UNIVERSITÄT DUISBURG ESSEN

Open-Minded

Psychologische Forschungsmethoden

Medienbasierte Wissenskonstruktion

University of Nottingham

UK | CHINA | MALAYSIA

Designing for human-AI collaboration: The effects of elaborateness and adaptability of explanations

Lydia Harbarth, Cora Weisenberger, Daniel Bodemer & Lenka Schnaubert

INTRODUCTION

- Successful human-AI collaboration requires an understanding of and **trust in AI systems.**
  → consider system and human factors when designing human-AI collaboration
- **System factors:** Trust in AI can be supported via system **transparency**, for example via **explanations** (Molina & Sundar, 2022). **Explainability** refers to the ability of a system to explain its functioning (Adadi & Berrada, 2018).
- **Human factors:** The **quality of explanations** can only be evaluated by users. **Causability** refers to perceived appropriateness of explanations to foster the users' understanding of the causal chain of system functioning (Holzinger et al., 2020).
  → integrate research from learning sciences and the explainable AI (xAI)
- **Individual characteristics** related to understanding:
  - **need or preference of differently elaborated explanations** (Putnam & Conati, 2019)
  - additional information may strain **cognitive resources** (Sweller, 2010)
- Users may benefit from
  - differently elaborated explanations (**elaborateness**)
  - the possibility to flexibly adjust (**adaptability**) the level of elaborateness of explanations

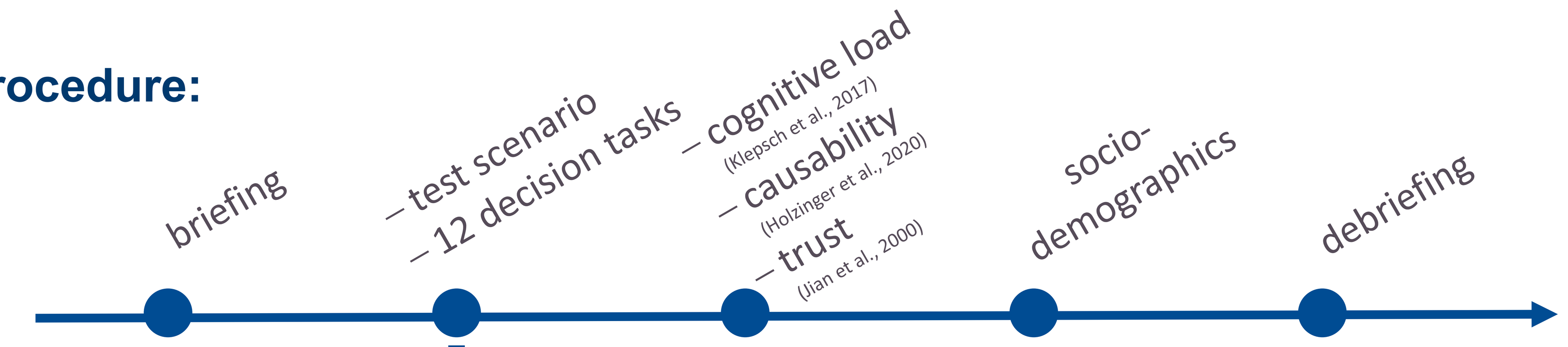
Research questions:

1. What are the effects of different types of instructional material (**level and adaptability of elaborateness of explanations**) on **causability** and **trust** in the system?
2. What is the role of **cognitive load** regarding **causability** and **trust**?

