The AI Accomplice: Exploring Generative Artificial Intelligence in Facilitating and Amplifying Deceptive Designs

Thomas Kosch
HU Berlin
Berlin, Germany
thomas.kosch@hu-berlin.de

Veronika Krauß
Responsible System Design
Ansbach University of Applied Sciences
Ansbach, Germany
veronika.krauss@hs-ansbach.de

Christopher Katins
HU Berlin
Berlin, Germany
christopher.katins@hu-berlin.de

Dominik Schön
Technical University of Darmstadt
Darmstadt, Germany
schoen@tk.tu-darmstadt.de

Mark McGill
School of Computing Science
University of Glasgow
Glasgow, United Kingdom
mark.mcgill@glasgow.ac.uk

Jan Gugenheimer
TU-Darmstadt
Darmstadt, Germany
jan.gugenheimer@tu-darmstadt.de

Abstract

As generative Artificial Intelligence (AI) becomes increasingly embedded and utilized for digital design, it presents both opportunities and risks. One major concern is its potential to facilitate and incorporate deceptive design patterns into computing technologies, which could manipulate or mislead users to their disadvantage. Similar to the concept of precedent-based design, a common approach in design theory that suggests reapplying previous design solutions to similar or identical problems, generative AI can integrate deceptive design patterns included in the training data a model has seen before. Our workshop explores how generative AI suggests and enacts deceptive design patterns in digital design. The goal of the workshop is to explore the ethical challenges of utilizing generative AI models and develop strategies to detect or prevent manipulative practices, thereby creating more transparent and equitable AI-generated experiences.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; **Interaction design**.

Keywords

LLMs, Deceptive Design, Dark Patterns, Manipulative Interfaces

ACM Reference Format:
Thomas Kosch, Veronika Krauß, Christopher Katins, Dominik Schön, Mark McGill, and Jan Gugenheimer. 2026. The AI Accomplice: Exploring Generative Artificial Intelligence in Facilitating and Amplifying Deceptive Designs. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3772363.3778770

CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3778770

1 Motivation

Generative Artificial Intelligence (AI) has become a powerful tool to streamline creative and productive processes. For example, Large Language Models (LLMs), such as ChatGPT, Gemini, Llama, and Claude, are highly effective tools for text generation through user prompts [23]. Quickly, generative AI became popular for creating other forms of media, including image [16], video [19], and audio [18] generation. These models are trained on large amounts of data, and their output essentially recombines pre-existing information and ideas. Similar to the concept of precedent-based design, a well-known strategy in design theory that focuses on reusing past design solutions for comparable or identical problems [4], generative AI can spread design knowledge by replicating and reapplying existing concepts. Consequently, generative AI emerged as a tool for rapid interaction design and content creation [20].

Human biases shape the output of generative AI. While reusing existing design concepts can save time and resources, generative AI models may also produce deceptive designs [6], sometimes without the designer's awareness [12]. Because users often lack skills in prompt design [22], they may overlook deceptive elements in the generated output [12], or worse, intentionally delegate unethical practices to the AI [9]. Since generative AI is trained on online content, including deceptive design practices, it can replicate, adapt, and even amplify these strategies found across the web [13, 14]. Additional risks include the hallucination of misleading "established" patterns or outdated, erroneous designs in several application areas [1, 10, 11, 17].

We assert that generative AI has the potential to significantly contribute to the creation of deceptive design patterns, which can deliberately or inadvertently manipulate user decisions through the design process by proliferating deceptive designs through previously seen data in the training process. This workshop examines the potential risks of generative AI, with the goal of laying the groundwork for research on mitigating these risks and establishing safeguards against deceptive actions. The workshop aims to develop future design principles that emphasize transparency and safeguard user autonomy by analyzing how generative AI can be used to manipulate perceptions and deceive users. The workshop will foster

interdisciplinary collaboration in creating proactive solutions that protect users from AI-enabled deceptions through design. This workshop builds on prior workshops addressing similar themes in HCI [2, 3, 5, 7, 8, 15, 21], but focuses on the emerging challenges posed by generative AI. The workshop will consolidate position papers and organize interactive sessions among participants to discuss the current state of using generative AI for creating deceptive design patterns. The overall goal of the workshop is to critically examine how generative AI can create and amplify deceptive design patterns while collaboratively developing strategies and principles to detect, prevent, and counteract such manipulative practices in the HCI community.

