

Examining the Effect of Explanations of AI Privacy Redaction in AI-mediated Interactions

Roshni Kaushik¹, Maarten Sap², Koichi Onoue¹

¹Fujitsu Research of America

²Carnegie Mellon University

rkaushik@fujitsu.com, msap2@andrew.cmu.edu, konoue@fujitsu.com

Abstract

AI-mediated communication is increasingly being utilized to help facilitate interactions; however, in privacy sensitive domains, an AI mediator has the additional challenge of considering how to preserve privacy. In these contexts, a mediator may redact or withhold information, raising questions about how users perceive these interventions and whether explanations of system behavior can improve trust. In this work, we investigate how explanations of redaction operations can affect user trust in AI-mediated communication. We devise a scenario where a validated system *removes* sensitive content from messages and *generates explanations* of varying detail to communicate its decisions to recipients. We then conduct a user study with 180 participants that studies how user trust and preferences vary for cases with different amounts of redacted content and different levels of explanation detail. Our results show that participants believed our system was more effective at preserving privacy when explanations were provided ($p < 0.05$, Cohen's $d \approx 0.3$). We also found that contextual factors had an impact; participants relied more on explanations and found them more helpful when the system performed extensive redactions ($p < 0.05$, Cohen's $f \approx 0.2$). We also found that explanation preferences depended on individual differences as well, and factors such as age and baseline familiarity with AI affected user trust in our system. These findings highlight the importance and challenge of balancing transparency and privacy in AI-mediated communications and suggest that adaptive, context-aware explanations are essential for designing privacy-aware, trustworthy AI systems.

Extended version — <https://arxiv.org/abs/2603.24735>

1 Introduction

As AI systems are increasingly integrated into everyday communication, researchers have been exploring their role not only as assistants that respond to direct user queries but also as intermediaries that shape the flow of information between people. For example, AI systems are already being used to support multilingual communication through real-time translation (Shahmerdanova 2025), to simplify specialized texts for broader audiences (Araújo and Aguiar 2023), and to remove sensitive information from shared records in

healthcare (Caine and Hanania 2013). In these cases, the AI is no longer a passive tool, but is an active mediator making decisions about what to share with whom and how to present those decisions. Increasing efforts to develop AI intermediaries suggest their potential to become an important participant in human-human interactions.

One particularly challenging and underexplored setting for AI-mediated communication is the preservation of privacy. In high-stakes domains, different parties may have different permissions to access sensitive data. For example, a clinician might share records with a specialist, while an AI system redacts personally identifiable information to comply with privacy regulations. In a workplace setting, an AI could help managers and employees collaborate while removing sensitive HR details. In research collaborations, an AI might enable data sharing between institutions by removing data that particular parties should not have access to. In these settings, the AI mediator does not simply summarize or transmit information, but actively redacts content. This creates a fundamentally different interaction paradigm where the AI alters messages while attempting to preserve trust and understanding between the humans involved. Unlike traditional settings where AI outputs are evaluated for accuracy, the system must balance both privacy protection and interaction satisfaction, raising questions about how users interpret and respond to AI-mediated redactions.

In this work, we examine how humans perceive AI interventions in privacy-preserving scenarios, focusing on AI explanations of redactions, which has been underexplored in previous work. This setting provides a new lens on trust in AI systems, and users must reason about what the AI shows and withholds. In contrast to standard explanation settings, where explanations justify observed model outputs, explanations in these scenarios must justify intentional transformations of the content. This shifts the user's task from evaluating the correctness of visible outputs to reasoning about potentially missing information and correctness of the system's redaction, which introduces additional challenges for calibrating system trust and reliance.

We focus on explanations as they are increasingly incorporated into AI outputs (e.g. reasoning models such as deepseek-r1 (Guo et al. 2025), openai-o1 (Jaech et al. 2024)) and serve as the primary mechanism for explaining the AI's intervention. Prior work suggests that explanations can act

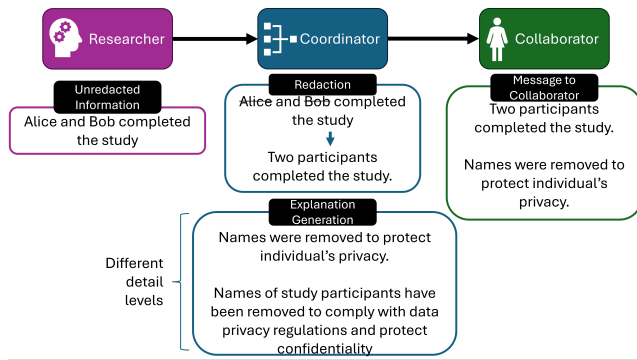


Figure 1: Conceptual overview of AI-mediated communication. The AI **coordinator** mediates between a **researcher** and **collaborator** by redacting sensitive information and providing an explanation to the recipient.

as double-edged swords; they may increase perceived transparency and trust, but they can also lead to over-reliance or reduce user scrutiny of the system (Park et al. 2025). This tension is particularly powerful in privacy-preserving scenarios, where users must decide whether to accept or question an AI’s redaction decisions.

While explanations can clarify why information was withheld and help users understand the system’s behavior, they also introduce design trade-offs. Explanations that are absent or too general may make the AI’s actions seem arbitrary or confusing, while explanations that are too detailed risk overwhelming users or inadvertently disclosing private information. This tension is particularly seen in privacy-sensitive contexts, where longer outputs increase the probability of exposing memorized or private content (Carlini et al. 2019), highlighting the conflict between transparency and privacy.

To study this, we craft a scenario where an AI mediator is tasked with removing any sensitive information that is sent by one party with information access before sending the redacted version to another party without information access. As previous work has shown that information asymmetry can negatively impact the trust of those with access to less information (Bertolin et al. 2008), we designed the AI mediator to provide an explanation to the party without access to improve their trust and explain the redaction actions that the system performed.

More concretely, we devised a communication setup between two parties: (1) an external **collaborator**, a human participant without access to private information, and (2) a **researcher**, simulated for this work and with access. An AI mediator (**coordinator**) facilitates communication between the two parties. As shown in the example in Figure 1, the **coordinator** redacts private information from the **researcher**’s messages and provides explanations of its decisions to the **collaborator**. We verified the technical robustness of our AI mediator by validating its privacy preservation capabilities in our setup.

Using our verified system, we conducted a study with 180 participants to explore how user trust, system understanding, and explanation preferences change by systematically vary-

ing both the amount of sensitive information required to be redacted and the level of explanation detail. Specifically, we ask three research questions:

- Does providing an explanation to the **collaborators** improve their trust in and understanding of the system?
- How does the amount of sensitive information removed from the **researcher**’s message affect the **collaborators**’ explanation preferences?
- Do individual differences (e.g., demographic or attitudinal factors) influence trust and explanation preferences?

