MULTIDISCIPLINARY RESEARCH INSTITUTE INDIA

Multidisciplinary Journal of Research in Engineering and Technology

ISSN: 2348-6953

Volume 13 Issue 02 (2026) | Pages: 358-364

The Rise of AI Chatbot Leaks: How Users Unknowingly Share Sensitive Information

Ravikant Zirmite [1], Sharvi Belsare[2], Prajakta Joshi[3], Shreya Joshi[4], Tejas Bhos[5], Siddhi Kawade[6]

[1,2,3,4,5,6]Department of MCA, MES' IMCC, Pune

[1]rsz.imcc@mespune.in, [2]sharvibelsare@gmail.com, [3]prajaktajoshi2003@gmail.com, [4]joshishreya2019@gmail.com, [5]tejasbhos448@gmail.com, [6]siddhikawade92@gmail.com

Peer Review Information

***Type**: Article*
***Received**: 20 March 2026*
***Revised**: 03 April 2026*
***Accepted**: 21 May 2026*
***Published**: 03 June 2026*

Abstract

While such AI chatbots like ChatGPT, Gemini, and Claude are increasingly becoming part of regular workflows and con- versations, privacy and data security concerns are amassing. Though the tools are immensely valuable to drive productivity, automation, and customer interactions, unknowing users are sharing close personal or business information during use. This study probes the prevalence of accidental data spillage by AI chatbots using the integration of a standardized survey of users alongside case studies predicated on real events. An 18–30 years' old user born in the online world is the target user sought to track the usage behavior, type of information exchanged, interpretation of responses given by chatbots, and sensibility towards potential privacy threats. Moreover, well- known data breach case studies are analyzed to outline the manner in which misuse or misconfiguration of AI tools leads to gross privacy breaches. Implications of primary and secondary data indicate an acute lag in awareness among users, emphasizing the demand for more protection, policy measures, and AI tool digital literacy. The paper concludes by propounding practical suggestions to mitigate threats while unleashing the potential of generative AI.

Keywords: AI Chatbots; Data Privacy; ChatGPT; Sensitive Information; User Behavior; Cybersecurity; Trust in Artificial Intelligence; Data Leaks; Prompt Injection; Generative Artificial Intelligence Risks.

How to Cite This Article

Zirmite, R., Belsare, S., Joshi, P., Joshi, S., Bhos, T., & Kawade, S. (2026). The rise of AI chatbot leaks: How users unknowingly share sensitive information. *Multidisciplinary Journal of Research in Engineering and Technology*, 13(2), 358–364.

Introduction

The sudden arrival of AI chatbots like ChatGPT, Google Bard, and Microsoft Copilot has transformed the business model and day-to-day routine of an individual, making it easier and more convenient than ever. The technologies are being increasingly used for everything from automating customer service to office process automation, content creation, and technical problem-solving assistance.

However, this innovation is also generating its own prob- lems—primarily data security and protection. Both customers and employees are unknowingly disclosing confidential in- formation while making transactions through these chatbots, without knowing anything about the repercussions. With the years that these AI technologies have been serving day-to- day needs, the leakage of unwanted information has multi- plied many times. The unintended disclosure of confidential, personal, or proprietary information has become an issue in most industries despite AI chatbots being highly supportive.

Security Risks and Threat Landscape

For example, employees writing cryptic messages using AI or developers asking code-level questions can unwittingly ex- pose sensitive information like source code, individual identi- fiers, or business strategies. It is not just a data privacy problem but a huge security risk. As such, breaches of confidence and exposure of businesses to cyber attacks are inevitable.

Adding to this problem is the fact that threat actors now em- ploy AI-generated responses for the purpose of executing more advanced phishing and social engineering attacks. These AI- generated attacks are more difficult to detect and are generally more convincing compared to their traditional counterparts, thus posing a serious threat in the security space. The problem in preventing these risks is the adaptive nature of generative AI technologies, which are likely to be beyond conventional security paradigms.

Growing Complexity in Data Protection

With the increasing deployment of AI chatbots such as ChatGPT in consumer and workplace spheres, the degree of sophistication required in data privacy and security needs has become more complex. With chatbots now capable of process- ing huge amounts of user data and responding instantly, they have inadvertently become de facto conduits for information leaks.

High-profile incidents, such as the inadvertent leakage of source code to AI chatbots by workers in some instances, un- veiled enormous gaps in user sensitivity and security protocols. These are just some of the more representative examples of the larger issue at hand: how to utilize effective data security protocols without taking away from the merits of AI.