2 Length of the Workshop

We plan a long workshop with two sessions. Each session is 90 minutes long.

3 Workshop Content & Activities

This workshop connects recent research around deceptive designs in AI generation to start, grow, and foster a community that creates awareness and investigates countermeasures. The participants will present their workshop submissions, participate in practical sessions where they generate deceptive designs, and establish awareness and countermeasures against potential generative AI-based perceptual attacks through generative AI.

We plan an in-person workshop led by the organizers. We will incorporate interactive elements, including group discussions, scenario-based thinking, and attempts at generalization. Further details on the submission formats will be provided during the call for participation. We plan interactive sessions where participants can engage with demos that showcase deceptive design concepts and implementations. We require one projector, tables, chairs, and power sockets for the participants and demos.

During the workshop, participants will actively engage in a series of interactive activities. We begin with an icebreaker exercise to foster collaboration, followed by short pitch presentations where attendees share their research statements on deceptive designs in generative AI. In the main interactive session, participants will be divided into groups and assigned a specific medium (e.g., text, images, audio, or video). Using curated generative AI models, each group will design and analyze deceptive patterns, critically reflecting on how these patterns influence user perception and behavior. They will then collaboratively develop countermeasures, such as detection strategies, literacy interventions, or interface design principles to mitigate deception. Each group will present its findings to the plenary, encouraging cross-group discussion, critique, and synthesis of the results. The workshop concludes with a moderated reflection to consolidate insights and establish directions for ongoing collaboration and dissemination.

4 Workshop Schedule

Pre-Workshop Plans: We will distribute information and materials on our workshop website. Information includes the intention, motivation, and potential outcomes of the workshop. Furthermore, the website serves as a platform to advertise and acquire potential workshop participants. A workshop website will be available for the participants. The website includes a workshop description, objectives, and possible submission topics. It also hosts a call for participation, a link to the submission system, the workshop schedule, additional organizational information, and details about the workshop organizers. Accepted papers will be made publicly available on the website prior to the conference, allowing for maximum preparation time for the workshop and fostering discussions. Finally, workshop participants can join our Slack channel to receive updates about the workshop and join our community. We will make the position papers available in advance for participants who wish to prepare for the workshop. At the same time, participants who can not participate can follow along with the research topics of the workshop by reading the position papers.

Workshop Plan: We plan a half-day workshop for around 20 participants and the following schedule:

(1) **Moderated icebreaker activity** (approx. 15 min): Workshop attendees participate in icebreaker sessions to get to know each other by physically grouping them.
(2) **Workshop introduction** (15 min): the organizers introduce themselves, the workshop topic, and the schedule.
(3) **Pitch presentations** of short papers and research statements (total up to 60 min): Up to 10 research statements, each 6 min long. These presentations aim to gather and showcase current findings on deceptive design patterns, with a focus on the role of generative AI in creating and amplifying these patterns. The goal is to collaboratively collect deceptive designs that emerge during the use of generative AI, outlining key mechanisms and tactics employed in deceptive design across various user interfaces.
(4) **Coffee break** (30 min)
(5) **Introduction of the deceptive design sessions** (10 min): the organizers introduce the interactive session.
(6) **Interactive session** (60 min): Group work and presentation of AI-generated deceptive designs and countermeasures. The workshop participants will be divided into groups, each focusing on a different content type (e.g., text, images, audio, or video) and utilizing generative models to explore deceptive designs. Each group will select a specific content type and choose from a curated list of generative AI models tailored for that medium, such as a text-based LLM for generating misleading articles or an image generation model for creating deceptive visuals. Groups will brainstorm and create examples of these deceptive designs, analyzing the mechanisms behind them and discussing their influence on perceptions and behaviors. Following this, they will collaboratively develop countermeasures to mitigate the impact of these deceptions, which may include strategies for detecting misleading content, promoting media literacy, or designing user interfaces that highlight potential misinformation. Finally, each group will present its findings, showcasing both the deceptive designs and the proposed countermeasures, fostering discussion and critique to deepen participants' understanding of the challenges posed by AI-generated content. This activity promises to be a fun and informative exploration of the intersection between AI, deception, and user awareness,

providing participants with practical skills and insights into the implications of AI-generated content.