Our results show that participants believed our system was more effective at preserving privacy when explanations were provided ($p < 0.05$, Cohen’s $d \approx 0.3$). We also found that contextual factors had an impact; participants relied more on explanations and found them more helpful when the system performed extensive redactions ($p < 0.05$, Cohen’s $f \approx 0.2$). Individual characteristics further influenced trust; for example, those with a baseline familiarity or trust in AI/LLM systems were also more likely to trust our system.

Explanations for AI-mediated communication that preserve privacy present a complex challenge, requiring a careful balance between transparency and protection of sensitive information. Our work highlights this difficulty by exploring how different redaction strategies and explanation styles affect user understanding and trust. By examining these interactions, we shed light on design principles and open questions that can guide future research on creating AI-mediated systems that are both privacy-aware and user-friendly.

2 Related Work

Previous research relevant to our work includes contributions from AI-mediated interaction, privacy-preserving communication, explanation design, and trust in AI systems. Together, these areas highlight interesting intersections between protecting sensitive content and maintaining user understanding.

2.1 AI Mediation of Human Communication

Research in HCI has begun to characterize AI not just as a tool but as an active intermediary that modifies and reshapes interpersonal exchanges. Hancock et al. offer a foundational framework for AI-mediated communication, where intelligent agents modify or generate messages on behalf of a human (Hancock, Naaman, and Levy 2020). They articulate key characteristics, such as the magnitude of involvement, media type (text, audio, video, etc.), optimization goal, autonomy, and role orientation (who the AI is operating on behalf of). This allows us to understand the various ways that AI can intervene in human communication. Other research has shown that trust perceptions in senders diminish when recipients know that AI is involved, but surprisingly trust increased when AI wrote interpersonally rather than transactionally (Liu et al. 2022).

Psychology studies have also revealed how humans attribute agency, responsibility, and trust in mediated contexts. Users often reassess blame when AI mediates, and trust in

the human communicator remains high. However, some responsibility can shift to the AI when interactions fail, known as the moral crumple zone phenomenon (Hohenstein and Jung 2020). Frameworks such as AI-Mediated Exchange Theory (Ma and Brown 2020) explain these types of trust shifts in the balance of social exchange. Research on human perceptions based on the author of certain texts found that only when participants thought they saw a mix of AI and human-written profiles, they mistrusted hosts whose profiles were labeled as or suspected to be written by AI (Jakesch et al. 2019). These works position AI mediation as not only a technical innovation, but also a social intervention that has implications for trust and sense of agency. However, prior work has not yet examined how AI mediation shapes trust and understanding when the system is actively altering messages for privacy reasons. Our work addresses this gap by studying explanations as a mechanism to support trust in privacy-preserving AI mediation.

2.2 Privacy-Preserving Communication

Privacy is a critical concern in mediated communication, especially in domains like healthcare, workplace collaboration, and research. Nissenbaum's foundational framework for contextual integrity argues that privacy is maintained when information flow follows contextual norms (Nissenbaum 2004). HCI research has explored many methodologies to protect sensitive information while enabling collaboration, including for selective redaction, access controls, and data anonymization (Caine and Hanania 2013).

Advances in natural language processing have enabled automatic detection and redaction of private information, but these methods face trade-offs between privacy and utility (Sousa and Kern 2023; Pal et al. 2024). Redacting too much information can reduce message clarity, but not redacting enough can lead to privacy leaks. Security and privacy research has begun to systematically evaluate these trade-offs. For example, recent work has investigated the vulnerabilities of LLMs in the extraction of personal information and the effectiveness of redaction as a countermeasure (Liu et al. 2024; Cheng et al. 2025). Studies have also explored how contextual integrity can serve as a framework for auditing privacy risks with language models (Miresghallah et al. 2023).

From an HCI perspective, there have been calls for research on how LLM design paradigms affect disclosure behaviors, mental models and preferences for privacy control, and the design of tools, systems, and artifacts that empower users to reclaim ownership over their personal data (Li et al. 2024). Setting the correct expectations for AI can drastically change the users' perspective of the system and interaction (Kocielnik, Amershi, and Bennett 2019), and adding design paradigms, such as explanations, must balance informativeness while preserving privacy. Other efforts have emphasized the need for user empowerment in the privacy preservation process through interactive AI tools (Sun, Zhou, and Jiang 2024). However, it remains unclear how different explanation strategies interact with the varying levels of redaction that a system might perform to impact user trust and comprehension. Our work addresses this open question

by systematically studying these factors in an AI-mediated, privacy-preserving communication scenario.

2.3 Explanations in AI-mediation

Explanations are a central mechanism to help users understand AI behavior. Work in explainable AI has shown that explanations influence how people understand, predict, and evaluate system decisions (Miller 2019). HCI researchers emphasize that explanations must align with users' goals and preferences to improve human-AI collaboration (Ehsan et al. 2019). Design-oriented reviews of human-XAI interaction further highlight that explanations should be interactive, context-aware, and tailored to the needs of different users to support effective decision-making (Chromik and Butz 2021). When AI acts as a mediator, explanations can clarify why information was altered or withheld and can help users maintain a coherent understanding of the interaction. Recent work has also found that AI-generated explanations could have a large impact on human trust, and factors such as the AI background of the user can heavily influence the interpretation of explanations (Ehsan et al. 2024). Prior research also suggests that individual user characteristics can meaningfully shape how people interpret and rely on explanations in decision-making contexts (Khadar et al. 2025). This motivates our addition of background and personality questions as part of our user questionnaire.

Beyond usability concerns, explanations are also closely tied to broader discussions of accountability and ethics in AI systems. Researchers argue that transparency mechanisms such as explanations are essential for holding algorithmic systems accountable and enabling meaningful oversight (Busuioc 2021). At the same time, ethical analyses of AI systems emphasize the importance of transparency alongside bias mitigation and responsible system design to ensure trustworthy deployment of AI technologies (Mensah 2023).

However, explanations in mediated communication can raise unique challenges. Explanations can act as double-edged swords by increasing perceived transparency and trust, but also promoting to over-reliance or reduce user scrutiny of the system's outputs (Park et al. 2025). This tension is particularly powerful in privacy-preserving scenarios, where users must decide whether to accept or question an AI's decision to redact information. Surveys of privacy-preserving explainable AI highlight that explanations themselves may leak sensitive details if not designed carefully (Dosilovic, Brcic, and Hlupic 2018). Studies on attacks against explanations from models have shown that adversaries can sometimes infer private information from outputs that appear to be anonymized (Shokri, Strobel, and Zick 2021). Reducing the amount of information disclosed to combat this can come at the cost of informativeness (Nguyen et al. 2024). Systems have been developed that help users identify sensitive information and summarize relevant privacy policies, which has improved users' understanding of data practices and privacy risks (Chen et al. 2025). Additionally, prior work (Krsek et al. 2025) has developed a tool to help users catch accidental private information leakage and found that the AI should account for context and user differences when providing explanations for privacy modifi-

cations.

