Leschanowsky A., Popp B., Peters N. (2022): Adapting Debiasing Strategies for Conversational AI. In: Proceedings of the International Conference on Privacy-friendly and Trustworthy Technology for Society – COST Action CA19121 - Network on Privacy-Aware Audio- and Video-Based Applications for Active and Assisted Living

Adapting Debiasing Strategies for Conversational AI

Anna Leschanowsky[1,2,3], Birgit Popp[2], Nils Peters[1,3]
[1]International Audio Laboratories Erlangen*, Germany
[2]Fraunhofer IIS, Germany
[3]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
anna.leschanowsky@fau.de, birgit.popp@iis.fraunhofer.de, nils.peters@fau.de

Abstract

Conversational AI (CAI) systems such as smart speakers or virtual assistants are widely adopted in our daily lives. While many users report privacy concerns, only few engage in privacy-protective strategies. This privacy paradox can leave users uncertain and frustrated. One explanation for the mismatch of behavior and attitudes could be that users' decision-making is subject to heuristics and biases. Debiasing strategies can help users to make rational decisions about their privacy that are aligned with their values. While nudging approaches have been applied in privacy research, little is known about other available debiasing strategies. We introduce debiasing strategies known from the medical field and show their applicability and usefulness in CAI systems.

* The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS.

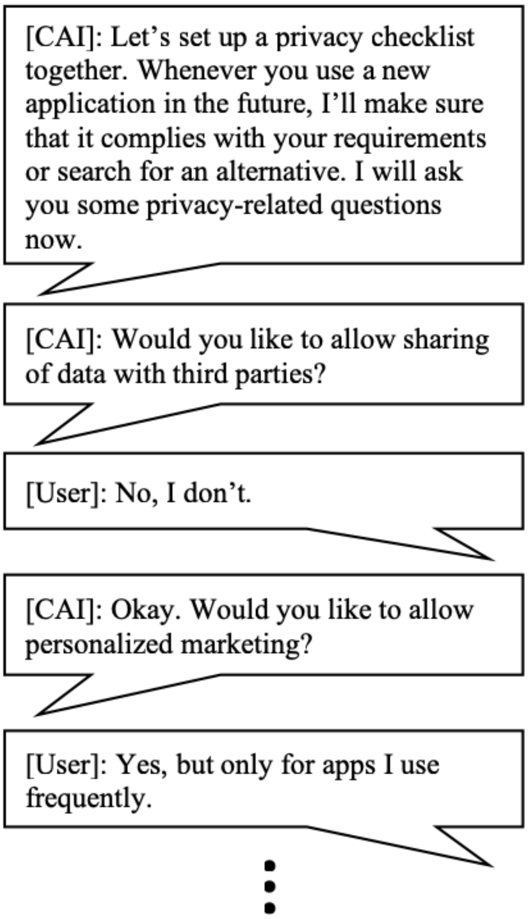
Introduction

The privacy paradox describes the discrepancy between people's attitudes towards privacy and their actual behavior and has sparked controversial debates in the field of privacy research (Kokolakis, 2017; Solove, 2021). It has been investigated in contexts such as e-commerce, social networks and CAI (Barth & de Jong, 2017; Konrad et al., 2020; Masur, 2019; Williams et al., 2017). Behavioral economics and decision research have been applied to investigate how heuristics and cognitive biases influence privacy decision-making (Acquisti, 2009). Differences in risk and benefit perception and judgement can lead users to weigh benefits higher than risks and thus engage less in privacy-protective strategies (Barth & de Jong, 2017; Leschanowsky et al., 2021). Previous research in the privacy context has shown the applicability of nudges and soft paternalism solutions as an appealing concept to improve security and privacy decisions (Acquisti, 2009).

