, we explain the contents of the study and how we respect your privacy.
A consent form is shown, and consent must be given to continue.

1. Do you consent to participating in this study?
   - Yes, continue to the survey.
   - No, return to Prolific.

2. Please input your Prolific ID below (note: this text-box should auto-populate)

A **Usage**

(1) AI chatbot usage product Which, if any, of the following AI chatbots or AI assistants have you used in the past three months? Please check all that apply.

- ChatGPT
- Bing AI
- Gemini by Google
- Bard by Google
- Claude by Anthropic
- Copilot by Microsoft
- Perplexity
- YouChat
- OpenAI Playground
- HuggingChat
- Sparrow by DeepMind
- ChatSonic
- PDF.AI
- Other, please specify: _____
- I have used an AI chatbot or AI language tool in the past three months, but I do not remember what it was.
- I have not used an AI chatbot or AI language tool in the past three months.

(2) AI chatbot usage task Which, if any, of the following have you used an AI chatbot or AI language tool to do or assist you with in the past three months? Please check all that apply.

- Answer a question, using the chatbot instead of a search engine
- Have it explain something
- Write, re-write, or edit something to accomplish a task
- Come up with ideas, e.g., for work, school, or private purpose
- Have it summarize a longer piece of text
- Get recommendations, such as where to go eat and so on
- Have a conversation with someone
- Translate something from one language to another
- Generate computer code
- Come up with travel plans
- Other: _____
- No particular task, I just wanted to see what it was like

(3) Frequency usage of AI chatbots

How often do you use an AI chatbot or AI language tool?

1 - Every day
2 - More than once a week
3 - Once a week
4 - Once a month
5 - Less than once a month
6 - Never

B **Attitudes**

Thank you! In the following, we would now like to learn more about your experience with AI chatbots and AI assistants, and what you think about them.

(1) **Risks**

Please think about your experience with AI chatbots. In your opinion, how likely is it that an AI chatbot generates the following outputs?
Measurement items reported in Table 2.
Additional item:

- (Attention check) I have been to every country in the world.

Participants indicated their response on the following scale:

1 - Very unlikely
2 - Unlikely
3 - Neutral
4 - Likely
5 - Very likely

(2) **Human-Computer Trust Scale (HCTS)**

Please think about your experience with AI chatbots. In your opinion, to what extent do you agree or disagree with the following statements about AI chatbots?
Measurement items reported in Table 9.
Additional item:

- (Attention check) A dolphin is an animal.

Participants indicated their agreement on the following scale:

1 - Strongly disagree
2 - Disagree
3 - Neither agree nor disagree
4 - Agree
5 - Strongly agree

(3) **Fairness**

Based on your experience, to what extent do you agree or disagree with the following statements on AI chatbots?

Measurement items reported in Table 1.

Participants indicated their agreement on the following scale:

1 - Strongly disagree
2 - Disagree
3 - Neither agree nor disagree
4 - Agree
5 - Strongly agree

C **Demographics**

(1) What is your gender? Check all that apply.

- Male
- Female
- Non-binary / third gender
- I identify with a different description of my gender, not listed
- Prefer not to answer

(2) Do you identify as transgender?

- Yes
- No
- Prefer not to answer

(3) How would you describe your sexual orientation? Check all that apply.

- Allosexual
- Asexual
- Bisexual
- Gay
- Lesbian
- Queer
- Heterosexual
- Homosexual
- Monosexual
- Pansexual/uid
- Polysexual
- Questioning
- I have a different description for my sexual orientation, not listed above
- Prefer not to answer

(4) What racial or ethnic groups do you identify with? Check all that apply.

- American Indian or Alaska Native
- Black or African-American
- East or Southeast Asian
- Indian subcontinent (including Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka)
- LatinX, Latino, Hispanic or Spanish Origin
- Middle Eastern or North African
- Native Hawaiian or other Pacific Islander
- White
- Other
- Prefer not to answer

(5) What is your age?

(6) What is the highest level of education you have completed?

- High school or lower, e.g., degree/diploma or GED
- Associate's degree
- Bachelor's degree
- Master's degree
- Professional degree (MD, PhD, JD, etc.)
- Prefer not to answer

(7) What is your total household annual income?

- $0 to $19,999
- $20,000 to $49,999
- $50,000 to $89,999
- $90,000 to $129,999
- $130,000 to $149,999
- $150,000 or more
- Prefer not to answer

(8) Do you identify with having any disability?

- Yes
- No
- Prefer not to answer

(9) Is English your first language?

- Yes
- No
- Prefer not to answer

If "No" was selected in the previous question

(10) What is your first language?
[Free text]

D **Conclusion**

(1) Is there any feedback you want to provide us?
[Free text]

B Demographics

Table 8: Participant demographic data (N=123).