(7) **Moderated discussion and closing** (20 min): the organizers moderate a discussion based on the pitch presentations and the interactive session. Finally, the workshop is closed.

Post-Workshop Plans: The keynote and talks will be recorded. These will be posted online as a YouTube series in future workshops. The accepted contributions will be encouraged to submit to preprint platforms such as arXiv or CEUR and will be posted on our workshop website. Participants will also be encouraged to develop their work further and submit a full paper to the CHI conference the following year. Everyone will be invited to join our online community of ongoing Slack channels for continued discussions and collaboration. We will publish an open-access paper summarizing the workshop results, the current state of deceptive design pattern generation in AI models, and an overview of the results from the interactive session. We will take notes of each session and publish them on the website for individuals interested in the workshop results.

5 Workshop Proceedings

After the workshop, we encourage authors to revise their publications based on the discussions and feedback received during the workshop. We will support researchers in submitting their final papers to arXiv[1] or CEUR[2] and as a proceedings collection on our website. Based on the workshop results, we will distill the critical aspects and outcomes into a position paper, which will be published with open access. The anticipated results address research questions concerning prototyping, study design, and the evaluation of deceptive designs generated by generative AI. The feedback from the workshop attendees accompanies these research questions, inspiring researchers interested in tackling the questions in collaboration with the workshop attendees. Based on the interest of the workshop attendees, we organize regular meetups. We plan to establish a long-term format with the potential for future invitations for authors to contribute to a journal.

6 Asynchronous Engagement & Accessibility

We will facilitate asynchronous engagement for those who are unable to attend in real-time. To achieve this, we will distribute the workshop slides and accepted papers on our website, allowing participants to explore the content at their own pace. We encourage participants to notify us of any accessibility requests in advance, as outlined in our call for papers and on our website, so that we can ensure a fully inclusive and accommodating environment. Whether it involves physical accessibility, captioning for virtual attendees, or any other specific needs, we are committed to making the workshop as accessible as possible for everyone.

7 Recruitment & Reviewing

The organizers use their social networks and mailing lists to disseminate the call for participation (see below). Submissions will be collected via EasyChair. Each submission will receive 2-3 reviews from PC members (i.e., the authors of this workshop proposal) and external reviewers.

8 Call for Participation

As generative Artificial Intelligence (AI) becomes central to digital design, it brings both opportunities and risks, particularly the potential for deceptive design patterns that can manipulate or mislead users. Similar to precedent-based design, generative AI may replicate harmful patterns embedded in its training data. Our workshop will explore how AI suggests and enacts deceptive design patterns, focusing on ethical challenges and strategies to prevent manipulative practices. We invite researchers and practitioners to join us in exploring these issues and developing approaches for more transparent and user-friendly AI-driven design solutions. The goal is to understand and recognize deceptive design patterns generated by generative AI, while establishing policies that mitigate the proliferation of such patterns. Submissions should follow the ACM two-column format with a length between two and four pages, excluding references. We solicit the following types of submissions: *position papers*, *research statements*, and *demonstrations*. We consider demonstrating an approach or technology that solicits deceptive designs. Workshop participants can discuss and try out the demonstration during the workshop breaks. The workshop website[3] has information about submitting papers. Contributions will be selected based on the merit of their contribution to the workshop. We encourage authors to make their research available on arXiv after the seminar. At least one author of each accepted submission is required to attend the workshop, and all participants are required to register.

9 Organizers

Thomas Kosch is a junior professor at Humboldt University of Berlin, Germany, where he studies how AI can facilitate an understanding between users and interfaces using computational interaction design. Furthermore, his work advances methodological rigor in human–AI interaction research by developing frameworks that improve reproducibility and transparency.