According to Thaler & Sunstein (2008) a nudge “is any aspect of the choice architecture that alters people’s behavior in a predictable way without forbidding any options or significantly changing their economic incentives”. Nudging approaches have been investigated in the field of mobile apps, app development and social media (Almuhimedi et al., 2015; Choe et al., 2013; Lambe et al., 2016; Wang et al., 2013). In addition, the medical field is particularly rich with empirically evaluated strategies that enable practitioners to overcome cognitive biases and avoid diagnostic errors. Thus, we will introduce four debiasing strategies known from the medical field that can improve privacy decision-making and show their applicability to CAI.

Debiasing Strategies

Checklists are a common tool to reduce cognitive failures as they provide consistency and ensure the completeness of a task. Diagnostic checklists or debiasing checklists have been investigated in the medical context resulting partly in fewer errors (Lambe et al., 2016). Usually, such checklists state possible alternative diagnoses, special diagnoses that should not be missed or provide step-by-step guidance to diagnosis (Ely et al., 2011). CAI allows the creation and management of checklists and recently a voice-controlled surgery checklist for anesthesiologists which ensures that critical safety steps are carried out has been developed (*Voice Controlled Checklist App | Softengi.Com*). A privacy-related checklist could be used to check user-specific privacy requirements before installing a new application (see Figure 1 for an example interaction). Such checklists can include items on privacy settings, can be context-dependent and present users with alternative applications. In the context of active and assisted living (AAL) technologies, privacy concerns and in particular the lack of privacy control have been shown to be one of the most prevalent barriers to acceptance



[CAI]: Let's set up a privacy checklist together. Whenever you use a new application in the future, I'll make sure that it complies with your requirements or search for an alternative. I will ask you some privacy-related questions now.

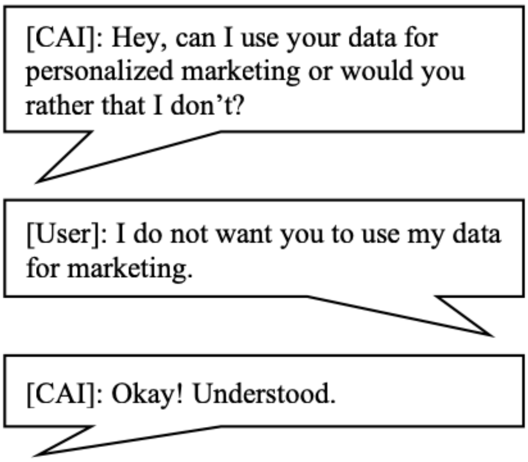
[CAI]: Would you like to allow sharing of data with third parties?

[User]: No, I don't.

[CAI]: Okay. Would you like to allow personalized marketing?

[User]: Yes, but only for apps I use frequently.

Figure 1. Example interaction for setting up a privacy-related checklist. Users can set their privacy requirements using natural language and CAI will ensure that only applications are used that comply with users' privacy requirements.



[CAI]: Hey, can I use your data for personalized marketing or would you rather that I don't?

[User]: I do not want you to use my data for marketing.

[CAI]: Okay! Understood.

Figure 2. Example interaction using a cognitive forcing strategy where users are presented with alternative options. Considering alternatives can help users to engage in privacy decision-making that is aligned with their values. Cognitive forcing strategies can also include reconsideration of decisions.

and adoption of the technology (Jaschinski & Ben Allouch, 2019; Schomakers & Ziefle, 2019). Privacy-related checklists can allow users to easily specify data recipients, data type and frequency of transmission. Due to the conversational nature of CAI, such checklists can be set up by care receivers without dedicated technological knowledge. Nevertheless, care receivers' safety may be negatively impacted when strong control settings are applied (Jaschinski & Ben Allouch, 2019). To counter this, CAI can communicate benefits and risks of specific data transmissions and implications for safety in natural language to the users and thus, a balance between privacy and safety can be ensured.

