Building an Appropriate Level of Trust with Conversational Agents as Team Partners for Learning

ANONYMOUS AUTHOR(S)

Conversational artificial intelligence (AI), such as ChatGPT, can be expected to provide new opportunities for educational purposes. For example, conversational AI can assist learners in writing texts, in learning topics from automatically generated summaries, and provide suggestions for further topics. Whereas previous research has often focused on improving trust towards AI (e.g., through anthropomorphic design), research on finding effective ways to develop an appropriate level of trust in AI-based educational scenarios is relatively scarce. Research in relation to AI-equipped decision support systems has already shown that forcing users to engage with the information provided through cognitive forcing functions may enhance decisions by creating an appropriate level of trust. Therefore, we aim to test whether these functions can also help to create an appropriate level of trust for conversational AI used for learning, even when the means of improving trust through anthropomorphic design to enhance use are implemented. Here, we present a planned study to investigate this, namely a 2 (cognitive forcing function: presenting the AI generated paper summary first vs. presenting the original paper first) x 2 (anthropomorphic design: lower vs. higher) between-subjects experiment.

Additional Key Words and Phrases: human-AI teaming, trust, cognitive forcing functions, anthropomorphic design

ACM Reference Format:
Anonymous Author(s). 2023. Building an Appropriate Level of Trust with Conversational Agents as Team Partners for Learning. 1, 1 (February 2023), 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 [20, 23]. They can compose written texts that are difficult to distinguish from actual human-written text, are described as eliciting sympathy, and are seen as having the potential to serve as innovator for a range of tasks. As a consequence, conversational AIs could serve as a readily available partner to humans, that could enhance their individual performance and learning.

Teaming between humans and AI is an promising way to solve current problems, associated with the use of technology as either a tool or as an automated replacement for human involvement [3, 21]. Some research has shown that combined decisions between humans and systems equipped with AI can be more effective than decisions of humans or AI systems alone.

Teaming has been investigated across a range of topics, for example in relation to increasing performance in work or learning tasks. In relation to education, opportunities for conversational AI have been indicated for in areas such as course selection [8], second language learning [1], and instructional scaffolding [26],

However, even though conversational AI can provide persuasive answers, the same answers could also be false, biased, or even malicious. For example, ChatGPT responds with wrong answers when asked to summarize a review [24], Google's Bard replied with a wrong answer in a promotional video [12], and Microsoft's conversational AI on

Bing has even replied with threats [11]. When used as a tool for writing essays, ChatGPT did not always improve the quality of students' essays [4]. This research indicates that using conversational AI effectively is a skill that must be learned. It cannot be expected and special care must be taken to prevent overtrust.

Since conversational AIs may give inaccurate or dangerous answers, it is important for humans to develop a level of trust (and skepticism) towards them that is appropriate to their competence. Previous attempts to use explainable AI have shown limited success in the area of decision support systems [2, 14, 15]. A likely explanation for this effect stem from dual-process theories [5, 10] that propose that information is more often processed in the fast, automatic system and compared to the slower, reflective system. To mitigate this shortcomings, some research has successfully used cognitive forcing functions [5]. 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) [9].

In this research proposal, we aim to investigate the effect of using a cognitive forcing function combined with anthropomorphic design on trust and learning to answer the following research question: *How can an appropriate level of trust towards conversational AI be created in an educational context?* This paper is organized as followed. First, we describe theoretical insights regarding trust in relation to AI and develop hypotheses. Second, we describe the experiment with its respective materials and measures in the methods section. Finally, we provide an outlook on how the expected results of our experiment could improve theoretical understanding.

2 THEORETICAL BACKGROUND AND HYPOTHESES

This section describes the theoretical background related to trust in AI, possibilities to affect this trust with different types of technological design variants, and the relevance of trust for learning and teaming. On the basis of this theoretical background, the hypotheses are developed. An overview on the constructs used in this paper is given in table 1.