Need for Awareness and Security Protocols

Even though there is growing use of AI chatbots, their use in unexpected areas is causing enormous breaches of data, most of which are caused due to user lack of awareness and lack of appropriate security protocols. Although used by more companies and users, with greater space to accommodate sensitive information to leak or be exploited, it is necessary to conduct studies as to why such data breaches occur and how improved data security frameworks can be created in the age of generative AI.

Importance of the Paper

This research is timely in the era of expanding uses of AI- based solutions in various industries. Learning from research on actual data breaches and phishing attacks on AI chatbots, this study will attract attention to the necessity of greater user awareness and enhanced security measures. Conquering these challenges will provide secure and confidential data handling and facilitate the secure and responsible use of AI technology.

Research Questions

- What are the most prevalent forms of sensitive informa- tion that users unwittingly provide to AI chatbots?
- How do AI chatbots play a role in the unintentional leakage of confidential data in business and personal environments?
- What are the main causes of data leaks through AI chatbots, and how can they be addressed?
- How are attackers using AI-produced content in social engineering campaigns and phishing attacks?
- What can be done by companies to reduce risks of data leaks and maximize security when using AI chatbots?

By answering the above questions, this research intends to provide effective insights into avoiding accidental disclosure of confidential information and maximizing security while using AI chatbots.

Literature Review

The adoption of AI chatbots in business workflows has been the focus of growing industry-oriented and academic research. They have been studied for adoption in different industries like healthcare, customer service, education, and software development, where they have shown

promise to organize work processes more efficiently and increase user interaction (Chen et al., 2021; Vryoni & B., 2021). The study focuses on how AI applications, such as chatbots, play a crucial role in workplace productivity and automation, especially in administrative and technical work. One common theme in scholarly literature is the extent to which users trust AI systems. Min et al. (2021) conducted a study that found users commonly rate AI chatbots as intelligent and trustworthy conversational partners. Such perceived intelligence may result in over-trust, as it may prompt users to reveal sensitive or private information without satisfactorily assessing the asso- ciated risks. Zhang et al. (2022) also complement this by illustrating how higher levels of trust in AI are directly related to higher disclosure rates within AI-enabled customer service settings. The "automation bias" phenomenon also aggravates this tendency, where users place excessive reliance on AI suggestions or features without critically evaluating them (Kim & Lee, 2023). On the cybersecurity side, a number of industry reports signal impending threats related to generative AI. The IBM X-Force Threat Intelligence Index (2024) points to AI- powered phishing and social engineering as emerging attack vectors, where cybercriminals employ AI-generated text to create human-like conversations for scamming. Likewise, the ENISA Threat Landscape Report presents increasing concerns regarding the abuse of AI in creating manipulative or deceptive content, such as impersonating customer support representa- tives or executives. Real-world examples provide a chilling look at the risks. One of the best-known incidents was when Samsung workers, in January 2023, employed ChatGPT to debug code—unwittingly sending proprietary and confidential information to an outside system (Vincent, 2023). This led the company to ban internal use of AI tools and triggered a corporate reevaluation about AI regulation. Other businesses have released internal policy or memos warning staff from disclosing work information to public AI tools.

In spite of the increasing number of studies, there is still a big gap in the current literature. The majority of current re- search centers on AI technical potential or enterprise adoption, with few studies dedicated to real user actions that result in unintentional data leaks. Empirical studies are needed that test the questions of how and why users provide sensitive data to AI solutions and how situational factors such as trust, urgency, or unawareness affect these choices (West et al., 2019). This paper fills that gap by examining actual incidents, probing user behavior patterns, and providing practical recommendations to reduce data leakage threats in the context of AI chatbot use.