Veronika Krauß is a professor at the University of Applied Sciences Ansbach, Germany. Her work investigates social challenges originating from UI design for and with emerging technologies. Her main focus lies on adapting deceptive designs to XR and the role of design practices and tools.

Christopher Katins is a PhD student and HCI researcher at the Humboldt University of Berlin. His research focuses on the challenges and dangers that Mixed Reality might pose in the future.

Dominik Schön is a PhD student at the Technical University of Darmstadt, Germany. In his research, he investigates Ergonomic-aware Extended Reality Environments. In this regard, he examines how AI and LLMs can facilitate the streamlined implementation of such functional design.

Mark McGill (https://augsoc-project.org) is a senior lecturer (associate professor) in the School of Computing Science, University of Glasgow. His research explores the future of spatial computing and a world where perception is mediated by AI through everyday augmented reality.

[1] https://arxiv.org
[2] https://ceur-ws.org
[3] www.hcistudio.org/ai_accomplice

Jan Gugenheimer (www.gugenheimer.com) is a Professor at the Technical University of Darmstadt. His research focuses on the emerging social challenges for mixed reality technology and how to integrate HMDs into the fabric of our daily lives.

Acknowledgments

This work was funded by the ERC (Project-ID 101116910) and UK Research and Innovation (UKRI) under the UK Government's Horizon Europe funding guarantee (AUGSOC) [EP/Z000068/1]. This work was also supported by the AI Safety Institute (AISI) under the WearAI project. Furthermore, this work was supported by the German Research Foundation (DFG), CRC 1404: "FONDA: Foundations of Workflows for Large-Scale Scientific Data Analysis" (Project-ID 414984028).

References

[1] William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. 2024. The Illusion of Artificial Inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 286, 12 pages. https://doi.org/10.1145/3613904.3642703

[2] Minsuk Chang, John Joon Young Chung, Katy Ilonka Gero, Ting-Hao Kenneth Huang, Dongyeop Kang, Vipul Raheja, Sarah Sterman, and Thiemo Wambsganss. 2024. Dark Sides: Envisioning, Understanding, and Preventing Harmful Effects of Writing Assistants - The Third Workshop on Intelligent and Interactive Writing Assistants. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 464, 6 pages. https://doi.org/10.1145/3613905.3636312

[3] Alicia Cork, Mike Richardson, Heide Karen Lukosch, Mohamed Khamis, Christopher Katins, Veronika Krauß, Lauren Ruffin, Sarah Papazoglakis, Victoria Sanchez, Xueni Pan, et al. 2025. Shaping the future: principles for policy recommendations for responsible innovation in virtual worlds. *Frontiers in Virtual Reality* 6 (2025), 1645330. https://doi.org/10.3389/frvir.2025.1645330

[4] Buthayna Hasan Eilouti. 2009. Design knowledge recycling using precedent-based analysis and synthesis models. *Design Studies* 30, 4 (2009), 340–368. https://doi.org/10.1016/j.destud.2009.03.001

[5] Colin M. Gray, Johanna T. Gunawan, René Schäfer, Nataliia Bielova, Lorena Sanchez Chamorro, Katie Seaborn, Thomas Mildner, and Hauke Sandhaus. 2024. Mobilizing Research and Regulatory Action on Dark Patterns and Deceptive Design Practices. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 482, 6 pages. https://doi.org/10.1145/3613905.3636310

[6] Colin M. Gray, Lorena Sanchez Chamorro, Ike Obi, and Ja-Nae Duane. 2023. Mapping the Landscape of Dark Patterns Scholarship: A Systematic Literature Review. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 188–193. https://doi.org/10.1145/3563703.3596635

[7] Jan Gugenheimer, Wen-Jie Tseng, Abraham Hani Mhaidli, Jan Ole Rixen, Mark McGill, Michael Nebeling, Mohamed Khamis, Florian Schaub, and Sanchari Das. 2022. Novel Challenges of Safety, Security and Privacy in Extended Reality. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 108, 5 pages. https://doi.org/10.1145/3491101.3503741

