EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Science

Psychology Department · Social Cognition and Decision Sciences

Advice Taking from Interactive, Self-Explanatory Generative AI

Tobias R. Rebholz, Alena Koop, & Mandy Hütter
Psychology Department, University of Tübingen

Background

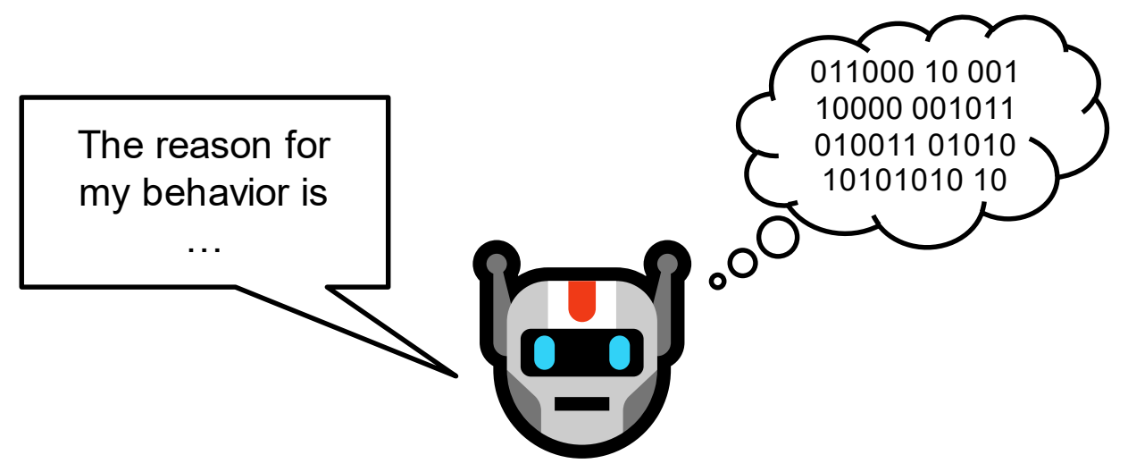
Algorithm Aversion

Individuals' general **preference** for interacting with humans rather than algorithms (Mahmud et al., 2022)

- **Versus:** Natural conversational style of contemporary chatbots (e.g., OpenAI's ChatGPT, Google's Gemini)

Explanation

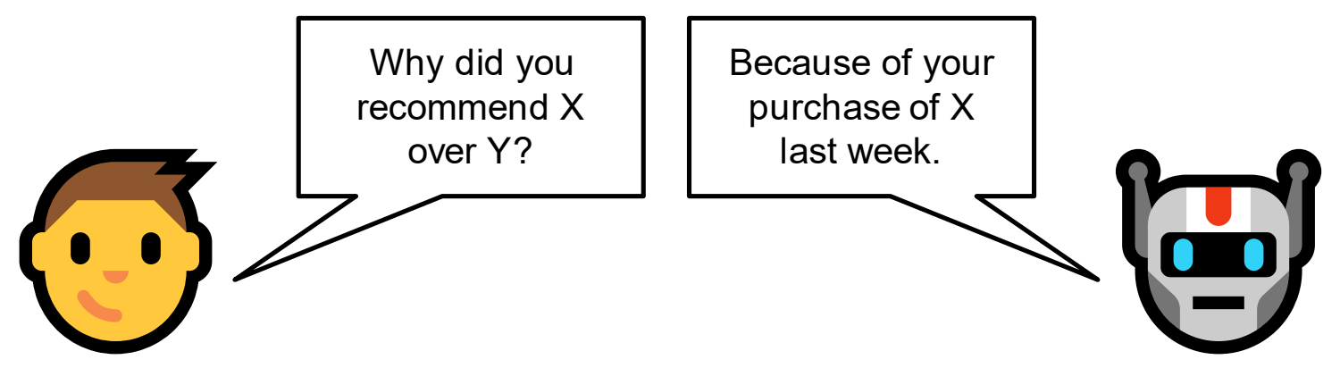
Increases **transparency** of the algorithmic judgment and decision-making process (Papamichail, 2003; van Dongen & van Maanen, 2013)



- Advice from explanatory algorithms is **weighted more strongly** (Gönül et al., 2006; Goodwin et al., 2013)