2.1 Definitions

Trust can refer to a diverse set of concepts, ranging from attitudes to behavior [16]. In this paper, we investigate trust in conversational AI in the form of trusting beliefs. Trusting beliefs refer to the degree to which an conversational AI is perceived as beneficial in relation to benevolence, competence, and integrity [13, 16]. In relation to this, an appropriate level of trust is reached when the individual's trusting beliefs in an conversational AI adequately reflect the capabilities of the AI.

Table 1. Definitions

Construct	
Conversational AI	Conversational AI can be defined as artificial agents using natural language to communicate with humans via auditory or text-based means [13, 22]
Trusting beliefs	We define trusting beliefs as the attitude that an conversational AI is beneficial to the trustee, comprised of the components competence, and integrity in a potentially risky situation [17].
Appropriate level of trust	We define an appropriate level of trust as the extent to which an individual's trusting beliefs towards an conversational AI and their capabilities match [16, 27].

2.2 Trust in Relation to Conversational AI

Previous research on trust towards AI has oftentimes investigated how trust can be improved [6, 13]. For example, increasing the anthropomorphic design of AI can improve trust in the context of a health chatbot [19], and increases trust resilience [7]. However, merely improving trust an AI can lead to overtrust. Therefore, it is important to develop an appropriate level of trust in AI. Previous research has already investigated explainable AI and disclosing confidence levels [27]. Additionally, cognitive forcing functions can improve appropriate levels of trust for AI in relation to decision making [5].

To the best of our knowledge, the effectiveness of cognitive forcing functions in combination with anthropomorphic design has not been investigated in educational contexts. Because anthropormorphic design can increase trust, and, consequently, use of conversational AIs it is likely that these features will be implemented in conversational AI. Therefore, it is relevant to look at how anthropomorphic design can influence trust and learning in combination with strategies aimed at creating appropriate levels of trust (i.e., cognitive forcing functions). Against this background, we propose the following hypotheses for scenarios in which learners are confonted with an conversational AI making errors:

Hypothesis 1. *Using a cognitive forcing function reduces trusting beliefs compared to not using it.*

Hypothesis 2. *Using a cognitive forcing function increases learning compared to not using it.*

Hypothesis 3. *Higher anthropomorphic design increases trust compared to lower anthropomorphic design.*

Hypothesis 4. *Higher anthropomorphic design decreases learning compared to lower anthropomorphic design.*

Hypothesis 5. *Trusting beliefs mediate the effect of a) using a cognitive forcing functions and b) anthropomorphic design on learning.*

3 METHODS

In this section, we describe the design of the planned experiment and the independent and dependent variables. Additionally, we explain the procedure and planned data analysis.

3.1 Participants and Design

We propose a 2 (cognitive forcing function: presenting the AI generated paper summary first vs. presenting 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 and will receive course credit as compensation for participation in our study.

3.2 Independent Variables

Cognitive forcing function is implemented by forcing the participant to either read the (alleged) summary of the paper from the conversational AI first or reading the original paper first. Thus, participants are either presented with the summary of the paper provided by the (alleged) conversational AI (i.e., ChatGPT) first. Alternatively, the paper they have to read will be presented first in a pdf viewer in the browser. As soon as the second part of the information has been presented, participants will be able to access both for the remaining learning period.

Similar to previous experiments on anthropomorphic design in relation trust [7], anthropomorphic design will be varied by either showing a digital representation of an actual human (higher anthropomorphic condition) or a digital agent with lesser anthropomorphic cues (lower anthropomorphic condition) next to the conversational AI.

3.3 Dependent Variables

We will measure trusting beliefs on a 7-point Likert scale using the three dimensions: benevolence (example item: The chatbot is interested in my well-being, not just its own.), competence (example item: The chatbot is truthful in its dealings with me.), and integrity (exampel item: Overall, the chatbot is a capable and proficient teacher.) of McKnight et al [18] adapted to the context of conversational AI (i.e., using the word ‘chatbot’ instead of the website name) and our learning context. Learning will be measured by a set of 20 multiple choice questions, half of them containing the wrong information of the conversational AI as distractor. As control variables, disposition to trust [18] and prior experience with conversational AI will be used.

