Between Protection and Control: AI-Based Privacy Obfuscation and Its Implications for User Agency

MARVIN STRAUSS, Human-Computer Interaction Group, University of Duisburg-Essen, Germany
STEFAN SCHNEEGASS, Human-Computer Interaction Group, University of Duisburg-Essen, Germany

1 MOTIVATION

Ubiquitous sensing technologies have become an integral part of contemporary digital systems, ranging from smartphones and social media platforms to Extended Reality (XR) environments. These systems continuously collect, process, and disseminate personal data about users and bystanders alike. As a result, questions of consent and privacy protection have become increasingly complex. Yet, many current systems still rely on binary consent mechanisms, offering users only coarse-grained choices such as allowing or denying data collection altogether [2].

Such binary models fail to capture the nuanced and situated nature of privacy preferences in continuously sensing environments. Consent is rarely a one-time decision. Rather, it is an ongoing negotiation that depends on context, audience, purpose, and anticipated consequences. In response to this limitation, data obfuscation has emerged as a promising alternative. Instead of preventing data use entirely, obfuscation selectively hides or alters sensitive information while preserving the utility of the remaining data [9].

Traditional obfuscation techniques typically introduce visible degradation, for example, through blurring or noise. While these approaches reduce data fidelity, they also make privacy protection perceptible to users, thereby maintaining a degree of transparency and control [13]. Recent advances in AI fundamentally change this dynamic. AI-based obfuscation can preserve realism, coherence, and aesthetic appeal, rendering privacy protection effectively invisible. Moreover, such systems can autonomously decide which attributes to obfuscate, how to modify them, and what information remains unchanged [4, 14, 16].

This shift transforms obfuscation from an explicit, user-visible intervention into an implicit, automated system behavior. While this promises improved usability and data utility, it simultaneously raises fundamental questions about informed consent, accountability, and user agency.

2 AI-BASED OBFUSCATION

Instead of offering binary choices of full access or complete denial, research proposed using a more fine-grained mechanism for user consent [9]. Traditional obfuscation techniques, such as pixelation in image data, can effectively conceal sensitive information. However, they often degrade perceptual quality and reduce the overall appeal and utility of the output [13]. Recent advances in AI-based methods offer promising alternatives that preserve the aesthetic while

Authors' addresses: Marvin Strauss, marvin.strauss@uni-due.de, Human-Computer Interaction Group, University of Duisburg-Essen, Essen, Germany; Stefan Schneegass, stefan.schneegass@uni-due.de, Human-Computer Interaction Group, University of Duisburg-Essen, Essen, Germany.

Manuscript submitted to ACM

protecting privacy. In the context of image obfuscation, generative approaches enable, for instance, the replacement of an identifiable face with a synthetic identity [15], or the substitution of sensitive regions such as captured documents or on-screen content with alternatives [16]. Beyond transformation, AI techniques can also support the automatic detection and selection of privacy-relevant content that requires obfuscation [14].

Recent research on human perception of such obfuscation techniques reports encouraging findings, demonstrating their potential to enhance privacy protection while preserving aesthetic quality, and rendering the manipulation imperceptible to observers [4, 5, 16]. Khamis et al. [4] investigated the perceived effectiveness of deepfakes for face obfuscation in an online study. Participants viewed images of obfuscated public figures. The results indicated that deepfakes were as effective as conventional methods in protecting privacy, while being more aesthetically pleasing. Similarly, Xu et al. [16] explored generative content replacement in images to conceal privacy-sensitive content, showing that deepfakes hindered identification while maintaining narrative coherence and visual harmony. Another study by Khamis et al. [5] compared the perspectives of both image owners and the individuals whose faces were obfuscated. Participants applied different obfuscation methods to themselves and others in their personal photos. Findings revealed that both groups regarded face-swapping as more effective than traditional methods, highlighting its superior integration within the visual context.

3 ETHICS, CONTROL, AND USER AGENCY

The ethical challenges of AI-based obfuscation closely parallel those discussed in the context of generative AI and deepfakes. In both cases, personal data is synthetically modified in ways that may be difficult to detect, understand, or contest. These challenges transfer directly to privacy protection systems that rely on AI-driven obfuscation [6].

A central concern lies in the opacity of AI decision-making. Users are often unable to understand why specific attributes are considered sensitive, how they are altered, or what synthetic information replaces them. This lack of intelligibility directly affects user agency: when users cannot meaningfully anticipate or influence system behavior, their capacity for intentional and informed action is diminished.

Ethical implications extend beyond the individual whose data is being protected. Recipients of obfuscated data, including other users, may be unaware that information has been modified. As a result, synthetic attributes may be interpreted as truthful representations, shaping judgments, interactions, or decisions based on altered realities. In such cases, AI-based obfuscation not only protects privacy but also actively constructs new representations of individuals [7].

At the same time, AI-based obfuscation is often justified by its ability to reduce users' cognitive burden. By delegating complex privacy decisions to automated systems, users are relieved from continuously selecting which information to protect. However, this convenience entails a redistribution of control from users to AI systems. Users may lose the ability to inspect, override, or contest how they are represented. In extreme cases, obfuscation may fail to protect information that users consider sensitive, introduce attributes they find undesirable, or portray them in ways that negatively affect their identity, reputation, or social standing [1].