Together, these findings suggest that explanations can have a powerful impact in privacy contexts, but issues such as context-dependence and accidental privacy leakage must also be considered. How people actually experience and evaluate explanations when an AI mediates their communication with another human is an underexplored area. Our work addresses this gap by examining how both explanation detail and redaction amount shape preferences, trust, and understanding in a privacy-sensitive, AI-mediated setting.

2.4 Trust in AI and Human-AI Collaboration

Trust has long been recognized as a critical factor in the way people interact with intelligent systems. Lee et al. (Lee and Moray 1992) explored how trust influences control strategies and how responsibility was allocated between humans and systems. Later frameworks such as the Intelligent Systems Technology Acceptance Model integrated transparency and acceptance into intelligent system models (Vorm and Combs 2022). In human-AI collaboration, maintaining the appropriate level of trust is critical. Over trust can lead to overreliance, while under trust can lead to misunderstandings and lower use of the system.

Recent HCI research explores these dynamics in AI-mediated decision making. Trust must be understood from the perspective of those who use AI systems and from those affected by their outputs, resulting in a complex multi-stakeholder dynamic (Vereschak et al. 2024). A systemic overview of models and trust measures in (Ueno et al. 2022) details the variety of approaches that can be applied in different scenarios. Explainability is often explored as a way to calibrate user trust and can help users understand system outputs and limitations (Kim et al. 2023). Prior work also distinguishes between attitudinal trust (a user’s belief that a system is reliable) and behavioral reliance (the extent to which users actually follow or act on the system’s outputs). Recent research on explainable reliance-aware user interfaces emphasizes designing explanations that help users decide when to accept, question, or overrule AI recommendations (de Souza Filho, Belloum, and de Oliveira 2024). We focus on attitudinal trust in this work.

However, most of this research considers trust in systems that provide information rather than those who also can choose to withhold it. When an AI redacts information for privacy reasons, users must assess the reliability of the remaining information under conditions of incomplete information, which then shapes their trust in the system. Our work addresses this open research question by investigating trust in AI mediation, where explanations and privacy-preserving redactions intersect.

3 AI-Redaction System Design

We devise a scenario where our system (Figure 2) enables human-AI collaboration and mediates between two human roles: **researcher** and external **collaborator**. The **researcher** has performed some research (such as a user study) that they would like to share with the **collaborator**. However, there is information in the results, such as names and

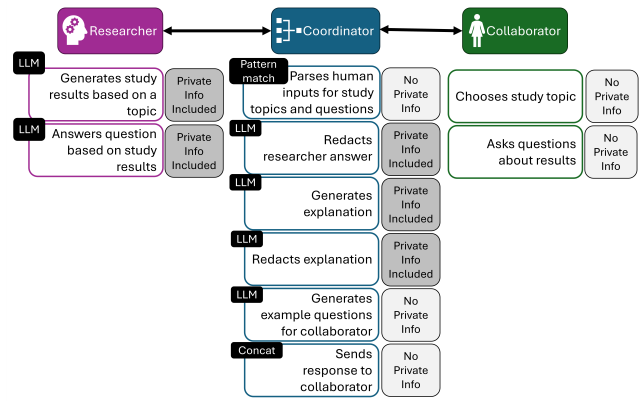


Figure 2: Diagram of the interactions between the **researcher**, **collaborator**, and **coordinator**, with tasks performed by each that include and do not include private information indicated.

genders, that the **collaborator** cannot have access to. Our system sits between these two roles and performs privacy-preserving operations on the information the **researcher** is sharing to ensure that any information the **collaborator** does not have access to is removed. The **collaborator** can ask questions about the information to the **researcher**, and our system will remove any private information in the **researcher**’s answer before sending it to the **collaborator**.

In the system we implemented for this work, the human is the external **collaborator**, and the **researcher** is simulated by an LLM, as we are interested in the **collaborator**’s understanding, trust, and explanation preferences. We focus on the **collaborator** perspective because **collaborators**, unlike **researchers**, must rely on AI explanations to interpret missing information, making their explanation needs particularly important in privacy-sensitive interactions (Miller 2019; Hoff and Bashir 2015). Additionally, we consider a relatively simple redaction task, the removal of Personally Identifiable Information (PII); however, our approach could be extended in future work to more complex privacy-preserving scenarios.

Figure 2 illustrates the tasks that the **researcher**, **collaborator**, and **coordinator** perform in this interaction, as well as which of these tasks include and do not include private information. The **researcher** and the **coordinator** both have access to the private information. The **researcher** shares the unredacted study results and answers to the **collaborator** questions. The **coordinator** takes that information, removes any private information, and generates an explanation for the **collaborator** that explains the privacy-preserving operations that were carried out on the **researcher**’s information. The **coordinator** also performs some tasks for the purposes of our particular interaction and study that do not include private information, such as parsing the **collaborator**’s input, compiling the redacted response for the **collaborator**, and generating some sample questions that could be interesting to ask the **researcher**. The extended version includes an excerpt from an example interaction between the **coordinator** and the **collaborator**. GPT 4.1 was used for all LLM calls in this work, and more prompting details are included in the extended version.

3.1 Study Topics and Original Information

We generated five different study topics that did not require specialized knowledge to comprehend and could be interesting/applicable to most people. These topics were (1) the impact of gardening on mental health, (2) the effects of exercise on sleep, (3) the relationship between diet and mood, (4) the influence of social media on self-esteem, and (5) the role of mindfulness in stress reduction.

For each study topic, we generate a set of mock but realistic user study results that included some private information about participants (e.g., names, ages, genders, ethnicities) and some general information that was not private (e.g., amount of sleep each night). We defined private information using standard NLP/HCI criteria (Schwartz and Solove 2011; Song et al. 2025), treating any attribute that could directly or indirectly identify a participant as private, and did not apply formal guarantees like k-anonymity. Both the study topic and study results were with GPT 4.1, and we manually verified that the generated information included both private and general information.

3.2 Questions and Amount of Sensitive Information

We then generated 25 questions in three different redaction categories: high, moderate, and low. These redaction categories represent the amount of sensitive information present in the **researcher**'s answer to a question that needs to be removed before it can be sent to the **collaborator**. Three examples of questions the **collaborator** are included below, one from each redaction category.

- **High:** What did the oldest participant say about how exercise affected their sleep?
- **Moderate:** How did women in the study describe their gardening experiences compared to indoor activities?
- **Low:** How did participants describe the difference in their sleep quality between the two conditions?

The answer to the high redaction question results in identifying this participant, so most/all of the answer should be removed. In contrast, the answer to the low redaction question requires little to no removal of sensitive information. All questions and answers were generated with GPT 4.1, and we manually verified that the answers to each question were correct based on the corresponding user study results. Additionally, we ensured that the questions were realistic and representative of plausible **collaborator** queries by verifying they were grounded in the types of analyses typically performed on user study data (e.g., comparing participant groups, examining condition differences, summarizing qualitative responses) (Braun and Clarke 2006).