METHOD

Sample: N = 109 participants (31 m, 76 f, 2 d), age: M = 26.89 (SD = 11.35)

Procedure:



Design:

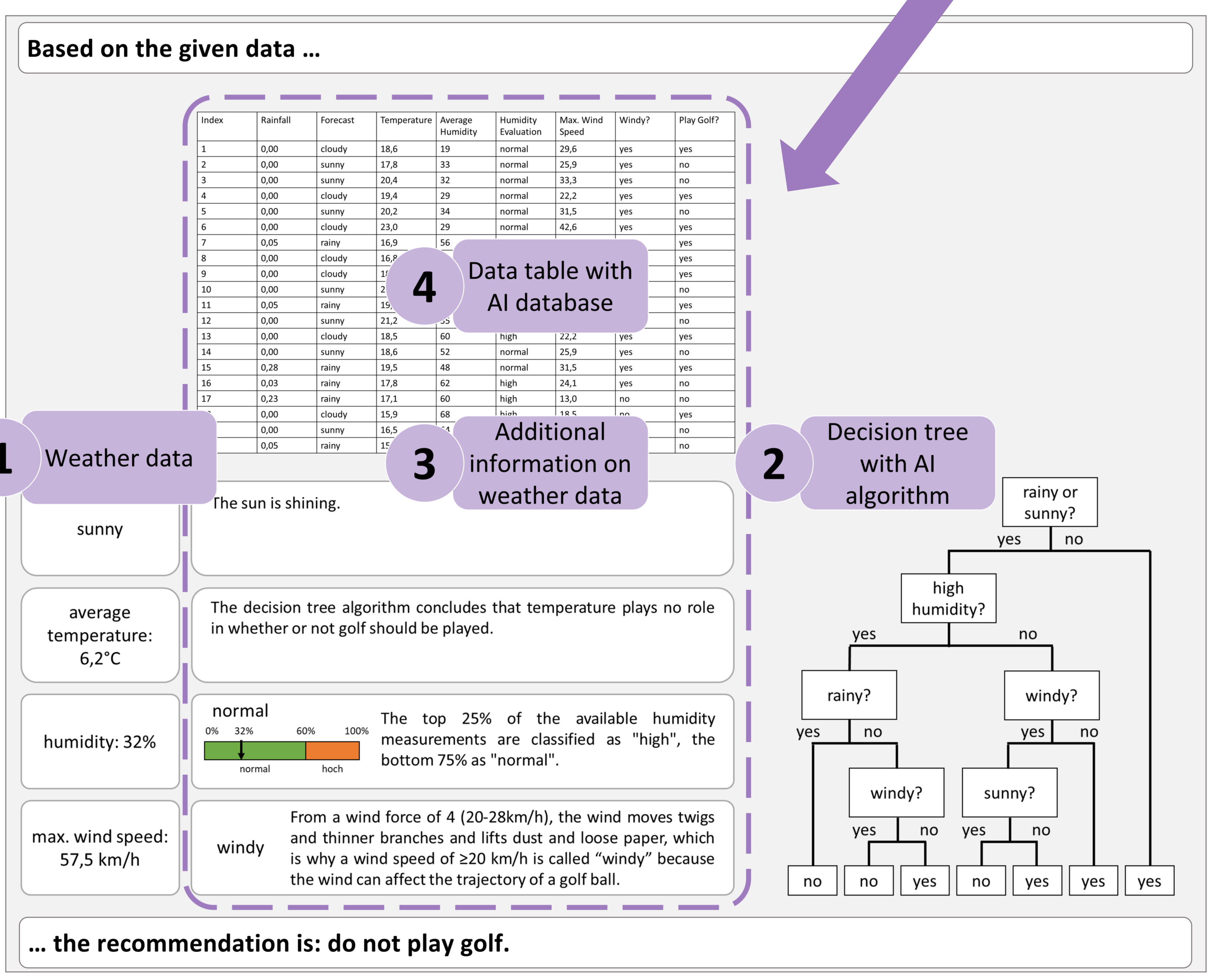
3-groups *between*-subjects design:

Experimental group 1 (E-): low elaborateness → 1 + 2

Experimental group 2 (E+): high elaborateness → 1 + 2 + 3 + 4

Experimental group 3 (EA): adaptable elaborateness → enable/disable 3 + 4

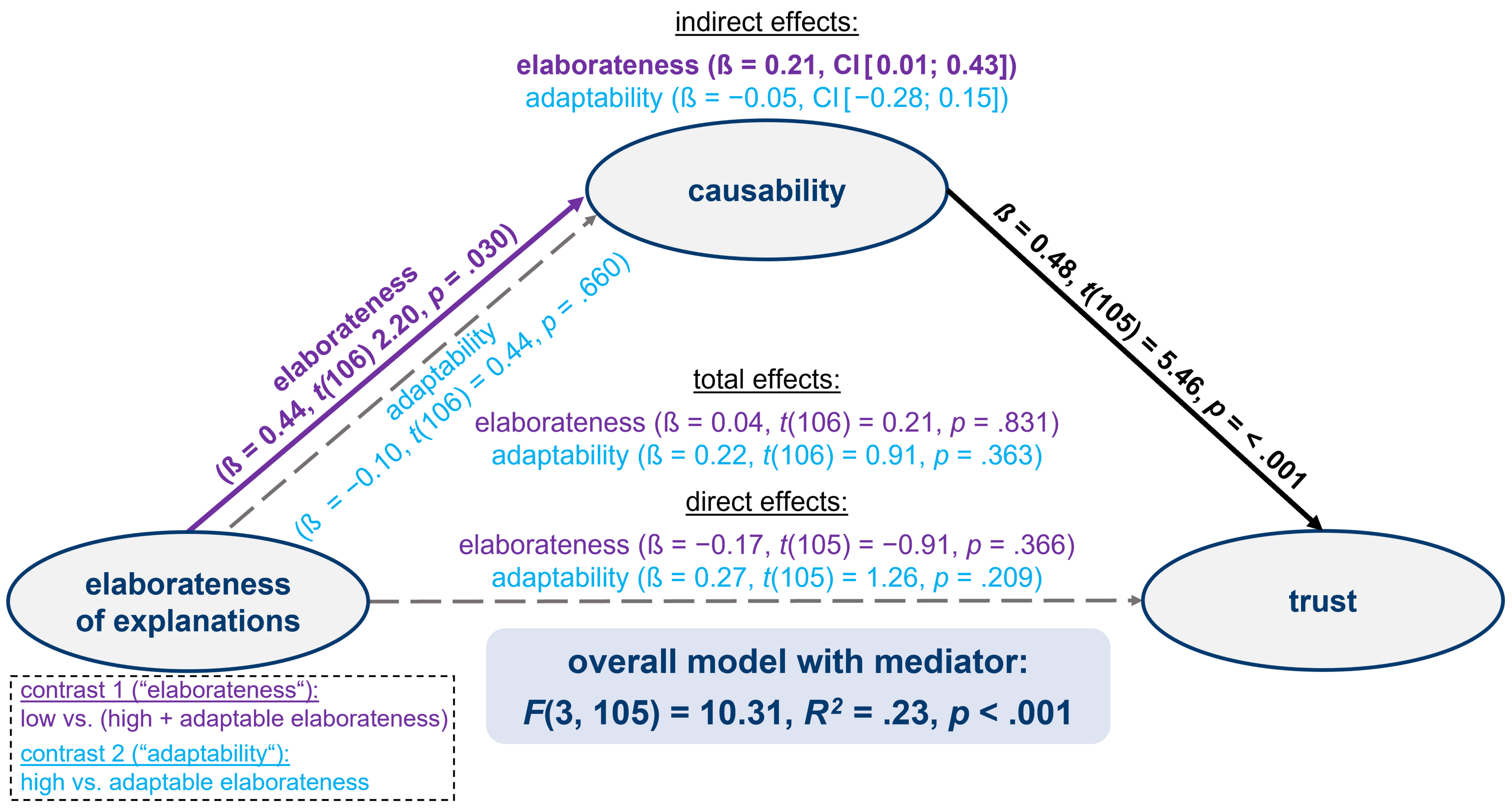
Scenario:



RESULTS

Mediation analyses (Helmert contrasts):

- Elaborateness and adaptability of explanations did not affect trust directly but indirectly via causability (R^2 = .23, p < .001).
- **Elaborateness** (contrast 1 significant)
  - higher elaborateness → higher causability → higher trust in AI system
- **Adaptability** (contrast 2 not significant)
  - adaptability → no further benefits regarding causability or trust



Explorative analyses on cognitive load:

- Significant negative Pearson correlations between cognitive load and causability ($-.71 \le r \le -.42$), and cognitive load and trust ($-.48 \le r \le -.42$).