Overall, these issues suggest that treating AI-based obfuscation as a purely technical privacy solution is insufficient. Instead, it should be understood as a socio-technical intervention that reshapes control, responsibility, and agency.

4 AUTHOR PERSPECTIVE

We approach these questions from research on privacy in XR systems, where sensing is continuous, embodied, and socially embedded [3, 10]. XR environments exemplify the challenges of contemporary privacy protection: they capture intimate behavioral, physiological, and contextual data and frequently affect not only primary users but also

bystanders [8]. Our work aims to raise awareness among XR stakeholders about these privacy implications, design usable control interfaces, and explore privacy-preserving mechanisms such as selective disclosure and obfuscation [11, 12]. While XR serves as our primary domain, the challenges discussed in this paper extend to adjacent contexts such as smartphones and social media, where AI-based obfuscation is increasingly plausible.

5 CONCLUSION

Advanced privacy protection mechanisms are essential for systems characterized by pervasive data collection. AI-based obfuscation offers a compelling alternative to binary consent by enabling selective protection while maintaining data usability and experiential quality. However, these benefits come with ethical costs, particularly regarding transparency, control, and user agency.

Designing privacy protection without simultaneously designing for agency risks reinforces existing power asymmetries between users and automated systems. We therefore argue for agency-aware approaches to AI-based obfuscation, including human-AI hybrid models, intelligible obfuscation strategies, and interaction mechanisms that allow users to inspect, influence, and contest AI decisions.

Ultimately, effective privacy protection should not only shield users from harm but also preserve their ability to understand, influence, and take ownership of how they are represented through data.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[2] Zheran Fang, Weili Han, and Yingjiu Li. 2014. Permission based Android security: Issues and countermeasures. *computers & security* 43 (2014), 205–218.

[3] Gonzalo Munilla Garrido, Vivek Nair, and Dawn Song. 2023. Sok: Data privacy in virtual reality. *arXiv preprint arXiv:2301.05940* (2023).

[4] Mohamed Khamis, Habiba Farzand, Marija Mumm, and Karola Marky. 2022. DeepFakes for privacy: Investigating the effectiveness of state-of-the-art privacy-enhancing face obfuscation methods. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces*. 1–5.

[5] Mohamed Khamis, Rebecca Panskus, Habiba Farzand, Marija Mumm, Shaun Macdonald, and Karola Marky. 2024. Perspectives on DeepFakes for Privacy: Comparing Perceptions of Photo Owners and Obfuscated Individuals towards DeepFake Versus Traditional Privacy-Enhancing Obfuscation. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia*. 300–312.

[6] Jan Kietzmann, Linda W Lee, Ian P McCarthy, and Tim C Kietzmann. 2020. Deepfakes: Trick or treat? *Business Horizons* 63, 2 (2020), 135–146.

[7] Stefan Larsson and Fredrik Heintz. 2020. Transparency in artificial intelligence. *Internet policy review* 9, 2 (2020), 1–16.

[8] Vivek Nair, Gonzalo Munilla Garrido, Dawn Song, and James F O'Brien. 2022. Exploring the privacy risks of adversarial VR game design. *arXiv preprint arXiv:2207.13176* (2022).

[9] Katarzyna Olejnik, Italo Dacosta, Joana Soares Machado, Kévin Huguenin, Mohammad Emtiyaz Khan, and Jean-Pierre Hubaux. 2017. Smarper: Context-aware and automatic runtime-permissions for mobile devices. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1058–1076.

[10] Viktorija Paneva, Marvin Strauss, Verena Winterhalter, Stefan Schneegass, and Florian Alt. 2024. Privacy in the Metaverse. *IEEE Pervasive Computing* 23, 3 (2024), 73–78.

[11] Viktorija Paneva, Verena Winterhalter, Naga Sai Surya Vamsy Malladi, Marvin Strauss, Stefan Schneegass, and Florian Alt. 2025. Usable Privacy in Virtual Worlds: Design Implications for Data Collection Awareness and Control Interfaces in Virtual Reality. *arXiv preprint arXiv:2503.10915* (2025).

[12] Marvin Strauss, Viktorija Paneva, Florian Alt, and Stefan Schneegass. 2024. Designing and Evaluating Scalable Privacy Awareness and Control User Interfaces for Mixed Reality. *arXiv preprint arXiv:2409.00739* (2024).

[13] Yasuhiro Tanaka, Akihisa Kodate, Yu Ichifuji, and Noboru Sonehara. 2015. Relationship between willingness to share photos and preferred level of photo blurring for privacy protection. In *Proceedings of the ASE BigData & SocialInformatics 2015*. 1–5.

[14] Terence E Taylor, Frank Keane, and Yaniv Zigel. 2021. A speech obfuscation system to preserve data privacy in 24-hour ambulatory cough monitoring. *IEEE Journal of Selected Topics in Signal Processing* 16, 2 (2021), 188–196.

[15] Tomasz Walczyna and Zbigniew Piotrowski. 2023. Quick Overview of Face Swap Deep Fakes. *Applied Sciences* 13, 11 (2023), 6711.

[16] Anran Xu, Shitao Fang, Huan Yang, Simo Hosio, and Koji Yatani. 2024. Examining Human Perception of Generative Content Replacement in Image Privacy Protection. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.