We next created an LLM judge to validate that the amount of sensitive information present in the answers to these questions was sufficiently different in each redaction category. Note that this is a different LLM judge to the one presented below to verify privacy protection. We found that the questions in each redaction category were sufficiently different; specifically the generated questions in the high redaction category contained more sensitive information than the moderate category and the moderate category contained more

sensitive information than the low category. Further details are included in the extended version.

3.3 Redacting Sensitive Information from Answers

We then redacted sensitive information from all the **researcher** answers using GPT4.1. We used a validated LLM judge (see Section 3.4) to test whether there was any private information remaining in the full set of redacted answers. We found that our system protected private information for at least 98% of the data in each redaction category (more details in the extended version).

3.4 Data Validation of Privacy Preservation

To begin the validation process, we first assigned two human coders to determine whether privacy was preserved for a subset of data (60 randomly selected pieces of data from each redaction category). The human coders agreed for all of the data they evaluated, which allows us to rely on their responses for this subset.

We next set up an LLM judge to test whether private information was present after redaction was performed, motivated by prior research that shows that LLMs are effective at identifying and handling PII (Krčo et al. 2026). We tested two LLM judges (GPT 4.1 and Claude Opus 4.7) and compared their evaluations to each other on the same data subset and found perfect agreement between judges (100%), indicating a lack of self-bias, which could occur when generating and evaluating with the same model. Finally, we verified alignment between the LLMs and human coders and found high agreement between them (specifically 93.9% agreement between the human coders and GPT 4.1). This indicates that the result of our LLM judge is reliable in addition to not being a result of self-bias.

3.5 Explanations of Different Types

Once we generated the answer and redacted answer to each question, we generated explanations in two distinct styles. Building on prior work (Larasati, De Liddo, and Motta 2020) which explored how explanation styles influence trust, we further elaborate the distinction between *general* and *thorough* explanations as two points along a broader explanation design spectrum. In particular, we distinguish these styles by length and their level of abstraction and completeness.

The *general* explanation is concise (1-2 sentences in our setting) and communicates high-level information about why information was withheld, without detailing the specific steps the system performed. In contrast, the *thorough* explanation is longer (3-5 sentences) and provides a more complete, step-by-step account of how the system identified and removed private information. This distinction allows us to compare explanations that prioritize conciseness against those that prioritize transparency and detail. Examples of these two explanation styles for the same answer and redacted answer are shown in the extended version.

We focus on these two explanation styles to explicitly explore the trade-offs introduced by increasing explanation detail. We hypothesize that more thorough explanations may

reduce user misunderstanding and increase perceived transparency. However, longer and more detailed explanations increase the likelihood that an LLM may inadvertently reveal sensitive information or overwhelm the user with too much detail. In contrast, more general explanations may better preserve privacy and reduce user cognitive load, but risk not including enough detail.

As we show in the extended version, private information is being leaked in the thorough explanations, but not in the general explanations, with the validation performed by the verified LLM judge in Section 3.4. The longer explanations led to a greater amount information leakage. We therefore decided to redact private information from the generated explanations using the same GPT4.1 prompt used to redact the answers. After information redaction, we found that there was no private information remaining in the explanations (see graph in the extended version).

4 User Study Design

To explore our research questions, we conducted a user study (protocol approved by an internal ethics committee). The participants interacted with an LLM-simulated **researcher** through the AI **coordinator**, which applied appropriate redactions and provided explanations. Our two independent variables were explanation style (no explanation, general, thorough) and redaction needed (high, moderate, low). Through a survey, we measured our dependent variables of perceived trust in the system, subjective understanding of the system, and explanation preferences.

4.1 Participants

We first performed a power analysis to determine the appropriate number of participants to recruit. With three explanation styles, we determined that 180 participants (60 per group) should allow us to detect between a small effect ($f = 0.20$ results in a power of 0.661) and medium effect ($f = 0.5$ results in a power of 1.000). We recruited all our participants from Prolific with a qualification of being fluent in English and compensation of \$4 per participant. Their average age was 33.0, with 1.1% aged 18 – 19, 45.7% aged 20 – 29, 30.9% aged 30 – 39, 13.1% aged 40 – 49, 6.3% aged 50 – 59, and 2.3% aged 60 – 69. 52.6% identified as male and 47.4% as female. 3.4% identified as Asian, 70.9% as Black, 18.3% as White, and 7.4% as Mixed.

4.2 Study Procedure

The participants first completed a brief introduction process that included instructions explaining their role as the **collaborator** interacting with a **researcher** through an AI **coordinator**. They were informed that some information might be withheld or modified from the **researcher's** responses and that the **coordinator** might provide some explanation for those modifications. They then completed a 5-question multiple choice quiz to ensure they understood the scenario, where they had to complete each question correctly (with multiple attempts allowed).

The participants were randomly sorted into one of three explanation conditions: no explanation, general explanation,

and thorough explanation. The **coordinator** then presented them with a list of topics and instructed them to choose one they found most interesting. The **coordinator** then responded with a summary of the findings on that topic and a list of questions that the participant could choose from to follow up. These questions were from one of the redaction categories (high, moderate, or low). The participant then chose a question from the provided list, and the **coordinator** responded with the redacted **researcher** answer, a redacted explanation based on the participant's explanation condition, and a set of new follow-up questions in a different redaction category. This process continued for three questions, and the order of the redaction categories was randomized for each participant. This resulted in a between-subjects design for the first independent variable of explanation detail and a within-subjects design for the second independent variable of redaction amount.

4.3 Measures and Hypotheses

After three questions and answers, the participant was redirected to the survey, which had three sections. For the first two sections, a portion of the evaluation items were modified from (Hoffman et al. 2019) and (Jian, Bisantz, and Drury 2000), including an explanation satisfaction scale, a trust scale for the explainable AI context, and a checklist for trust between people and automation. The full questionnaires are included in the extended version.

Evaluation intended to explore explanation detail preferences The first section was a series of evaluation items per interaction, one for each of the question-answer-explanation interactions the participant experienced. Each question, modified researcher answer, and explanation (if any) given were shown to the participant (example in Table 1) to refresh their memory, and we asked five items (four 5-pt Likert, one free form) for each of these interactions, resulting in a total of 15 items in this section of the survey. These items were intended to explore how the amount of redaction in an interaction affected the participant's trust and understanding in the system and explanation preferences.

Evaluation intended to explore explanation detail and redaction amount interaction In the second section, we asked participants 8 evaluation items (7 5-pt Likert, 1 free form) about their overall experiences with the system to gauge their general explanation preferences and how the explanations that they received affected their trust and system understanding.

Evaluation intended to explore interaction between user characteristics and trust in our system In the last section, we included demographic evaluation items and some evaluation items about their attitudes and experience with AI. Previous research has found that background can impact trust in automated systems and the effect of explanations on that trust (Ehsan et al. 2024), so we wanted to obtain background information on our participants to see if there were correlations between their responses to these evaluation items and their experiences in the study.