Research Methodology

Purpose of Research The overall purpose of this research is to determine how often users inadvertently give away sensitive information to AI chatbots and to measure their level of awareness about privacy and security threats. With the increasing usage of conversational AI tools, it becomes more important to know how users behave and might be vulnerable. Research Design This research adopts a quantitative, cross-sectional survey design to examine user interactions with AI chatbots and the privacy threats that come with them. Data was elicited via a structured Google Form with solely close-ended questions to facilitate ease of completion and efficient analysis. Participant Demographic The age group targeted were people between the ages of 18–30 years, and they were chosen because they make extensive use of digital tools within both academic and professional settings. This age category is more susceptible to using AI chatbots for various purposes—ranging from expressing emotions to professional communication—and thus potentially exposing themselves to increased risks of oversharing sensitive data without even noticing it. Survey Structure A cross-sectional, quantitative survey was employed within this study to determine the extent to which users unknowingly share sensitive data with AI chatbots and their level of awareness regarding potential privacy risks. The majority of data were collected through a structured web-based survey designed to offer quantifiable information on users' behavior, trust, and attitudes towards conversational agents powered by AI. Structure of the survey form The survey was designed to thoroughly study different facets of user interaction with AI chatbots. It was divided into four broad categories: The first part, AI Usage Patterns, was targeted at determining the most commonly used chatbot platforms, such as ChatGPT, Gemini, Claude, Copilot, and DeepSeek. The respondents also had to determine the main usage scenarios of these platforms, such as coding support, content creation, emotional support, or research purposes. This part was targeted at establishing a point of initiation of how AI chatbots are being incorporated into users' daily routines. The second subject, Nature of Shared Data, investigated the type of information users tend to share when interacting. Forms ranged from common information-seeking questions to more personal inputs like individual thoughts, professional content, proprietary code, and even sensitive information like passwords or financial data. This part of the survey aimed to uncover potential oversharing behavior and unintentional data leaks. The third, Trust and Behavior, was designed to measure the degree to which users trusted chatbot responses, felt emotionally connected to the tools, and how much oversharing happened based on perceived human-like intelligence of AI technology. Participants were asked to report how believable or helpful chatbot responses appeared, and whether realism in the conversation influenced the willingness to share information. The final section, Privacy Awareness and Risk Perception, asked how familiar participants were with chatbot data policies and how well they knew about privacy consequences. This section of the survey also inquired as to whether respondents ever noticed suspicious activity—i.e., suggestive recommenda- tions or uncharacteristically correct answers—that might be an indication of data exploitation or algorithmic profiling. Data Collection Methods Primary data for the study was collected using a designed Google Form survey, shared through social media platforms and scholarly networks to access the targeted audience. The survey was comprised of all close-ended ques- tions to allow for easy analysis and ensure the ease of response completion. In addition to primary data, the study relied on secondary data sources, including high-profile case studies and cybersecurity reports. These sources included media coverage of incidents like the Samsung data breach involving ChatGPT, industry reports from IBM X-Force and the European Union Agency for

Cybersecurity (ENISA), and academic literature on AI-driven phishing and data mishandling. Sampling and Tools A convenience sampling approach was employed, targeting individuals aged 18 to 30. This demographic was selected due to their frequent use of AI tools in both academic and professional environments. The survey link was circulated among peer groups, institutional mailing lists, and relevant online communities. Data collection was facilitated through Google Forms, while Google Sheets and Microsoft Excel were used for data organization and analysis. The final sample size was 45 responses. Ethical Considerations To guarantee ethical research compliance, some precautions were observed. The survey was anonymous, and no personally identifiable information was gathered. A disclaimer on the front of the form made clear that responses were being gathered exclu- sively for an academic research study as part of an MCA curriculum. Individuals were told their data would be kept confidential and the survey would only take a few minutes to complete. Limitations The study, despite the applicability of the results, has some limitations. It may overlook some viewpoints from age groups older than 18 to 30 because they might be using AI chatbots in ways that are not comparable. Second, self-reporting data is subject to the potential biases of underreporting dangerous behavior or exaggerating familiarity with privacy hazards. The convenience sampling approach also restricts the generalizability of the findings, as the sample would not reflect the larger population.