3.4 Materials

Because ChatGPT has previously made errors when summarizing the meta analysis on cognitive behavioral therapy of van Dis et al. [24], we will use this meta analysis [25] as learning content. The summary of this paper will be generated by ChatGPT and adapted by the authors to contain ten errors.

3.5 Procedure

The experiment will be conducted online. After participants have received information on the experiment and given informed consent, 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. According to 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 conversational AI will be embodied according to the anthropomorphic design condition. Next, they will answer the questionnaire and the multiple choice questions. Finally, participants will be thanked and debriefed.

3.6 Data Analysis

We will use a 2 x 2 ANOVA to test hypotheses 1-4. Additionally, we will use regression-based causal mediation analysis to test hypothesis 5.

4 OUTLOOK

The proposed experiment is expected to propose insights on how cognitive forcing functions may affect learning. The results can be used to design future experiments related to specific design principles for creating learning material. The next steps consist of creating the learning materials, implementing and conducting the experiment.

REFERENCES

[1] Emmanuel Ayedoun, Yuki Hayashi, and Kazuhisa Seta. 2019. Adding Communicative and Affective Strategies to an Embodied Conversational Agent to Enhance Second Language Learners' Willingness to Communicate. *International Journal of Artificial Intelligence in Education* 29, 1 (March 2019), 29–57. https://doi.org/10.1007/s40593-018-0171-6

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. 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*. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445717

[3] Vernol Battiste, Joel Lachter, Summer Brandt, Armando Alvarez, Thomas Z. Strybel, and Kim-Phuong L. Vu. 2018. Human-Automation Teaming: Lessons Learned and Future Directions. In *Human Interface and the Management of Information. Information in Applications and Services (Lecture Notes in Computer Science)*, Sakae Yamamoto and Hirohiko Mori (Eds.). Springer International Publishing, Cham, 479–493. https://doi.org/10.1007/978-3-319-92046-7_40

[4] Željana Bašić, Ana Banovac, Ivana Kružić, and Ivan Jerković. 2023. Better by You, Better Than Me? Chatgpt-3 as Writing Assistance in Students' Essays. *EdArXiv. February* 9 (2023).

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 188:1–188:21. https://doi.org/10.1145/3449287

[6] Shalini Chandra, Anuragini Shirish, and Shirish C. Srivastava. 2022. To Be or Not to Be …Human? Theorizing the Role of Human-Like Competencies in Conversational Artificial Intelligence Agents. *Journal of Management Information Systems* 39, 4 (Oct. 2022), 969–1005. https://doi.org/10.1080/07421222.2022.2127441 Publisher: Routledge _eprint: https://doi.org/10.1080/07421222.2022.2127441.

[7] Ewart J. de Visser, Samuel S. Monfort, Ryan McKendrick, Melissa A. B. Smith, Patrick E. McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22 (2016), 331–349. https://doi.org/10.1037/xap0000092 Place: US Publisher: American Psychological Association.

[8] Massimiliano Dibitonto, Katarzyna Leszczynska, Federica Tazzi, and Carlo M. Medaglia. 2018. Chatbot in a Campus Environment: Design of LiSA, a Virtual Assistant to Help Students in Their University Life. In *Human-Computer Interaction. Interaction Technologies (Lecture Notes in Computer Science)*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 103–116. https://doi.org/10.1007/978-3-319-91250-9_9

[9] Claudia Caroline Dobler, Allison S. Morrow, and Celia C. Kamath. 2019. Clinicians' cognitive biases: a potential barrier to implementation of evidence-based clinical practice. *BMJ Evidence-Based Medicine* 24, 4 (Aug. 2019), 137–140. https://doi.org/10.1136/bmjebm-2018-111074 Publisher: Royal Society of Medicine Section: EBM analysis.

[10] Jonathan St B. T. Evans. 2010. Intuition and Reasoning: A Dual-Process Perspective. *Psychological Inquiry* 21, 4 (Nov. 2010), 313–326. https://doi.org/10.1080/1047840X.2010.521057 Publisher: Routledge _eprint: https://doi.org/10.1080/1047840X.2010.521057.