Question from You:	In what ways did physical activity or being outdoors influence participants' experiences?
Researcher Answer Modified by Coordinator:	Physical activity and being outdoors during gardening appeared to significantly enhance participants' experiences, with several noting increased calmness, happiness, and reduced stress. Participants described feeling energized, uplifted, and refreshed after gardening, emphasizing the positive effects of fresh air, sunlight, engagement with plants, and movement. These benefits were not consistently observed in the indoor relaxing activities, which were often described as less satisfying and less effective in providing stress relief or mood improvement.
Explanation Given by Coordinator:	In the privacy preservation process, sensitive or personally identifiable details are removed or generalized to protect participant anonymity while maintaining the integrity of the message. Verification ensures that the modified message continues to accurately reflect the original content's intent and factual basis, without introducing distortion or losing key concepts. Summarization is applied to condense the text, omitting repetitive or overly detailed descriptions, but retaining all major findings and comparative points in the experiences shared. The overall process flow begins with reviewing the original message for private or excess information, followed by a review to verify accuracy. This ensures that the final message is concise, privacy-safe, and still conveys the essential information about the effects of gardening and indoor activities on participant well-being.

Table 1: Example of the question, redacted answer, and explanation (if provided) that are included for each per-interaction evaluation group

Hypotheses We tested three hypotheses through our user study that address our research questions.

- (H1): People will prefer an explanation over no explanation
- (H2): Explanation preferences will vary between different redaction categories
- (H3): People will have different levels of trust in the system, based on their background and experiences

5 Results and Discussion

5.1 Explanations vs. No Explanations

We first analyzed responses from the second part of the survey (Section 4.3), where participants answered Likert questions about their overall experiences with the system. Participants were divided into two groups for this analysis: those who received an explanation and those who did not. We normalized the (1 – 5) responses to a 0 – 1 scale for plotting simplicity, and mean ratings with standard deviation bars are shown in Figure 3. We performed a Welch ANOVA (due to unequal variances) to compare between these two groups. We found a significant effect of explanation on participants' trust in the system's ability to preserve privacy ($F(1, 101) = 3.95, p < 0.05$, Cohen's $d = 0.337$). Responses to the remaining questions did not reach significance, although trends were consistent with expectations. Participants who received explanations tended to report higher confidence in the system, perceived outputs as more predictable, considered the system more reliable and safe, and rated it as outperforming a novice user, although these differences were not significant.

We also examined participants' engagement from their free-text responses. We defined engagement by identifying minimal responses, defined as containing no response or keywords such as 'no comment', 'none', 'n/a', 'nothing', 'no comments', and 'no additional'. In addition to identifying minimal responses, we also conducted a sentiment analysis using the Natural Language Toolkit (Ma, Peng, and Cambria 2018). Results, summarized in Table 2, show that participants who received explanations were more engaged,

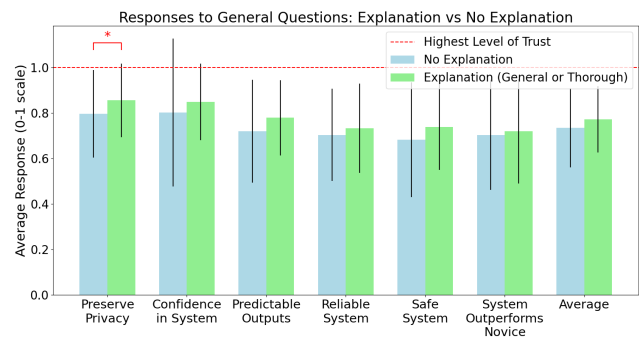


Figure 3: Responses of participants to Survey Section 4.3 about their overall experiences with the system, normalized to 0 – 1. The last set of bars illustrate the average across all questions. The value corresponding to the most positive response is shown in red. Significant differences are shown in red with one asterisk for $p < 0.05$.

less likely to provided minimal responses, and more likely to write positive comments. They were also less likely to provide negative comments, indicating that explanations not only influence quantitative trust ratings but also positively affect subjective evaluations and engagement.

H1: People will prefer an explanation over no explanation These results provide partial support for the hypothesis. Participants who received explanations rated the sys-

	No Explanation	General + Thorough Explanations
Minimal	34.0%	24.0%
Positive	56.4%	60.2%
Neutral	33.3%	35.2%
Negative	10.3%	4.5%

Table 2: Analysis of the minimal responses and positive/neutral/negative responses from free-response, separated by participants who received and did not receive an explanation.

tem as more effective in protecting privacy than those who did not, indicating that explanations can improve trust in privacy-preserving systems.

The pattern was reinforced by participants' open-ended responses. Table 2 shows that explanations increased engagement, encouraged positive comments, and reduced negative feedback. Several participants who did not receive explanations expressed a desire for more information ("The system sometimes gave useful and detailed answers but other times withheld too much information without explanations."), whereas others were not bothered by limited answers ("I had an easy experience with the system. [I]t did what it was supposed to despite not answering that one question that had to do with age."). Participants who received explanations highlighted the value of transparency: one noted, "Better to honestly say that information cannot be revealed", while another said, "The system appears to be trustworthy with the user's private information." These comments suggest that explanations signal reasoning and intent, which reassure users and support trust.

These findings are consistent with prior research in explainable AI showing that explanations increase trust by improving perceived transparency (Kizilcec 2016). Additionally, our results suggest that even for a relatively simple privacy-preservation task, such as PII removal, providing explanations can positively influence how users interpret and interact with the system. Beyond supporting user trust, explanations also support user agency by enabling users to assess whether system decisions align with their expectations. From a socio-technical perspective, this highlights that explanations shape not only perceptions of the system, but also the broader relationship between users and AI systems. When relevant information is withheld or explanations are unclear, users' ability to critique the system's assumptions or identify potential errors is limited. Explanations could therefore be used as mechanisms for increasing trust and as tools to support user investigation, accountability, and more informed human-AI interaction.

5.2 Preference Variances by Amount of Redaction Needed

We analyze responses from the first part of the survey (Section 4.3), where participants answered Likert questions for each of the three interactions they experienced in the study, corresponding to low, moderate, and high redaction cases. Participants were grouped according to the type of explanation they received: none, general, or thorough. Their responses (1–5) were normalized to a (0–1) for plotting simplicity, and Figure 4 shows the mean ratings with standard deviation bars for each of the four questions, illustrating how the perceptions changed across redaction levels.

Because the groups exhibited unequal variances, we used a Welch ANOVA with Games–Howell post hoc tests to identify significant differences, which are marked in the plot. Significant differences emerged for the first question (top-left plot), which asked how well the modified answer matched participants' expectations. For participants who did not receive an explanation, there was a significant effect of redaction level on perceived expectation-match

($F(2, 115.63) = 5.29, p < 0.01$, Cohen's $f = 0.218$), with high redaction cases matching expectations significantly more than low redaction cases ($p < 0.01$). Participants who received a general explanation showed a similar pattern ($F(2, 105.21) = 10.35, p < 0.001$, Cohen's $f = 0.305$), again rating high redaction cases as significantly more aligned with their expectations than low redaction cases ($p < 0.001$). Across both groups, greater system intervention was associated with a stronger sense that the modified answer was appropriate or expected.