Interactivity

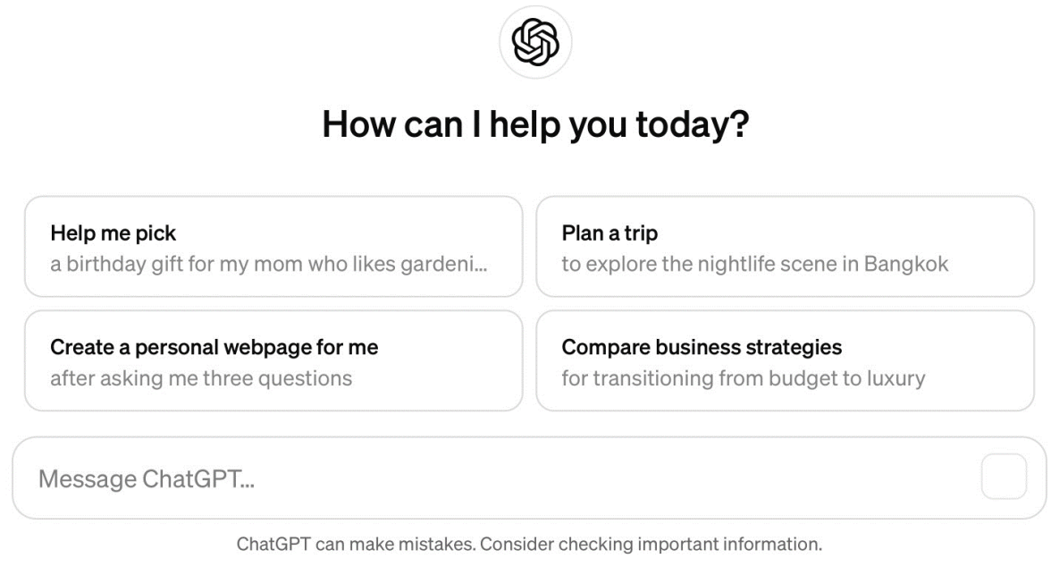
Enhances **trust calibration** and satisfies users' **desire for control** (e.g., Westphal et al., 2023)


Why did you recommend X over Y?
Because of your purchase of X last week.

- More control over the behavior of an algorithm **increases** users' willingness to rely on its output (Dietvorst et al., 2018; van Dongen & van Maanen, 2013)

Conversational User Interfaces

By providing information **upon request**, parties can reduce **informational asymmetry** (van Dongen & van Maanen, 2013)



How can I help you today?

Help me pick
a birthday gift for my mom who likes gardeni...

Plan a trip
to explore the nightlife scene in Bangkok

Create a personal webpage for me
after asking me three questions

Compare business strategies
for transitioning from budget to luxury

Message ChatGPT...

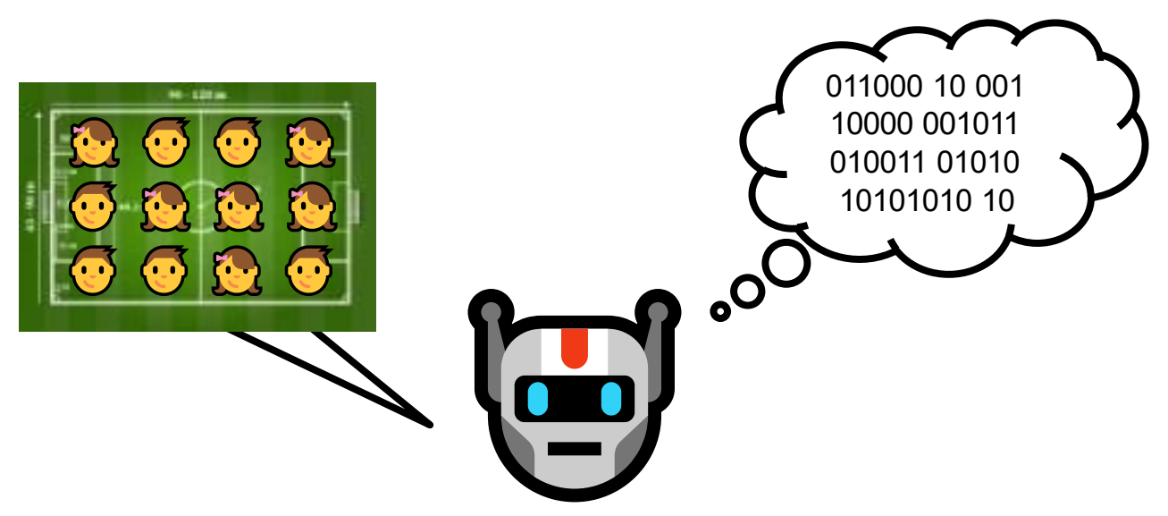
ChatGPT can make mistakes. Consider checking important information.