Cognitive forcing strategies – "a specific debiasing technique that introduces self-monitoring of decisionmaking [sic!]" – can be applied to broaden clinicians' views during diagnostic processing and allow them to consider alternative diagnoses (Croskerry, 2003). While cognitive forcing strategies originated from the medical education field, they have been applied as workplace strategies that can support clinicians at the time of decision-making (Lambe et al., 2016). It was found that when clinicians were asked to consider alternative diagnoses or reconsider diagnoses compared to diagnosing based on first impression, diagnostic accuracy increased (Lambe et al., 2016). In the privacy context, cognitive forcing strategies can be applied to support the process of rational cost-benefit analysis. Instead of making fast and intuitive decisions about disclosure or storage of one's personal information, CAI can present users with alternatives (see Figure 2 for an example interaction) or offer them the option to reconsider their decision to share data. While cognitive forcing strategies can help users to overcome their cognitive biases and consider costs and benefits more rationally, assessing costs can still be extremely difficult as privacy harms might only become apparent in the future due to new ways of data aggregation and analysis (Solove, 2012). CAI could be used to monitor data usage, inform users if necessary, and offer them the option to reconsider their decisions at the time of actual usage of their data.

Guided reflection refers to a concept in "which the practitioner is assisted by a mentor (or 'guide') in a process of self-enquiry, development, and learning through reflection" and has led to increased diagnostic accuracy when applied in the medical field (Johns, 2010; Lambe et al., 2016). The reflective practice should lead to more critical thinking of one's decision-making process. Studies on guided reflection have also used sets of procedures to diagnose a case (Lambe et al., 2016). Different to checklists where one might be reminded of possible alternative diagnoses, in studies on guided reflection participants were given detailed instruction on what to consider e.g. "list findings that support this hypothesis" (Mamede et al., 2008). In the CAI privacy context, CAI offers unique possibilities to function as a guide and to assist users in their development and privacy decision-making. Especially with the adoption of large language models such as OpenAI's GPT-3 or Meta AI's OPT-175B (Brown et al., 2020; *Meta AI is sharing OPT-175B*), CAI can be capable of acting as a guide to users that otherwise do not have access to human mentors and reflective practices. However, it needs to be ensured that language models can be trusted and that mentoring on decision-making is unbiased. Moreover, CAI could trigger reflective reasoning by asking privacy-related questions. Asking users to find and list privacy-related information themselves, could be seen as an educational strategy that raises awareness for the topic and could lead to a state where users automatically engage in reflective reasoning before disclosing personal information.

Lastly, **instructions** were used by researchers in the medical context to reduce diagnostic errors (Lambe et al., 2016). Instructions covered dual-process reasoning, a list of clinical features and thoughtful diagnosis (Lambe et al., 2016). In the CAI privacy context, instructions can be easily applied to interrupt users' intuitive decision-making and make them think more carefully about privacy decisions. For example, CAI could instruct users to consider the types of information that are collected or ask them to think thoroughly about how their information will be used before installing a new application. While these instructions can help to overcome cognitive biases, conversational systems need to provide ways to answer possible follow-up questions from the users. Thus, similarly to easily understandable privacy labels (Kelley et al., 2009), CAI should be able to efficiently communicate privacy policies and their implications.

Conclusion and Future Work

We introduced four debiasing strategies known from the medical research field and showed their applicability and usefulness in the context of privacy and conversational AI. Debiasing strategies can support users to overcome the discrepancy between their behavior and their values regarding the disclosure of personal information. Due to the conversational and human-like capabilities and its accessibility, CAI could uniquely ensure that users engage in decision-making aligned with their values. Therefore, future research is needed to investigate debiasing techniques for privacy decision-making in the context of CAI. Moreover, debiasing strategies could not only be applied at the time of decision-making but could be used as educational strategies to raise awareness and spark discussions around the topic. Future work could consider and investigate the long-term influences of debiasing strategies on users' privacy decision-making.

References

Acquisti, A. (2009): 'Nudging Privacy: The Behavioral Economics of Personal Information'. *IEEE Security & Privacy Magazine*, vol. *7*, no. 6, 2009, pp. 82–85.

Almuhimedi, H., Schaub, F., Sadeh, N., Adjerid, I., Acquisti, A., Gluck, J., Cranor, L. F., & Agarwal, Y. (2015): 'Your Location has been Shared 5,398 Times! A Field Study on Mobile App Privacy Nudging'. in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, USA, 2015, pp. 787–796.