Significant differences also emerged for the fourth question (bottom-right plot), which examined how well the explanation helped participants judge the system's accuracy, reliability, and trustworthiness. Among participants who received a general explanation, redaction level significantly affected their ability to assess the system ($F(2, 107.25) = 3.26, p < 0.05$, Cohen's $f = 0.2035$), with moderate redaction cases yielding better assessments than low redaction cases. Participants who received a thorough explanation showed a similar effect ($F(2, 112.02) = 3.47, p < 0.05$, Cohen's $f = 0.189$), with high redaction cases producing better assessments than low redaction ones. In both explanation conditions, the explanation was more helpful when the system performed more extensive modifications.

H2: Explanation preferences will vary between different redaction categories Taken together, these results support our hypothesis. Participants relied more on explanations and found them more helpful, when the system performed more redaction. When the system intervened heavily, explanations provided crucial context that aided participants in evaluating whether the output was trustworthy and appropriate. In contrast, when the system performed minimal modification, explanations were less necessary, offering little additional benefit. This pattern indicates that the usefulness of explanations is sensitive to the degree of system intervention and that provided explanations should scale with the amount of redaction performed.

These findings align with prior research showing that effective explanations must be tailored to context (Lim, Dey, and Avrahami 2009). Too little explanation can leave users uncertain about why a system behaved as it did, while too much can overwhelm or feel redundant when system behavior is already clear. Our results extend this principle to privacy-preserving systems, demonstrating that the optimal level of explanation detail shifts based on how substantially the system modifies content. This suggests the existence of a "sweet spot" in which explanations provide just enough insight to support user understanding without creating unnecessary cognitive load. More broadly, this pattern is consistent with work on progressive or selective transparency, which proposes that explanation detail should be revealed gradually depending on user needs and the system's level of intervention (Muralidhar, Belloum, and Ashok 2025). In our setting, explanations were particularly valued when redaction substantially changed the content, suggesting that explanation mechanisms for privacy-preserving systems may benefit from adapting based on the amount of modification performed by the system.

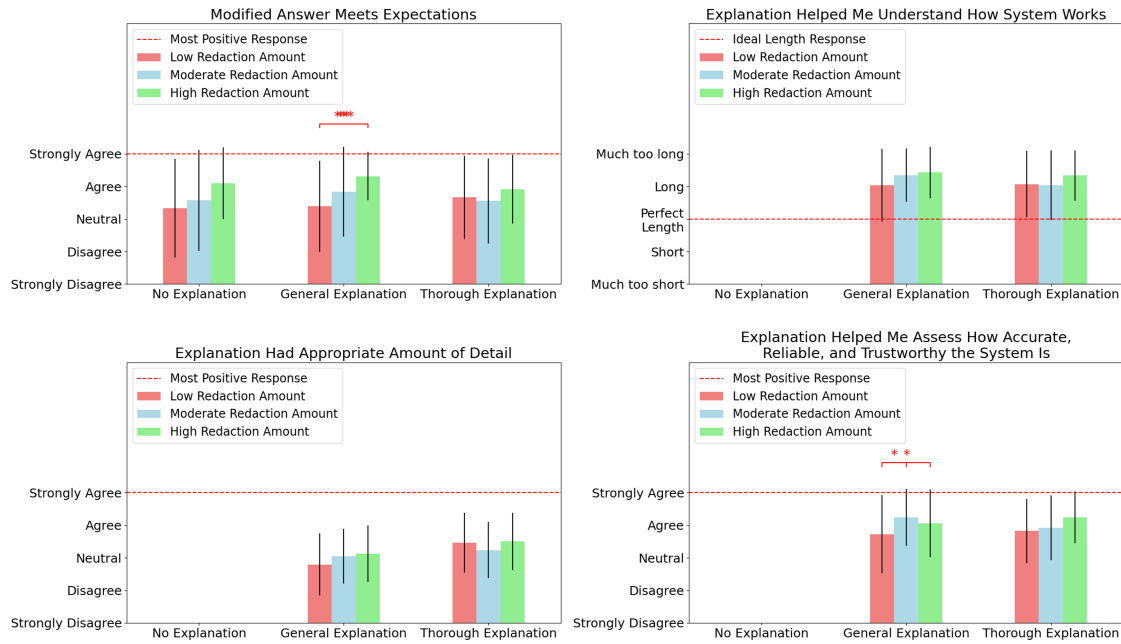


Figure 4: Responses of participants to Survey Section 4.3 about their overall experiences with the system, normalized to 0 – 1. Responses are separated by the explanation received by participants and by the amount of redaction needed for the answer. The value corresponding to the most positive response is shown by a dotted line. Significant differences are shown in red with one, two, and three asterisks for $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

5.3 Effect of Background and Experiences

We finally analyze the results of the final part of the survey (Section 4.3) to determine if there were any correlations between the background/experiences of the participants and their trust in the system. We used an aggregated trust score (averaged across the questions in Section 4.3) and grouped participants by explanation condition (no explanation, general explanation, and thorough explanation). Table 3 shows the F-values, effect sizes, and p-values for each explanation style and for each demographic question.

H3: People will have different levels of trust in the system based on their background and experiences In general, these findings provide partial support for this hypothesis. Some individual difference factors, e.g., baseline trust in technology and AI, were strongly predictive of trust. These effects were consistent across multiple measures and conditions, suggesting that participants enter AI-mediated tasks with predefined dispositions that shape their evaluations.

However, many other background variables did not produce significant effects, and the most significant findings were isolated to the no explanation condition. This pattern suggests that explanations play a stabilizing role on user trust, reducing the degree to which individual differences drive trust. When explanations were provided (general or thorough), trust levels became more uniform across demographic groups. Therefore, while certain personal dispositions matter, demographic characteristics alone (without considering explanations) are not predictors of trust, and

their influence can be greatly minimized by an effective explanation design.

Our findings align with prior research showing that user characteristics, such as familiarity with technology, trust in AI systems, and comfort with automation, shape how individuals respond to AI-driven tools (Ehsan et al. 2024). In our study, trust was consistently higher among participants already predisposed to accept or rely on automation, regardless of the explanation condition. This pattern highlights a dual challenge for designers of privacy-preserving systems:

- How do we maintain and appropriately calibrate the trust of users with a high baseline of confidence in AI without overwhelming them with unnecessary explanation?
- How do we build trust among users with lower baseline confidence, who may need more support or tailored explanations to feel comfortable with the system?

These findings suggest that adaptive or personalized explanation strategies may be essential. Systems that adjust explanation detail to user characteristics and provide accessible and contextually relevant information may better serve the goal of supporting trust across diverse user profiles.