- Greater **salience** of influencing algorithms' behavior for actively requested than passively provided explanations
- No trust building through explanation if the opportunity to interact is **not used** to solicit an explanation (Goodwin et al., 2013)

Future Research

Multimodal Reasoning

Visual explanations improve users' objective **understanding** of complex algorithms (Cheng et al., 2019)



Method

Design

2 (**explanation**: provided vs. not provided) × 2 (**interactivity**: possible vs. not possible)

Participants

N = 472 university students (313 female, 154 male)

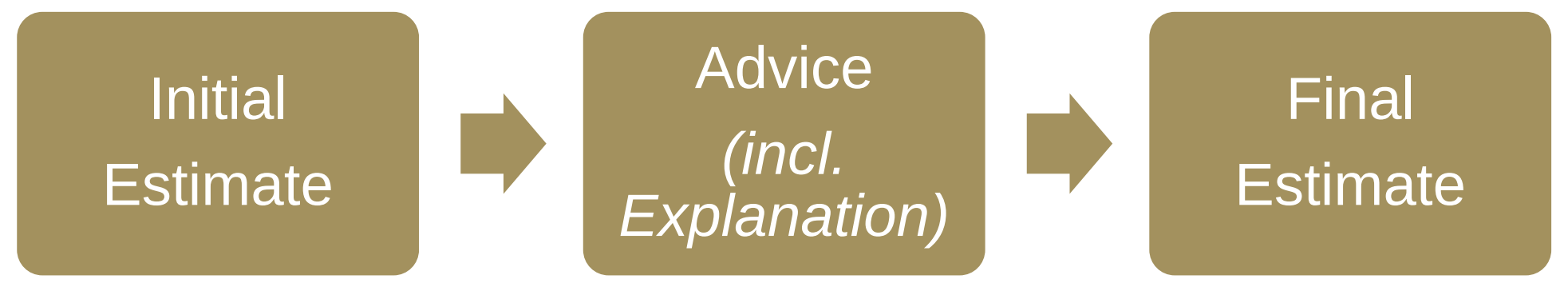
Material

10 Fermi problems (i.e., numerical judgment tasks)

- **Example:** How many soccer pitches would it take to accommodate all the inhabitants of Germany?

Procedure

Extended Judge-Advisor System (JAS) (Sniezek & Buckley, 1995)



Initial Estimate → Advice *(incl. Explanation)* → Final Estimate

- **Advice:** Pre-generated output from ChatGPT (gpt-3.5-turbo model with temperature = 0)