Barth, S., & de Jong, M. D. T. (2017): 'The privacy paradox – Investigating discrepancies between expressed privacy concerns and actual online behavior – A systematic literature review'. *Telematics and Informatics*, vol. *34, no.* 7, pp. 1038–1058.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.. (2020): 'Language Models are Few-Shot Learners', In Larochelle, H. and Ranzato, M. and Hadsell, R. and Balcan, M.F. and Lin, H., *Advances in neural information processing systems,* Curran Associates, Inc., 2020, pp. 1877 - 1901

Choe, E. K., Jung, J., Lee, B., & Fisher, K. (2013): 'Nudging People Away from Privacy-Invasive Mobile Apps through Visual Framing'. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M.

Winckler, *Human-Computer Interaction – INTERACT 2013*. Springer Berlin Heidelberg, 2013, pp. 74-91

Croskerry, P. (2003): 'Cognitive forcing strategies in clinical decisionmaking'. *Annals of Emergency Medicine*, vol. *41, no.* 1, pp. 110–120

Ely, J. W., Graber, M. L., & Croskerry, P. (2011): 'Checklists to Reduce Diagnostic Errors'. *Academic Medicine*, vol. 86, no. 3, 2011, pp. 307–313

Jaschinski, C., & Ben Allouch, S. (2019): 'Listening to the ones who care: Exploring the perceptions of informal caregivers towards ambient assisted living applications'. *Journal of Ambient Intelligence and Humanized Computing*, vol. *10, no.* 2, pp. 761–778.

Johns, C. (2010). *Guided reflection: A narrative approach to advancing professional practice* (2nd ed). Johns, C. Blackwell Pub.

Kelley, P. G., Bresee, J., Cranor, L. F., & Reeder, R. W. (2009): 'A „nutrition label“ for privacy'. In *Proceedings of the 5th Symposium on Usable Privacy and Security - SOUPS '09*, Association for Computing Machinery, New York, USA, 2009, pp. 1 - 12

Kokolakis, S. (2017): 'Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon.' *Computers & Security*, vol. *64*, 2017, pp. 122–134.

Konrad, M., Koch-Sonneborn, S., & Lentzsch, C. (2020): 'The Right to Privacy in Socio-Technical Smart Home Settings: Privacy Risks in Multi-Stakeholder Environments'. In C. Stephanidis & M. Antona (Hrsg.), *HCI International 2020—Posters*, Springer International Publishing, pp. 549 - 557

Lambe, K. A. O'Reilly, G., Kelly, B. D., & Curristan, S. (2016): 'Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review'. *BMJ quality & safety*, vol. 25, no. 10, 2016, pp. 808–820.

Leschanowsky, A., Brüggemeier, B., & Peters, N. (2021): 'Design Implications for Human-Machine Interactions from a Qualitative Pilot Study on Privacy'. *2021 ISCA Symposium on Security and Privacy in Speech Communication*, pp. 76–79

Mamede, S., Schmidt, H. G., & Penaforte, J. C. (2008): 'Effects of reflective practice on the accuracy of medical diagnoses'. *Medical Education*, vol. *42*, no. 5, 2008, pp. 468–475.

Masur, P. K. (2019): *Situational Privacy and Self-Disclosure: Communication Processes in Online Environments* (1st ed. 2019). Springer International

Meta AI is sharing OPT-175B, Retrieved May 13 2022, from https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/

Schomakers, E.-M., & Ziefle, M. (2019): 'Privacy Perceptions in Ambient Assisted Living', In *Proceedings of the 5th International Conference on Information and Communication Technologies for Ageing Well and E-Health*, SCITEPRESS - Science and Technology Publications, Crete, Greece, 2019, pp. 205–212

Solove, D. J. (2021): 'The Myth of the Privacy Paradox'. *89 George Washington Law Review 1 (2021), GWU Legal Studies Research Paper No. 2020-10, GWU Law School Public Law Research Paper No. 2020-10*