6 Conclusion

As AI systems increasingly act as intermediaries in human communication, designing mechanisms that preserve privacy while maintaining user trust has become critical. This work examined how explanations can support trust in AI-mediated communication, particularly when sensitive infor-

	No Explanation			General Explanation			Thorough Explanation		
	F-value	Cohen's f	p-value	F-value	Cohen's f	p-value	F-value	Cohen's f	p-value
Age Group							$F(5, 52) = 3.12$	0.548	*
Education									
Gender									
Familiar with modern technology	$F(3, 55) = 3.20$	0.416	*						
Trust Technology	$F(4, 54) = 5.08$	0.613	*						
Trust AI/LLM	$F(3, 55) = 3.55$	0.440	**	$F(3, 54) = 3.10$	0.415	*			
Early technology adopter	$F(3, 55) = 5.19$	0.532	**						
Confident in learning new technologies	$F(3, 55) = 4.11$	0.474	*						
Trust output of system that they do not understand	$F(4, 54) = 10.33$	0.875	***	$F(4, 53) = 5.05$	0.618	**			
Importance of understanding how a system works									
Rely on AI system without understanding its reasoning				$F(4, 53) = 4.48$	0.582	**			
Changed decisions based on AI output	$F(4, 54) = 3.49$	0.508	*						

Table 3: Significant Results from the Background and Experience Survey Questions. Significant differences are indicated with one, two, and three asterices for $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

mation is redacted from messages. Our contributions included system development, technical validation, and user evaluation to explore the interplay between redaction, explanations, and individual differences in trust.

We demonstrated that providing explanations significantly improves participants' trust in the system's ability to preserve privacy. Participants who received explanations also engaged more and viewed their experience more positively with the system compared to those who did not receive explanations. Additionally, our results indicate that the effectiveness of explanations is context-dependent. Explanations are most valuable when the system performs extensive redactions, helping users understand the scope of the modifications and the reasoning behind them. When fewer modifications are necessary, the explanations are less critical, suggesting that optimal explanation strategies should adapt to the degree of system intervention.

Individual differences further affect trust in AI-mediated communication. Participants with more familiarity and trust in AI a priori were more likely to trust our system, but these differences were much more pronounced when explanations were not provided. Explanations were a stabilizing influence on user trust; trust levels became more uniform across demographic groups when explanations were provided. These findings highlight the importance of personalized or adaptive explanation strategies, which can tailor the level of detail to user characteristics and maximize trust without overwhelming or under-informing users.

In general, our results provide actionable insights for designing privacy-preserving AI intermediaries. Effective explanations strike a balance between transparency and privacy, offering enough detail to build trust while avoiding unnecessary exposure of sensitive information.

Future work should explore explanations in more diverse settings, including for different domains and more complex privacy protection systems, examine long-term trust evolution, and investigate adaptive explanation strategies that dy-

namically adjust to user characteristics and contextual information. Real-world trust dynamics may differ from the interactions explored in this work, particularly when users depend on redacted information to complete high-stakes tasks. Bridging this gap will require studying AI-mediated communication in more realistic environments, such as task-based settings where users must act on partially redacted information with real data constraints. Incorporating interactive and longitudinal evaluations can further capture how users adapt to redactions and explanations over time. By addressing these challenges, AI mediators can be designed to support privacy-preserving communication using explanations that improve user trust and system understanding.

Alternative approaches could include providing a standardized explanation along with labels indicating the categories of PII removed, or personalizing explanations by incorporating individualized information such as perceived threat severity or worst-case privacy scenarios. While these strategies may increase transparency, our findings suggest that explanation effectiveness strongly depends on contextual factors and individual differences. This highlights the value of developing adaptive explanation approaches that can select or tailor explanation styles based on the user and the context. Such mechanisms could enable AI mediators to support privacy-preserving communication while maintaining user trust and understanding.

6.1 Limitations

The user study focused on a controlled online setting with simulated communication scenarios, which may not fully capture the complexity of real-world interactions. We also collected the background and experience of the participants at the end of the study to avoid these questions affecting the responses of the participants during the study, but their background/experience responses were then affected by their study experience.

Ethics Statement

Our user study protocol was reviewed and approved by an internal ethics committee. All online study participants were compensated for their time, and their Prolific IDs were collected solely for compensation purposes.

References

- Araújo, S.; and Aguiar, M. 2023. Simplifying specialized texts with AI: a ChatGPT-based learning scenario. In *International Conference in Information Technology and Education*, 599–609. Springer.
- Bertolin, R. V.; Santos, A. C. d.; Lima, J. B. d.; and Braga, M. J. 2008. Information asymmetry and trust in cooperative interactions. *Revista de Administração Contemporânea*, 12: 59–81. ISBN: 1415-6555 Publisher: SciELO Brasil.
- Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101. ISBN: 1478-0887.
- Busuioac, M. 2021. Accountable artificial intelligence: Holding algorithms to account. *Public administration review*, 81(5): 825–836. ISBN: 0033-3352.
- Caine, K.; and Hanania, R. 2013. Patients want granular privacy control over health information in electronic medical records. *Journal of the American Medical Informatics Association*, 20(1): 7–15. ISBN: 1067-5027 Publisher: Oxford Academic.
- Carlini, N.; Liu, C.; Erlingsson, U.; Kos, J.; and Song, D. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, 267–284. ISBN 1-939133-06-8.
- Chen, C.; Zhou, D.; Ye, Y.; Li, T. J.-j.; and Yao, Y. 2025. CLEAR: Towards Contextual LLM-Empowered Privacy Policy Analysis and Risk Generation for Large Language Model Applications. ArXiv:2410.13387 [cs].
- Cheng, S.; Meng, S.; Xu, H.; Zhang, H.; Hao, S.; Yue, C.; Ma, W.; Han, M.; Zhang, F.; and Li, Z. 2025. Effective PII Extraction from LLMs through Augmented Few-Shot Learning. 8155–8173. ISBN 978-1-939133-52-6.
- Chromik, M.; and Butz, A. 2021. Human-XAI interaction: a review and design principles for explanation user interfaces. In *IFIP Conference on Human-Computer Interaction*, 619–640. Springer. ISBN 3-030-85616-X.
- de Souza Filho, J. C.; Belloum, R.; and de Oliveira, K. M. 2024. Where are we and where can we go on the road to reliance-aware explainable user interfaces? In *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 282–288. IEEE. ISBN 979-8-3503-6613-6.
- Dosilovic, F. K.; Brcic, M.; and Hlupic, N. 2018. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210–0215. Opatija: IEEE. ISBN 978-953-233-095-3.
- Ehsan, U.; Passi, S.; Liao, Q. V.; Chan, L.; Lee, I.-H.; Muller, M.; and Riedl, M. O. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–32. Honolulu HI USA: ACM. ISBN 979-8-4007-0330-0.
- Ehsan, U.; Tambwekar, P.; Chan, L.; Harrison, B.; and Riedl, M. O. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th international conference on intelligent user interfaces*, 263–274.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; and Bi, X. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Hancock, J. T.; Naaman, M.; and Levy, K. 2020. AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1): 89–100. ISBN: 1083-6101 Publisher: Oxford University Press.
- Hoff, K. A.; and Bashir, M. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3): 407–434. ISBN: 0018-7208 Publisher: Sage Publications Sage CA: Los Angeles, CA.
- Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2019. Metrics for Explainable AI: Challenges and Prospects. ArXiv:1812.04608 [cs].
- Hohenstein, J.; and Jung, M. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior*, 106: 106190. ISBN: 0747-5632 Publisher: Elsevier.
- Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; and Carney, A. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Jakesch, M.; French, M.; Ma, X.; Hancock, J. T.; and Naaman, M. 2019. AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–13.
- Jian, J.-Y.; Bisantz, A. M.; and Drury, C. G. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1): 53–71. ISBN: 1088-6362 Publisher: Taylor & Francis.
- Khadar, M.; Timma, A.; Ludden, L.; and Kaur, H. 2025. Explain it like I’m me: User Characteristics in XAI Decision-Making. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 1729–1744.
- Kim, S. S. Y.; Watkins, E. A.; Russakovsky, O.; Fong, R.; and Monroy-Hernández, A. 2023. “Help Me Help the AI”: Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17. Hamburg Germany: ACM. ISBN 978-1-4503-9421-5.