Dependent Variable

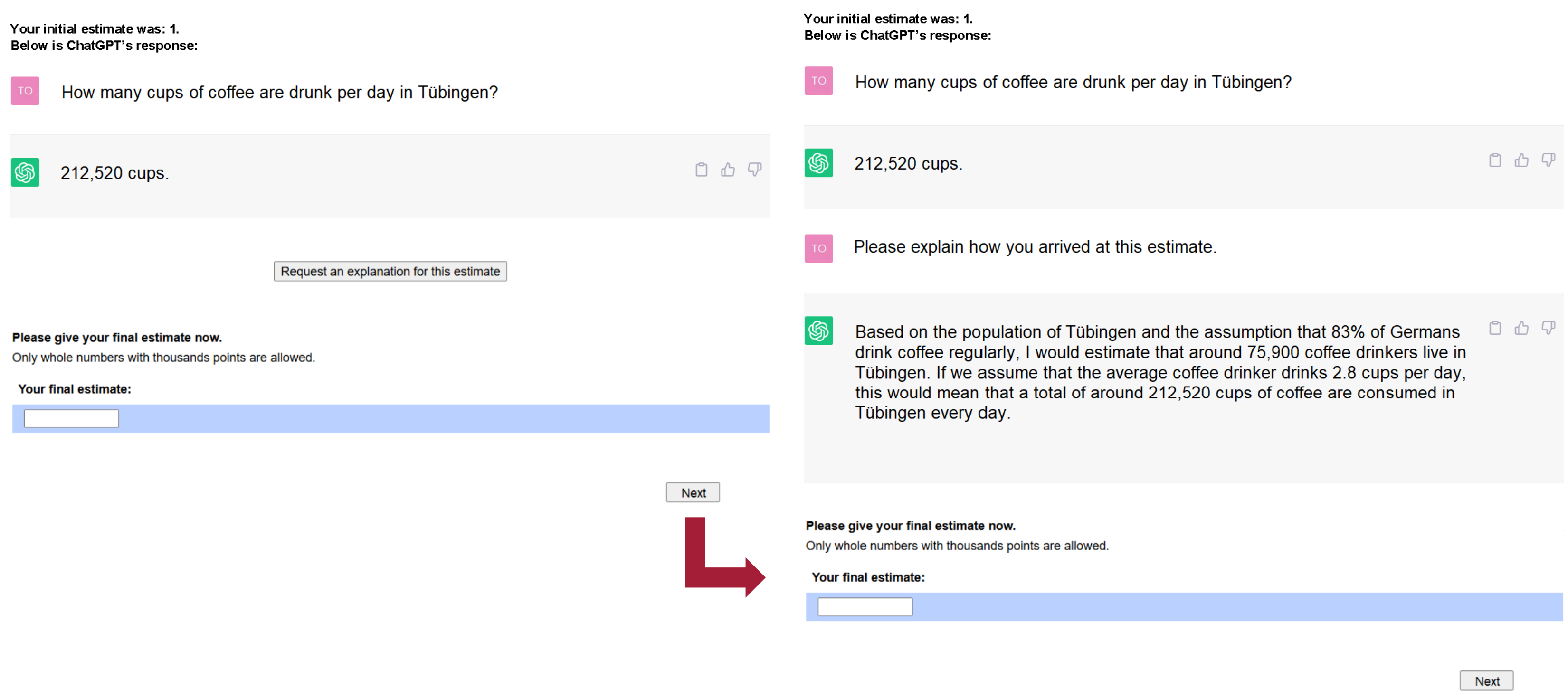
Weight of Advice (WOA) (Harvey & Fischer, 1997)

$$\text{WOA} = \frac{\text{Final Estimate} - \text{Initial Estimate}}{\text{Advice} - \text{Initial Estimate}}$$

Limitations

Interactivity

Limited to **pressing a button** for more information



Your initial estimate was: 1.
Below is ChatGPT's response:

TO How many cups of coffee are drunk per day in Tübingen?

212,520 cups.

Request an explanation for this estimate

Please give your final estimate now.
Only whole numbers with thousands points are allowed.

Your final estimate:

Next

Your initial estimate was: 1.
Below is ChatGPT's response:

TO How many cups of coffee are drunk per day in Tübingen?

212,520 cups.

TO Please explain how you arrived at this estimate.

Based on the population of Tübingen and the assumption that 83% of Germans drink coffee regularly, I would estimate that around 75,900 coffee drinkers live in Tübingen. If we assume that the average coffee drinker drinks 2.8 cups per day, this would mean that a total of around 212,520 cups of coffee are consumed in Tübingen every day.