	1 Causability	2 Trust	3 ECL	4 ICL
2 Trust	r = .46***	-		
3 ECL	r = −.71***	r = −.48***	-	
4 ICL	r = −.42***	r = −.42***	r = .66***	-
5 Overall CL	r = −.59***	r = −.43***	r = .89***	r = .85***

***p <.001. ECL = extrinsic cognitive load. ICL = intrinsic cognitive load.

CONCLUSION

General findings:

For enhancing **trust** in AI both system factors and human factors need to be considered:

- Higher **elaborateness** of explanations **increases trust** when users perceive these explanations as appropriate to understand the output of AI systems (**causability**).
- **Adaptability** provides **no further benefits** regarding causability or trust in the AI system.
- High **cognitive load** is associated with **lower causability** as well as **lower trust** in AI.

Implications:

The adoption of a **human-centric approach** is crucial for research and practice:

- From a cognitive and educational perspective, (x)AI research can gain insights on how to develop human-interpretable explanations.
- Engaging **users** into the design process and gather their **feedback** helps to tailor explanations to their needs and preferences.

Outlook:

- Trust, explanations, and actual understanding: How does **causability** affect **actual understanding** of AI systems?
- User expertise and experience with AI: How does **prior knowledge** influence the effectiveness of explanations in building trust in and understanding of AI? (novices vs. experts, see expertise reversal effect, Kalyuga, 2007).
- How does **cognitive load** impact **causability** and **trust** in AI systems? The causal relationship remains unclear, though research suggests a potential link to mistrust (Samson & Kostyszyn, 2015).

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052 **Holzinger, A., Carrington, A., & Müller, H. (2020).** Measuring the quality of explanations: The system Causability Scale (SCS). *KI - Künstliche Intelligenz, 34*(2), 193–198. https://doi.org/10.1007/s13218-020-00636-z **Jian, J.-Y., Bisantz, A. M., Drury, C. G., & Llinas, J. (2000).** Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04 **Kalyuga, S. (2007)**. Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509–539. https://doi.org/10.1007/s10648-007-9054-3 **Klepsch, M., Schmitz, F., & Seufert, T. (2017).** Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology, 8*(1), 1–18. https://doi.org/10.3389/fpsyg.2017.01997 **Molina, M. D., & Sundar, S. S. (2022).** When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication, 27*(4), 1–12. https://doi.org/10.1093/jcmc/zmac010 **Putnam, V., & Conati, C. (2019).** Exploring the need for Explainable Artificial Intelligence (XAI) in Intelligent Tutoring Systems (ITS). In C. Trattner, D. Parra, & N. Riche (Eds.), *Joint Proceedings of the ACM IUI 2019 Workshops* (Vol. 2327). http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-19.pdf **Samson, K., & Kostyszyn, P. (2015)**. Effects of cognitive load on trusting behavior – An experiment using the trust game. *PLOS ONE, 10*(5), e0127680. https://doi.org/10.1371/journal.pone.0127680 **Sweller, J. (2010).** Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*(2), 123–138. https://doi.org/10.1007/s10648-010-9128-5

University of Duisburg-Essen (Germany)
Research Methods in Psychology – Media-based Knowledge Construction
Department of Computer Science and Applied Cognitive Science
Faculty of Engineering
www.uni-due.de/psychmeth
Lotharstr. 65, D - 47057 Duisburg, Germany
+49 203 379-4143
Lydia Harbarth (lydia.harbarth@uni-due.de)