[11] Forbes. 2023. Bing Chatbot's 'Unhinged' Responses Going Viral. https://www.forbes.com/sites/siladityaray/2023/02/16/bing-chatbots-unhinged-responses-going-viral/ Section: Business.

[12] The Guardian. 2023. Google AI chatbot Bard sends shares plummeting after it gives wrong answer. *The Guardian* (Feb. 2023). https://www.theguardian.com/technology/2023/feb/09/google-ai-chatbot-bard-error-sends-shares-plummeting-in-battle-with-microsoft

[13] Peng Hu, Yaobin Lu, and Yeming (Yale) Gong. 2021. Dual humanness and trust in conversational AI: A person-centered approach. *Computers in Human Behavior* 119 (June 2021), 106727. https://doi.org/10.1016/j.chb.2021.106727

[14] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry* 11, 1 (Feb. 2021), 1–9. https://doi.org/10.1038/s41398-021-01224-x Number: 1 Publisher: Nature Publishing Group.

[15] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. https://doi.org/10.1145/10.1145/3313831.3376873 arXiv:2001.05871 [cs].

[16] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 Publisher: SAGE Publications Inc.

[17] D. Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F. Clay. 2011. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems* 2, 2 (2011). https://doi.org/10.1145/1985347.1985353

[18] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research* 13, 3 (Sept. 2002), 334–359. https://doi.org/10.1287/isre.13.3.334.81 Publisher: INFORMS.

[19] Nico Pietrantoni, Alfred Benedikt Brendel, R Stefan Greulich, and Fabian Hildebrandt. 2022. Follow Me If You Want to Live - Understanding the Influence of Human-Like Design on Users' Perception and Intention to Comply with COVID-19 Education Chatbots. ICIS 2022 Proceedings (2022).

[20] Jürgen Rudolph, Samson Tan, and Shannon Tan. 2023. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching* 6, 1 (Jan. 2023). https://doi.org/10.37074/jalt.2023.6.1.9 Number: 1.

[21] R. Jay Shively, Joel Lachter, Summer L. Brandt, Michael Matessa, Vernol Battiste, and Walter W. Johnson. 2018. Why Human-Autonomy Teaming?. In *Advances in Neuroergonomics and Cognitive Engineering (Advances in Intelligent Systems and Computing)*, Carryl Baldwin (Ed.). Springer International Publishing, Cham, 3–11. https://doi.org/10.1007/978-3-319-60642-2_1

[22] Xinmeng Song and Ting Xiong. 2021. A Survey of Published Literature on Conversational Artificial Intelligence. In *2021 7th International Conference on Information Management (ICIM)*. 113–117. https://doi.org/10.1109/ICIM52229.2021.9417135

[23] Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T. Hickey, Ronghuai Huang, and Brighter Agyemang. 2023. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments* 10, 1 (Feb. 2023), 15. https://doi.org/10.1186/s40561-023-00237-x

[24] Eva A. M. van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L. Bockting. 2023. ChatGPT: five priorities for research. *Nature* 614, 7947 (Feb. 2023), 224–226. https://doi.org/10.1038/d41586-023-00288-7 Bandiera_abtest: a Cg_type: Comment Number: 7947 Publisher: Nature Publishing Group Subject_term: Computer science, Research management, Publishing, Machine learning.

[25] Eva A. M. van Dis, Suzanne C. van Veen, Muriel A. Hagenaars, Neeltje M. Batelaan, Claudi L. H. Bockting, Rinske M. van den Heuvel, Pim Cuijpers, and Iris M. Engelhard. 2020. Long-term Outcomes of Cognitive Behavioral Therapy for Anxiety-Related Disorders: A Systematic Review and Meta-analysis. *JAMA Psychiatry* 77, 3 (March 2020), 265–273. https://doi.org/10.1001/jamapsychiatry.2019.3986

[26] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376781

[27] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852