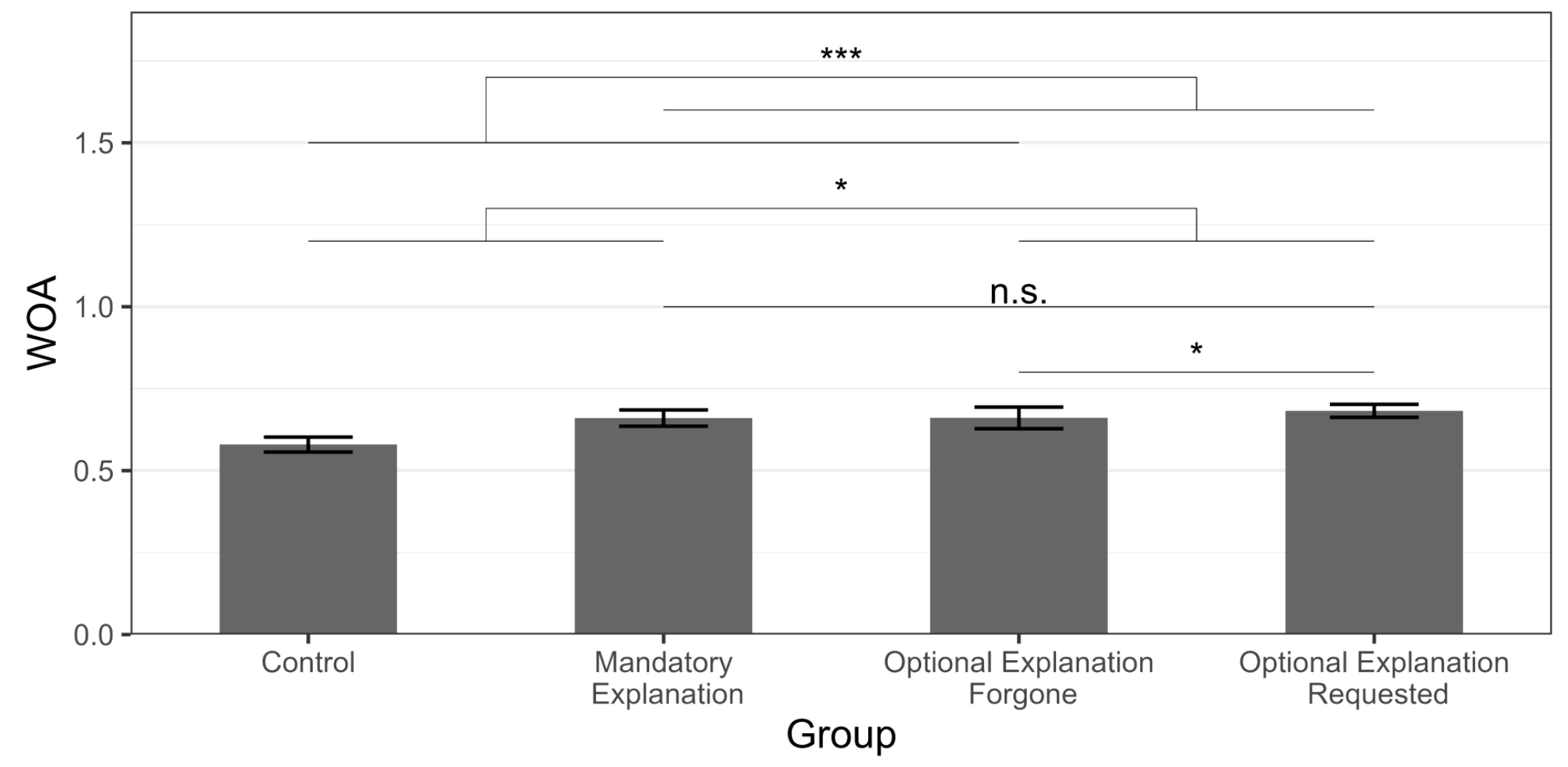
Please give your final estimate now.
Only whole numbers with thousands points are allowed.

Your final estimate:

Next

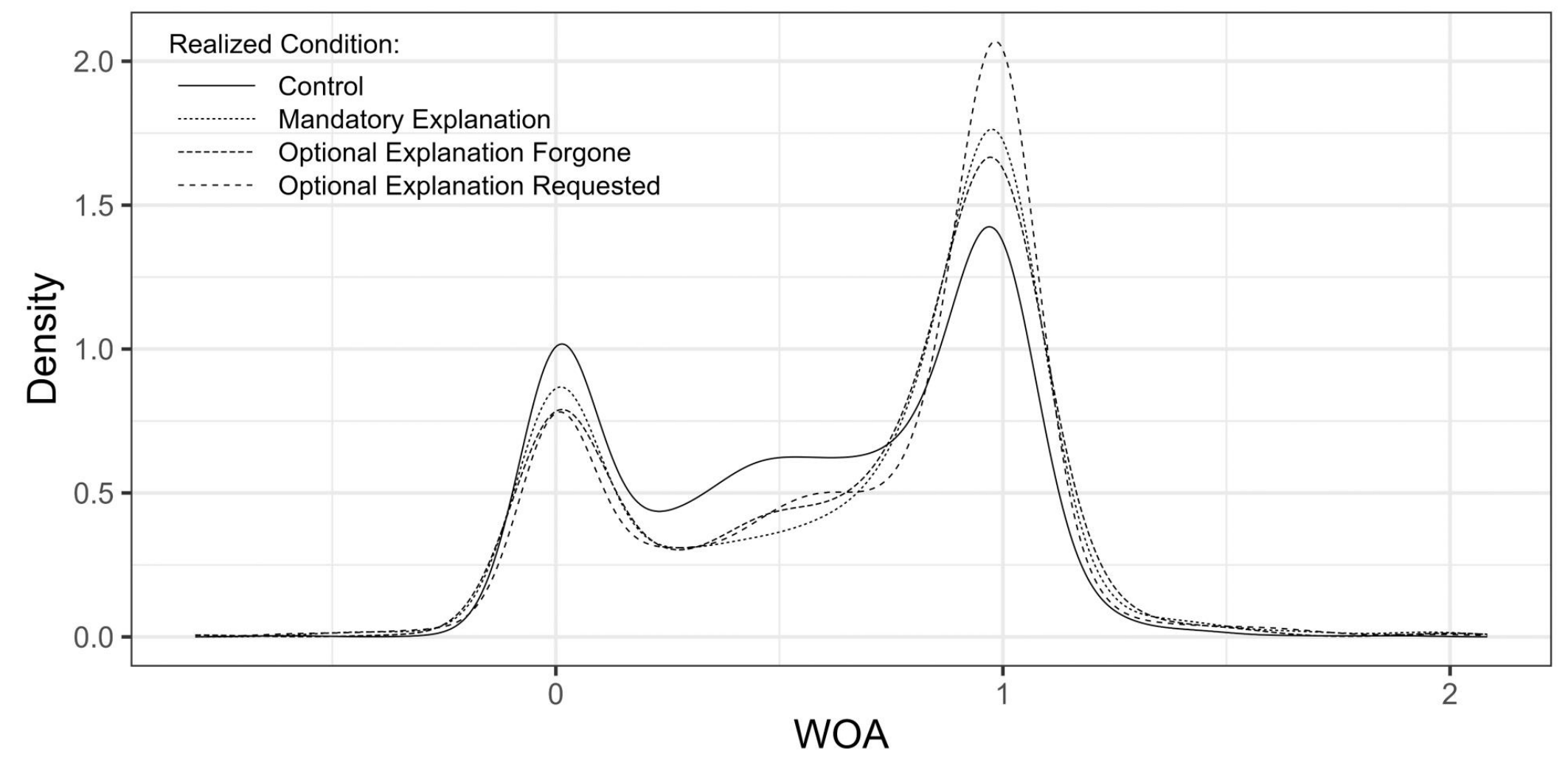
Results

Mean WOA per Group



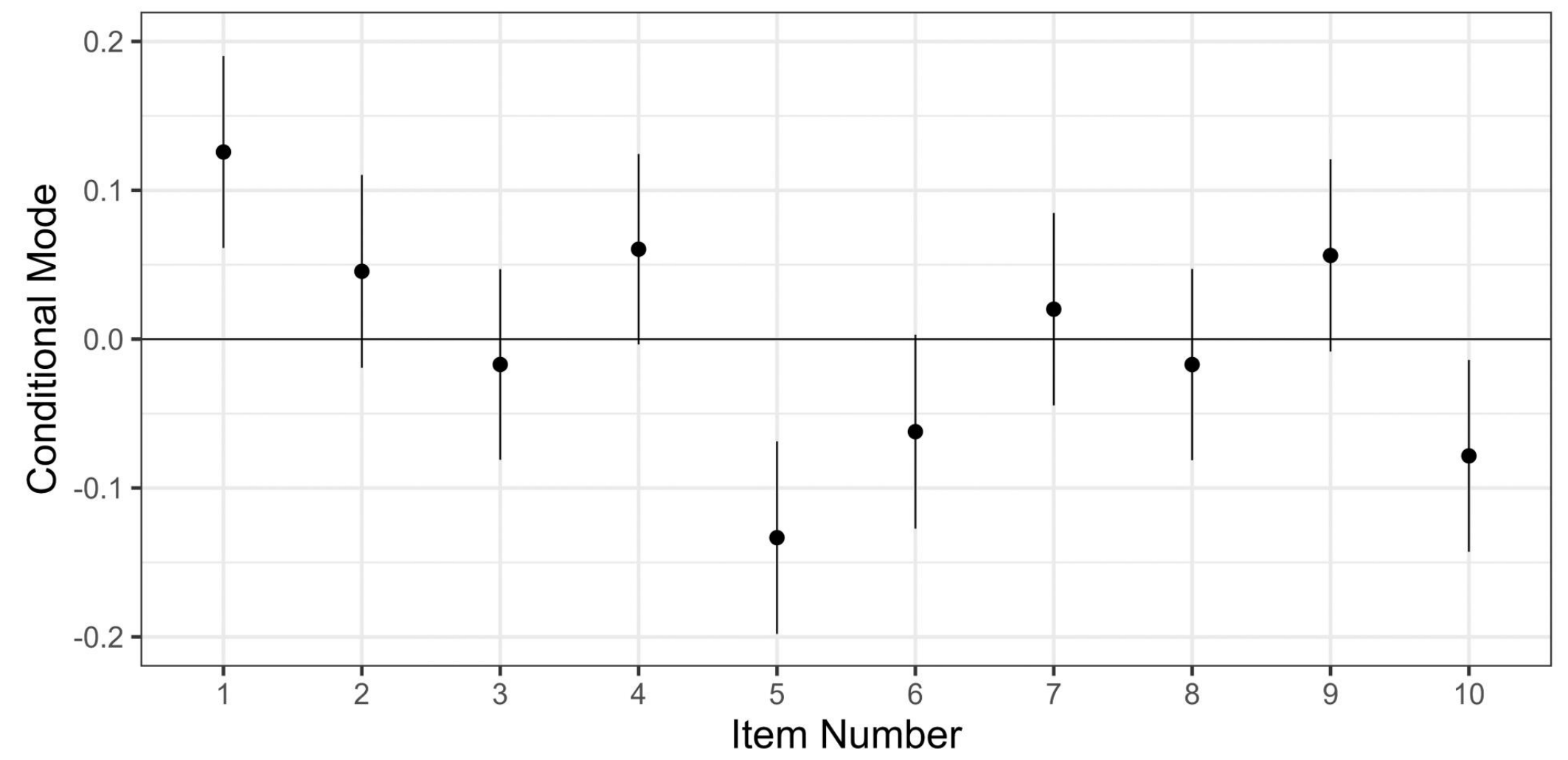
Note. Error bars show the 95% CI. Outliers of WOA are excluded based on Tukey's (1977) fences. * p < .05, ** p < .01, *** p < .001, two-sided. Adapted from Rebholz et al. (2024). CC BY 4.0.

Distribution of WOA per Group



Note. Gaussian kernel density plots with the bandwidth chosen