Findings and Analysis

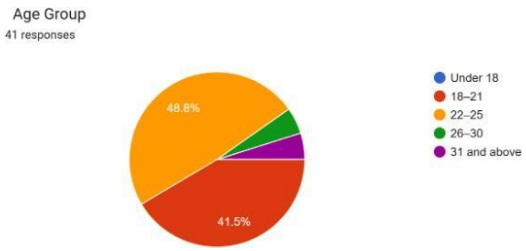

Fig. 1. *Age Group*

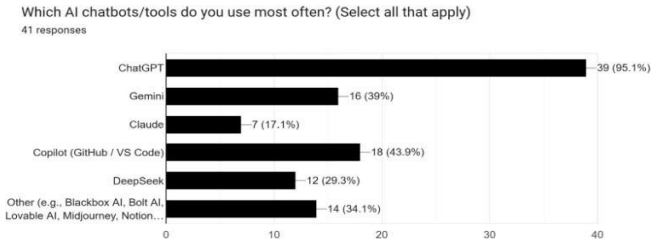

Fig. 2. *Most used AI Chatbots*

The findings of this study reveal interesting trends in the way young digital natives interact with AI chatbots, both their benefits and inherent privacy risks. The majority of the participants belong to the 18–30 age bracket—a generation heavily steeped in digital worlds and inclined to use AI tools for many tasks. This generation which is AI-ready engages daily with AI-powered chatbots, the majority of which use software like ChatGPT, Google Gemini, Microsoft Copilot, and Claude, most particularly to create content, coding help, and productivity aid. But this repeated use is accompanied by an undertone of threat. An overwhelming majority of participants admitted they had plagiarized work-related or sensitive information into chatbot inputs—like codebases and internal reports to client-based data. What spurs this is less malintent and rather the convenience and efficiency the tools bring about when breaking difficult projects.

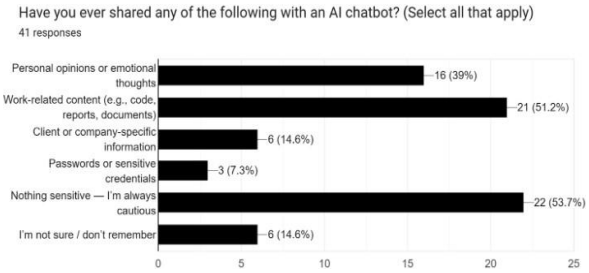

Fig. 3. *Info. shared with chatbots*

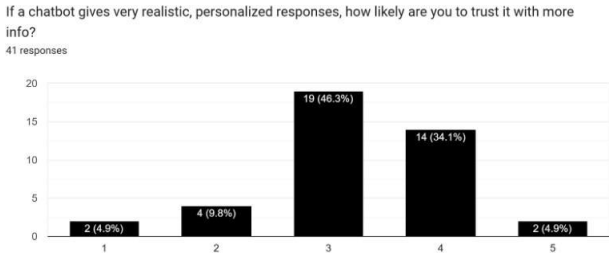

Fig. 4. *Trust with info.*

Most interestingly, perceived intelligence and realism are most highly related to trusting chatbots. The individuals who said they were provided with very contextual or affectively sensitive responses were most likely to give more information, and tend to do so without perceiving the potential threat. This agrees with psychological studies on anthropomorphism in technology—once consumers begin treating machines as dialogue partners or assistants, they are more at ease.

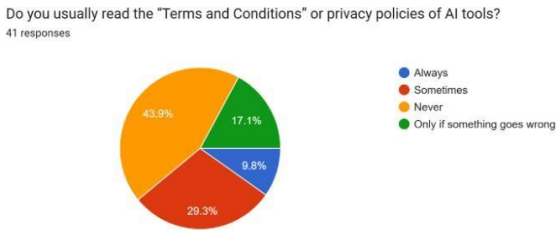

Fig. 5. *Terms and conditions*

The survey also suggests a disquieting disconnect between awareness of privacy. The majority of respondents indicated a high degree of privacy when using chatbots, although the majority of AI websites monitor and even analyze user input to improve their models. And the majority indicated that they never or seldom read terms of service and privacy policies of AI sites, which suggests a broader cultural pattern of dismissing online permission and data rights.

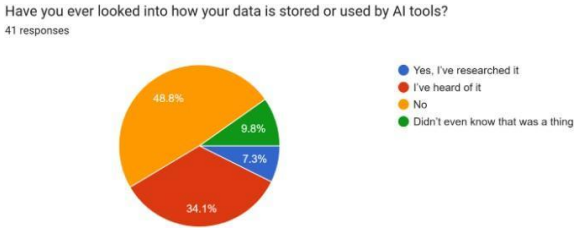

Fig. 6. *Data Storage*

These behav- ioral tendencies are an echo of fresh high-profile real-world situations, such as the Samsung scandal, where employees exposed confidential source code through ChatGPT, believing the discussion to be secure. These cases depict the harsh realities of relying too heavily on AI tools without enough exposure to how information is stored, accessed, and perhaps utilized against them.