[8] Christopher Katins, Jannis Strecker, Jan Hinrichs, Pascal Knierim, Bastian Pfleging, and Thomas Kosch. 2025. Ad-Blocked Reality: Evaluating User Perceptions of Content Blocking Concepts Using Extended Reality. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 626, 18 pages. https://doi.org/10.1145/3706598.3713230

[9] Nils Köbis, Zoe Rahwan, Raluca Rilla, Bramantyo Ibrahim Supriyatno, Clara Bersch, Tamer Ajaj, Jean-François Bonnefon, and Iyad Rahwan. 2025. Delegation to Artificial Intelligence can increase dishonest behaviour. *Nature* (2025), 1–9. https://doi.org/10.1038/s41586-025-09505-x

[10] Thomas Kosch and Sebastian Feger. 2024. Risk or Chance? Large Language Models and Reproducibility in Human-Computer Interaction Research. arXiv:2404.15782 [cs.HC] https://arxiv.org/abs/2404.15782

[11] Thomas Kosch and Sebastian Feger. 2025. Prompt-Hacking: The New p-Hacking? *arXiv preprint arXiv:2504.14571* (2025). https://doi.org/10.48550/arXiv.2504.14571

[12] Veronika Krauß, Mark McGill, Thomas Kosch, Yolanda Maira Thiel, Dominik Schön, and Jan Gugenheimer. 2025. "Create a Fear of Missing Out" - ChatGPT Implements Unsolicited Deceptive Designs in Generated Websites Without Warning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 857, 20 pages. https://doi.org/10.1145/3706598.3713083

[13] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 81 (Nov. 2019), 32 pages. https://doi.org/10.1145/3359183

[14] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 360, 18 pages. https://doi.org/10.1145/3411764.3445610

[15] Mike Richardson, Alicia G Cork, Danaë Stanton Fraser, Michael J Proulx, Xueni Pan, Veronika Krauß, Mohamed Khamis, and Heide Lukosch. 2024. Shaping The Future: Developing Principles for Policy Recommendations for Responsible Innovation in Virtual Worlds. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 488, 6 pages. https://doi.org/10.1145/3613905.3636306

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

[17] Michael T. Rücker, Carolin Büchting, and Thomas Kosch. 2025. Understanding the Effect of Risk Perception on the Acceptance and Use of Large Language Models Among University Students. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW512 (Oct. 2025), 21 pages. https://doi.org/10.1145/3757693

[18] Robin San Roman, Yossi Adi, Antoine Deleforge, Romain Serizel, Gabriel Synnaeve, and Alexandre Defossez. 2023. From Discrete Tokens to High-Fidelity Audio Using Multi-Band Diffusion. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 1526–1538. https://proceedings.neurips.cc/paper_files/paper/2023/file/054f771d614df12fe8def8ecdbe4e8e1-Paper-Conference.pdf

[19] Rui Sun, Yumin Zhang, Tejal Shah, Jiahao Sun, Shuoying Zhang, Wenqi Li, Haoran Duan, Bo Wei, and Rajiv Ranjan. 2024. From Sora What We Can See: A Survey of Text-to-Video Generation. arXiv:2405.10674 [cs.CV] https://arxiv.org/abs/2405.10674

[20] Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 378, 22 pages. https://doi.org/10.1145/3613904.3642466

[21] Anastasiya Zakreuskaya, Tobias Münch, Henrik Detjen, Sven Mayer, Passant Elagroudy, Bastian Pfleging, Fiona Draxler, Benjamin Weyers, Uwe Gruenefeld, Jonas Auda, Waldemar Titov, Wendy E. Mackay, Daniel Buschek, and Thomas Kosch. 2024. Workshop on Generative Artificial Intelligence in Interactive Systems: Experiences from the Community. Mensch und Computer 2024 - Workshopband. In *Proceedings of Mensch und Computer 2024*. Gesellschaft für Informatik e.V. https://doi.org/10.18420/muc2024-mci-ws09-123

[22] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. https://doi.org/10.1145/3544548.3581388

[23] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL] https://arxiv.org/abs/2303.18223