according to Silverman's (1986) rule of thumb. Outliers of WOA are excluded based on Tukey's (1977) fences. From Rebholz et al. (2024). CC BY 4.0.

Deviations from Mean WOA per Item

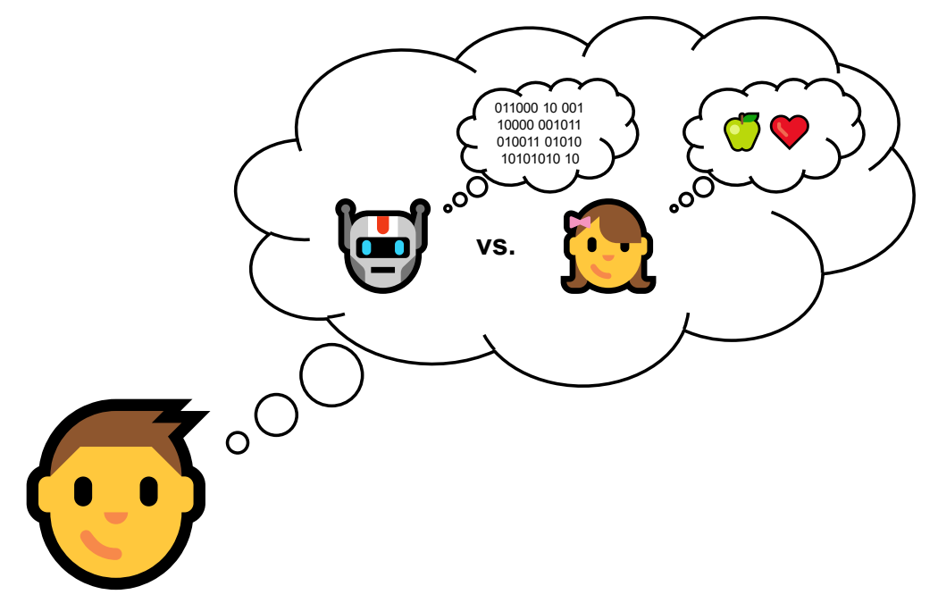


Note. Error bars show the 95% *CI*. The underlying extended multilevel model includes participant and item random slopes of explanation. From Rebholz et al. (2024). CC BY 4.0.