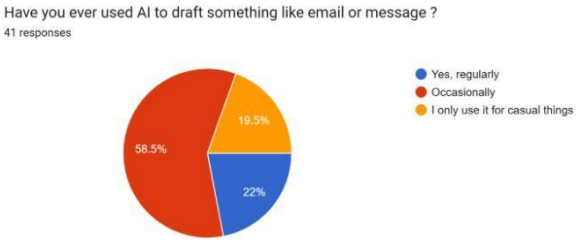

Fig. 7. *AI usage for email writing*

In short, although AI chatbots offer tremendous value in process automation and making information universally ac- cessible, they also introduce with them a subtle but potent vector of data leakage. The study finds that oversharing is less a matter of thoughtlessness and more a consequence of the emotional and contextual realism that these systems elicit. Since computerized artificial intelligence increasingly encroaches on quotidian digital action, user training, design- sourced transparency, and greater user control need to be taken along in the hope of safeguarding sensitive data. Such research would be replicated at larger population bases or through addition of behavior tracking in conjunction with self-report as a means to validate oversharing in real-time.

Case Studies

1. Samsung's ChatGPT Leak (2023)-Sources: BBC, The Verge In 2023, Samsung employees shared confidential infor- mation while using ChatGPT. They were attempting to trou- bleshoot some code, so they copied and pasted confidential in- formation including semiconductor code and internal meeting notes into ChatGPT. The employees shared this information not realizing ChatGPT is a generative artificial intelligence application that stores and learns from the information it is provided creating the danger of private company secrets being leaked or misused. Following this incident, Samsung banned the use of ChatGPT for work-related tasks and created their own private AI tool with better security. After this events many companies realized that they need to train employees on the proper use of AI and implement strict policies to protect sensitive information. Following the Samsung incident, major companies like Apple and JPMorgan banned the use of ChatGPT.

2. WotNot AI Chatbot Data Breach (2024)-Sources: TechCrunch, Cybernews In 2024, WotNot, a chatbot platform used by various companies for customer inquiry-handling, suffered a deep data breach due to misconfigured Google Cloud storage and accidentally exposing over 346,000 files to public access. The files had sensitive personal information like resumes, passports, and medical records mostly from

users of WotNot's free version. This is particularly worrisome regarding security with inexpensive AI services. In addition, because of slide regulations, many businesses face little or no accountability. This incident was a reminder to organizations to exercise caution when purchasing AI services and regularly evaluate security. One of the biggest takeaways from this lap- top was the hope of operationalizing a "zero-trust" approach, where no one is automatically trusted, and audit everything.

3. Deepfake CEO Scam at WPP (2024)-Sources: Financial Times, WPP Statement In 2024, hackers attempted to deceive a WPP employee utilizing ai to produce a deepfake voice of the CEO in a video call. The attackers impersonated the CEO using AI voice cloning in order to convince the employee to share confidential company correspondence with the scammers. Fortunately, the deception was caught before any negative consequence could occur. In similar incidents however, scammers have successfully stolen $25 million, such as in Hong Kong, utilizing deepfake AI videos. These attacks are successful in part because humans generally trust the sounds or voices they hear and see on video. For this reason, traditional security measures such as voice recognition fail when used alone as a protective measure. Companies now should use stronger protections such as multi-factor authen- tication or complementary technologies that include scanners that detect fake videos or voices. This example demonstrates that as the standard of AI generated content improves, we must approach the use of AI with care.

4. AI-Generated Phishing Emails-Sources: IBM X-Force 2023, Proofpoint Cybercriminals are moving into the level of artificial intelligence (AI) tools (ChatGPT and WormGPT) to write fake emails that look very real. These phishing emails attempt to trick people into these emails containing malicious links or giving out personal info. Although writing a phishing email was never difficult, if the writer was able to imitate the writing of an organization or person, it's no surprise that AI can copy a style enough to duplicate the personality of someone an employee knows, trust, or potentially respects. Recent studies show individuals that received emails written by AI are 30 percent more likely to click on a link or provide sensitive information versus an email written by a novice human being. AI is making it easy, as even beginners (utilizing a chat tool) can create dear organizations worth of emails critically representing the organization. In response to the threat, organizations are using AI-based email filters to identify potentially forged or altered messages. Some organizations are also implementing test phishing drill to practice spotting suspicious emails. Overall, we see how humans must stay vigilant, continually educate, and leverage technology.

