Article Identifier
DOI: 10.22032/dbt.58916
URN: urn:nbn:de:gbv:ilm1-2023isc-131:5

60th ILMENAU SCIENTIFIC COLLOQUIUM
Technische Universität Ilmenau, 4 – 8 September 2023
URN: urn:nbn:de:gbv:ilm1-2023isc:1

BUILDING AN APPROPRIATE LEVEL OF TRUST WITH CONVERSATIONAL AGENTS AS TEAM PARTNERS FOR LEARNING

Katharina Jahn, Felix Krieglstein, Lewis L. Chuang, Günter Daniel Rey

Institute for Media Research, Faculty of Humanities, Chemnitz University of Technology

1. INTRODUCTION

Conversational artificial intelligence (AI), such as OpenAI's ChatGPT, are greatly expected to enable effective teaming between humans and AIs, which will increase performance and efficiency across a wide range of tasks. For instance, conversational AI introduces opportunities in education for individualized tutoring and has been integrated into popular self-learning platforms (e.g., Khanmigo, https://khanacademy.org). Whilst conversational AI can provide highly persuasive answers, these answers can nonetheless be false, biased, or even malicious. For example, ChatGPT can create false "facts" and references [1] and might even generate hostile responses. When used as a tool for writing essays, it may not necessarily improve the existing writing quality [2]. Thus, the effective use of conversational AI is, itself, a skill that must be learned before it can be used reliably. Its application should be assumed to lead to improved performance and special care must be taken to prevent humans from trusting or relying on it more than its actual performance. Since conversational AIs can return inaccurate and inappropriate answers, it is important for humans to develop a level of trust (and skepticism) toward them that is appropriate to their actual competence. Previous attempts to use explainable AI have shown limited success in the area of decision support systems [3]. A likely explanation for this effect stems from dual-process theories [4], which propose that information is more often processed in the fast, automatic system compared to the slower, reflective system. To mitigate these shortcomings, some research has successfully used cognitive forcing functions. Cognitive forcing functions enforce cognitive processing through various means (e.g., by presenting the AI only after the human team partner has accomplished a certain task) [5]. Here, we propose a study of the effect of using a cognitive forcing function with anthropomorphic design on trust and learning. Our research question is: *How can an appropriate level of trust towards conversational AI be created in an educational context?*

2. METHOD

Design: We propose a 2 (cognitive forcing function: learning with the AI generated paper summary first vs. learning with the original paper first) x 2 (anthropomorphic design: lower vs. higher) between-subjects experiment. After conducting a power analysis using G*Power with a medium effect size of f = .25, we aim for 128 participants for a power of 80 %. Participants will be recruited from the local university's master's program to ensure they are familiar with the learning material and will receive course credit as compensation for participating in our study.
Materials: Because ChatGPT has previously made errors when summarizing the meta-analysis on cognitive behavioral therapy by van Dis et al. [1], we will use this meta-analysis [6] as learning content. The summary of this paper will be generated by ChatGPT and adapted by the authors to contain ten errors.

Procedure: After participants filled in an informed consent form, they will be asked for demographics and be given information on to the conversational AI. Next, they will be given information on the research paper they have to learn about. Depending on the condition, participants will be either prompted to learn with the conversational AI (containing ten errors) or to learn with the research paper. Afterwards, the other information prompt will be provided. The human-likeness of the text the conversational AI presents will be designed according to the anthropomorphic design condition. Next, they will answer the questionnaire and the multiple-choice questions. Finally, participants will be thanked and debriefed.

3. EXPECTED RESULTS, LIMITATIONS, AND OUTLOOK

The results of this experiment will contribute to understanding how teaming between humans and AI can take place in learning-relevant contexts. Thus, learning scenarios can be designed accordingly in relation to *when* and *how* an AI provides information and to design future experiments related to specific design principles for creating learning material. Although limitations exist regarding prescribing specific interaction times, which might not be feasible in all real-world contexts, the proposed experiment can inform future research on how human-AI teaming should be structured in learning contexts. The next steps consist of receiving feedback for the study implementation, creating the materials and subsequently conducting the experiment.

REFERENCES

[1] E. A. M. van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting, "ChatGPT: five priorities for research," *Nature*, vol. 614, no. 7947, pp. 224–226, 2023.

[2] Ž. Bašić, A. Banovac, I. Kružić, and I. Jerković, "Better by You, Better Than Me? Chatgpt-3 as Writing Assistance in Students' Essays," *EdArXiv Febr.*, vol. 9, 2023.

[3] G. Bansal *et al.*, "Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan, pp. 1–16, 2021.

[4] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, p. 188:1-188:21, 2021.

[5] C. C. Dobler, A. S. Morrow, and C. C. Kamath, "Clinicians' cognitive biases: a potential barrier to implementation of evidence-based clinical practice," *BMJ Evid.-Based Med.*, vol. 24, no. 4, pp. 137–140, 2019.

[6] E. A. M. van Dis *et al.*, "Long-term Outcomes of Cognitive Behavioral Therapy for Anxiety-Related Disorders: A Systematic Review and Meta-analysis," *JAMA Psychiatry*, vol. 77, no. 3, pp. 265–273, 2020.

CONTACTS

Dr. Katharina Jahn — email: katharina.jahn@phil.tu-chemnitz.de
ORCID: https://orcid.org/0000-0002-9943-5279

Felix Krieglstein — email: felix.krieglstein@phil.tu-chemnitz.de
ORCID: https://orcid.org/0000-0002-1324-3816

Prof. Dr. Lewis L. Chuang — email: lewis.chuang@phil.tu-chemnitz.de
ORCID: https://orcid.org/0000-0002-1975-5716

Prof. Dr. Günter Daniel Rey — email: guenter-daniel.rey@phil.tu-chemnitz.de
ORCID: https://orcid.org/0000-0001-9717-8478