Implications

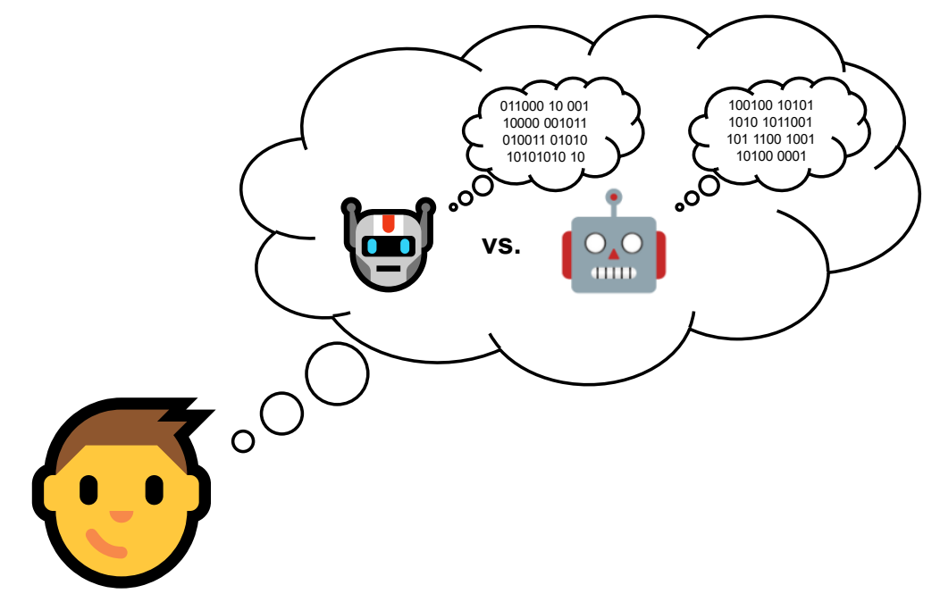
Theory of Machine

(Logg, 2022)



Theory of Machine 2.0

(Rebholz, 2024)



References:

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300789

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Gönül, M. S., Önkal, D., & Lawrence, M. (2006). The effects of structural characteristics of explanations on use of a DSS. *Decision Support Systems*, *42*(3), 1481–1493. https://doi.org/10.1016/j.dss.2005.12.003

Goodwin, P., Sinan Gönül, M., & Önkal, D. (2013). Antecedents and effects of trust in forecasting advice. *International Journal of Forecasting*, *29*(2), 354–366. https://doi.org/10.1016/j.ijforecast.2012.08.001

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, *70*(2), 117–133. https://doi.org/10.1006/obhd.1997.2697

Logg, J. M. (2022). The psychology of big data: Developing a "Theory of Machine" to examine perceptions of algorithms. In S. C. Matz (Ed.*), The psychology of technology: Social science research in the age of big data* (pp. 349–378). American Psychological Association. https://doi.org/10.1037/0000290-011

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, *175*, 121390. https://doi.org/10.1016/j.techfore.2021.121390

Papamichail, K. (2003). Explaining and justifying the advice of a decision support system: A natural language generation approach. *Expert Systems with Applications*, *24*(1), 35–48. https://doi.org/10.1016/S0957-4174(02)00081-7

Rebholz, T. R. (2024). *Theory of Machine 2.0: Artificial versus artificial intelligence*. PsyArXiv. https://doi.org/10.31234/osf.io/ekz9a

Silverman, Bernard. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall.

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174. https://doi.org/10.1006/obhd.1995.1040

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

van Dongen, K., & van Maanen, P.-P. (2013). A framework for explaining reliance on decision aids. *International Journal of Human-Computer Studies*, *71*(4), 410–424. https://doi.org/10.1016/j.ijhcs.2012.10.018

Westphal, M., Vössing, M., Satzger, G., Yom-Tov, G. B., & Rafaeli, A. (2023). Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance. *Computers in Human Behavior*, *144*, 107714. https://doi.org/10.1016/j.chb.2023.107714

STATISTICAL MODELING in PSYCHOLOGY
FREIBURG HEIDELBERG LANDAU MANNHEIM TÜBINGEN

PsyArXiv