5. AI Voice Cloning Bank Fraud ($35 Million Loss) - Source: Wall Street Journal In one of the largest AI-related scams to date, cybercriminals impersonated the voice of an ex- ecutive to commit $35 million worth of fraud. The fraudulent firm manufactured fake voices of executives at the company and used phone calls to instruct company employees to transfer money. Because the voices sounded real, the employees fell for their deception. This demonstrated that even systems that use voice recognition for security checks can be fooled. Companies now need better technology that can discern fake voices, and to be on the lookout for voices that sound flawless or unnatural, and they have to apply multiple checks on large transactions. This situation highlights the dangers that AI is capable of when used incorrectly with the added bonus of the ability to alleviate the perpetrators of accountability with voice cloning and other AI technologies, which is why even more vigilance will be required in security online.

Conclusion and Future Recommendations

As AI chatbots such as ChatGPT, Gemini, and Claude become integral to both personal and professional This study draws attention to a growing but little-known problem as AI chatbots like ChatGPT, Gemini, and Claude become more and more integrated into daily life: the unintentional sharing of private data. The results of the secondary case analysis and structured survey show that many users, especially digital natives between the ages of 18 and 30, regularly use these tools without fully comprehending the privacy implications. Oversharing behaviors are influenced by the trust that is placed in chatbot responses, which is frequently fueled by their perceived intelligence and human-like interactions.

The research also uncovers a disturbing lack of user and organizational protection. In spite of high-profile events and alerts issued by cybersecurity bodies, numerous users are still uninformed about the ways in which their data might be stored, utilized, or even manipulated via generative AI tools. What this suggests is that there's an immediate requirement for proactive responses at every level—individual, institutional, and policy-based.

1. Programs for Digital Literacy and User Education:
   - Launch awareness campaigns about data privacy in AI interactions.
   - Incorporate lessons that describe the ramifications of exchanging data with AI systems into digital literacy curricula.
2. Governance and Policy:
   - Companies should establish explicit guidelines limiting the use of open AI tools for private or work-related purposes.
   - Standardized rules for data handling and disclosure in AI systems must be established by regulatory bodies..
3. Integrated Privacy Features:

- AI chatbot developers ought to incorporate opt-in consent procedures, data anonymization, and privacy reminders into the user interface.
- When potentially sensitive data is found in user input, real-time flags or alerts could notify users.

4. Additional Research:
   - To obtain broader insights, future studies could examine the behavior of older populations or professionals in regulated industries (such as healthcare, law, or finance).
   - Longitudinal studies could evaluate how user behavior changes over time, particularly following awareness campaigns.
5. . Cooperation Among Stakeholders:
   - To create morally sound and open AI systems, promote cooperation among tech companies, academic institutions, and cybersecurity companies.
   - Encourage the creation of open-source tools that provide safe substitutes for for-profit AI chatbots.

In conclusion, even though generative AI offers incredible potential, its responsible application necessitates teamwork. We can better utilize AI's advantages while lowering its risks by raising awareness, bolstering governance, and putting user safety first.

References

1. Vryoni, V., & Bouvin, B. (2021). Chatbots in healthcare: Towards AI-enabled general diagnosis and medical support. University of Piraeus Repository. https://doi.org/10.26267/UNIPI_DIONE/1065
2. Min, F., Fang, Z., He, Y., & Xuan, J. (2021, January 15). Research on users' trust of chatbots driven by AI: An empirical analysis based on system factors and user characteristics. International Conference on Consumer Electronics. https://doi.org/10.1109/ICCECE51280.2021.9342098
3. Chen, Y., Prentice, C., Weaven, S. K. W., & Hsiao, A. (2021). A systematic literature review of AI in the sharing economy. Journal of Global Scholars of Marketing Science. https://doi.org/10.1080/21639159.2020.1808850
4. IBM X-Force. (2024). Threat Intelligence Index 2024. IBM Security. https://www.ibm.com/reports/threat-intelligence
5. Zhang, K., Zhao, K., Chen, L., & Chen, Y. (2022). Trusting artificial intelligence: The impact of user trust on disclosure behavior in AI-assisted customer service. *Computers in Human Behavior, 134*, 107328. https://doi.org/10.1016/j.chb.2022.107328
6. Kim, S., & Lee, J. (2023). The role of automation bias in human–AI collaboration: Implications for decision-making and privacy. *Journal of Human–Computer Interaction, 39*(1), 85–102. https://doi.org/10.1080/10447318.2022.2112657
7. Vincent, J. (2023, April 3). Samsung bans use of generative AI tools like ChatGPT after internal data leak. *The Verge*. https://www.theverge.com/2023/4/3/23667355/samsung-chatgpt-ban-privacy-leak
8. West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race, and power in AI. AI Now Institute. https://ainowinstitute.org/discriminatingsystems.pdf