Age			Gender			Ethnicity			Education		
	N	%		N	%		N	%		N	%
18–24	19	15.5%	Female	61	49.6%	Black/Af-Am	16	13.0%	High school or less	46	37.4%
25–34	28	22.8%	Male	61	49.6%	Hispanic	6	4.9%	Associate's degree	20	16.3%
35–44	33	26.8%	Transgender	2	1.6%	Asian Indian	4	3.3%	Bachelor's degree	35	28.5%
45–54	24	19.5%				White	78	63.4%	Master's degree	14	11.4%
55–64	11	8.9%				Other/Mixed	19	15.5%	Professional degree	7	5.7%
65+	8	6.5%									
Income			**Disability Status**			**Sexual Preference**			**AI Chatbot Use**		
	N	%		N	%		N	%		N	%
$0–19,999	13	10.6%	Yes	20	16.3%	Non-Heterosexual	27	22.0%	Never	3	2.4%
$20k–49,999	35	28.5%							< Once/month	12	9.8%
$50k–89,999	30	24.4%							Once/month	16	13.0%
$90k–129,999	21	17.1%							Once/week	18	14.6%
$130k–149,999	7	5.7%							> Once/week	50	40.7%
$150k+	14	11.4%							Daily	24	19.5%

C Scales

Table 9: Human-Computer Trust Scale (HCTS) [25] adapted to AI chatbots.

Item Number	Item
hcts_scale_1	I believe that there could be negative consequences when using AI chatbots. (R)
hcts_scale_2	I feel I must be cautious when using AI chatbots. (R)
hcts_scale_3	It is risky to interact with AI chatbots. (R)
hcts_scale_4	I believe that AI chatbots will act in my best interest.
hcts_scale_5	I believe that AI chatbots will do its best to help me if I need help.
hcts_scale_6	I believe that AI chatbots are interested in understanding my needs and preferences.
hcts_scale_7	I think that AI chatbots are competent and effective in helping me with what I use them for.
hcts_scale_8	I believe that AI chatbots have all the functionalities I would expect from them.
hcts_scale_9	If I use AI chatbots, I think I would be able to depend on them completely.
hcts_scale_10	I can always rely on AI chatbots for the things I use them for.
hcts_scale_11	I can trust the information presented to me by AI chatbots.

Table 10: Factor loadings of the adapted Human-Computer Trust Scale (HCTS) [25] when trying to replicate the original factor structure, with loadings $> |.40|$ indicated as bold.

Item Number	Sub-construct	Factor 1	Factor 2	Factor 3	Factor 4	Communality
hcts_scale_1	Risk	.22	**.66**	-.03	-.05	.49
hcts_scale_2	Risk	.21	**.67**	-.04	-.02	.50
hcts_scale_3	Risk	-.25	**.86**	.08	.09	.82
hcts_scale_4	Benevolence	.11	.16	**.58**	-.01	.37
hcts_scale_5	Benevolence	-.08	.06	**.89**	-.03	.81
hcts_scale_6	Benevolence	.29	-.16	**.59**	.07	.46
hcts_scale_7	Competence	.01	.05	.00	**.97**	.95
hcts_scale_8	Competence	**.63**	-.10	.15	-.05	.43
hcts_scale_9	Reciprocity	**.75**	.11	-.06	-.09	.58
hcts_scale_10	Reciprocity	**.66**	-.03	-.04	.22	.49
hcts_scale_11	Reciprocity	**.55**	.14	.11	.04	.34

Table 11: Factor loadings of the developed instrument on perceived fairness in AI chatbots informed by Colquitt and Rodell [10], with loadings $> |.40|$ indicated as bold.

Item Number	Factor 1	Factor 2	Communality
fairness_scale_1	**.62**	.12	.40
fairness_scale_2	**.75**	-.07	.57
fairness_scale_3	**.59**	.01	.35
fairness_scale_4	**.81**	.07	.66
fairness_scale_5	**.72**	.12	.53
fairness_scale_6	**.77**	.11	.60
fairness_scale_7	.08	**.64**	.42
fairness_scale_8	-.14	**.92**	.86
fairness_scale_9	**.90**	-.19	.85
fairness_scale_10	**.45**	.29	.28

Table 12: Factor loadings of the developed instrument on perceived risk in AI chatbots informed by Weidinger et al. [74] and Goyal et al. [22], with loadings $> |.40|$ indicated as bold.

	Factor 1	Factor 2	Factor 3	Communality
risks_scale_1	-.02	**.92**	-.03	.85
risks_scale_2	**.62**	.25	-.05	.45
risks_scale_3	-.14	-.05	**.84**	.72
risks_scale_4	.07	.03	**.74**	.55
risks_scale_5	.07	.15	**.59**	.38
risks_scale_6	.11	**.63**	.15	.43
risks_scale_7	**1.0**	-.12	-.02	1.0
risks_scale_8	-.16	**.80**	.06	.68
risks_scale_9	.05	.00	**.77**	.59
risks_scale_10	.18	.40	.37	.33
risks_scale_11	**.91**	-.19	.04	.87
risks_scale_12	.24	**.55**	.10	.38
risks_scale_13	**.69**	.22	-.05	.52