- Kizilcec, R. F. 2016. How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395. San Jose California USA: ACM. ISBN 978-1-4503-3362-7.
- Kocielnik, R.; Amershi, S.; and Bennett, P. N. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. Glasgow Scotland Uk: ACM. ISBN 978-1-4503-5970-2.
- Krsek, I.; Kabra, A.; Dou, Y.; Naous, T.; Dabbish, L. A.; Ritter, A.; Xu, W.; and Das, S. 2025. Measuring, Modeling, and Helping People Account for Privacy Risks in Online Self-Disclosures with AI. *Proc. ACM Hum.-Comput. Interact.*, 9(2): CSCW131:1–CSCW131:31.
- Krčo, N.; Yao, Z.; Meeus, M.; and de Montjoye, Y.-A. 2026. RAT-Bench: A Comprehensive Benchmark for Text Anonymization. *arXiv preprint arXiv:2602.12806*.
- Larasati, R.; De Liddo, A.; and Motta, E. 2020. The effect of explanation styles on user’s trust. In *Proceedings of IUI workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies*, volume 2582. CEUR Workshop Proceedings.
- Lee, J.; and Moray, N. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10): 1243–1270. ISBN: 0014-0139 Publisher: Taylor & Francis.
- Li, T.; Das, S.; Lee, H.-P.; Wang, D.; Yao, B.; and Zhang, Z. 2024. Human-Centered Privacy Research in the Age of Large Language Models. *ArXiv:2402.01994 [cs]*.
- Lim, B. Y.; Dey, A. K.; and Avrahami, D. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2119–2128. Boston MA USA: ACM. ISBN 978-1-60558-246-7.
- Liu, Y.; Jia, Y.; Jia, J.; and Gong, N. Z. 2024. Evaluating LLM-based Personal Information Extraction and Countermeasures. *arXiv preprint arXiv:2408.07291*.
- Liu, Y.; Mittal, A.; Yang, D.; and Bruckman, A. 2022. Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, 1–13.
- Ma, X.; and Brown, T. W. 2020. AI-Mediated Exchange Theory. *ArXiv:2003.02093 [cs]*.
- Ma, Y.; Peng, H.; and Cambria, E. 2018. Targeted aspect-based sentiment analysis via embedding common-sense knowledge into an attentive LSTM. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Mensah, G. B. 2023. Artificial intelligence and ethics: a comprehensive review of bias mitigation, transparency, and accountability in AI Systems. *Preprint, November*, 10(1): 1.
- Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38. ISBN: 0004-3702 Publisher: Elsevier.
- Miresghallah, N.; Kim, H.; Zhou, X.; Tsvetkov, Y.; Sap, M.; Shokri, R.; and Choi, Y. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*.
- Muralidhar, D.; Belloum, R.; and Ashok, A. 2025. Operationalizing selective transparency using progressive disclosure in artificial intelligence clinical diagnosis systems. *International Journal of Human-Computer Studies*, 103591. ISBN: 1071-5819.
- Nguyen, T. T.; Huynh, T. T.; Ren, Z.; Nguyen, T. T.; Nguyen, P. L.; Yin, H.; and Nguyen, Q. V. H. 2024. A Survey of Privacy-Preserving Model Explanations: Privacy Risks, Attacks, and Countermeasures. *ArXiv:2404.00673 [cs]*.
- Nissenbaum, H. 2004. Privacy as Contextual Integrity. *Washington Law Review*, 79.
- Pal, A.; Bhargava, R.; Hinsz, K.; Esterhuizen, J.; and Bhat-tacharya, S. 2024. The empirical impact of data sanitization on language models. *arXiv preprint arXiv:2411.05978*.
- Park, E.; Deng, W. H.; Varadarajan, V.; Yan, M.; Kim, G.; Sap, M.; and Eslami, M. 2025. Critical or Compliant? The Double-Edged Sword of Reasoning in Chain-of-Thought Explanations.
- Schwartz, P. M.; and Solove, D. J. 2011. The PII problem: Privacy and a new concept of personally identifiable information. *NYUL rev.*, 86: 1814.
- Shahmerdanova, R. 2025. Artificial Intelligence in Translation: Challenges and Opportunities. *Acta Globalis Humanitatis et Linguarum*, 2(1): 62–70. ISBN: 3030-1718.
- Shokri, R.; Strobil, M.; and Zick, Y. 2021. On the Privacy Risks of Model Explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 231–241. Virtual Event USA: ACM. ISBN 978-1-4503-8473-5.
- Song, Q.; Wu, Y.; Hernandez, R. H.; Li, Y.; Kou, Y.; and Gui, X. 2025. Understanding Users’ Perception of Personally Identifiable Information. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–24.
- Sousa, S.; and Kern, R. 2023. How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing. *Artificial Intelligence Review*, 56(2): 1427–1492. ISBN: 0269-2821 Publisher: Springer.
- Sun, B.; Zhou, Y.; and Jiang, H. 2024. Empowering users in digital privacy management through interactive llm-based agents. *arXiv preprint arXiv:2410.11906*.
- Ueno, T.; Sawa, Y.; Kim, Y.; Urakami, J.; Oura, H.; and Seaborn, K. 2022. Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–7. New Orleans LA USA: ACM. ISBN 978-1-4503-9156-6.
- Vereschak, O.; Alizadeh, F.; Bailly, G.; and Caramiaux, B. 2024. Trust in AI-assisted Decision Making: Perspectives from Those Behind the System and Those for Whom the Decision is Made. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–14. Honolulu HI USA: ACM. ISBN 979-8-4007-0330-0.

Vorm, E. S.; and Combs, D. J. Y. 2022. Integrating Transparency, Trust, and Acceptance: The Intelligent Systems Technology Acceptance Model (ISTAM). *International Journal of Human-Computer Interaction*, 38(18-20): 1828–1845.