Initial evidence for biased decision-making despite human-centered AI explanations

Nicolas Scharowski
Center for Cognitive Psychology and Methodology, University of Basel
Basel, CH-4055, Switzerland
nicolas.scharowski@unibas.ch

Klaus Opwis
Center for Cognitive Psychology and Methodology, University of Basel
Basel, CH-4055, Switzerland
klaus.opwis@unibas.ch

Florian Brühlmann
Center for Cognitive Psychology and Methodology, University of Basel
Basel, CH-4055, Switzerland
florian.bruehlmann@unibas.ch

CHI'21, May 8–13, 2021, Yokohama, Japan
ACM 000-0-0000-0000-0/00/00.

Abstract

In explainable artificial intelligence (XAI) research, explainability is widely regarded as crucial for user trust in artificial intelligence (AI). However, empirical investigations of this assumption are still lacking. There are several proposals as to how explainability might be achieved and it is an ongoing debate what ramifications explanations actually have on humans. In our work-in-progress we explored two post-hoc explanation approaches presented in natural language as a means for explainable AI. We examined the effects of human-centered explanations on trust behavior in a financial decision-making experiment (N = 387), captured by weight of advice (WOA). Results showed that AI explanations lead to higher trust behavior if participants were advised to *decrease* an initial price estimate. However, explanations had no effect if the AI recommended to *increase* the initial price estimate. We argue that these differences in trust behavior may be caused by cognitive biases and heuristics that people retain in their decision-making processes involving AI. So far, XAI has primarily focused on biased data and prejudice due to incorrect assumptions in the machine learning process. The implications of potential biases and heuristics that humans exhibit when being presented an explanation by AI have received little attention in the current XAI debate. Both researchers and practitioners need to be aware of such human biases and heuristics in order to develop truly human-centered AI.

Author Keywords
XAI; HCXAI; AI Trust; AI Transparency

Introduction
Recent breakthroughs in artificial intelligence (AI) have led to it being increasingly used in a variety of everyday applications such as video surveillance, autonomous driving, online customer support and product recommendations. Because of this general applicability and potential manifold consequences, voices are being raised that AI should satisfy criteria like fairness, reliability, accountability and transparency [6]. What can happen if AI is not built around those criteria is illustrated by the example of Amazon's AI based recruiting tool [1]. It learnt to prefer male applicants over female applicants. The reason for this was not that gender per se was used as a feature, but that the AI discovered a pattern in which applications that used more female related terms were less likely to be hired. The AI penalized CV's including the term "women" because such words were seen less in applications that had been selected in the past. Thus, the data used to train the AI was biased, meaning that already existing human biases (i.e. selecting fewer applications from women) were carried over to the AI. Due to examples like this, there have been extensive debates about how to prevent biased AI in the effort to develop human-centered AI [7]. This call for human-centered AI has also led to the multidisciplinary research field of explainable artificial intelligence (XAI). Among other things XAI explores methods that make the predictions or decisions of AI transparent to humans and aims to give meaningful information by explaining how a specific output was reached, thus making opaque AI models comprehensible to humans. However, there are still few empirical studies that evaluate the impact of explainability on factors like trust in the human-AI collaboration, especially for end-users. Consequently, there is limited understanding of the effect of AI explanations on people and many research opportunities are yet to be explored. We argue that the challenge of biases in human-centered explanations may not have been adequately recognized and articulated. If the results of our preliminary study prove to be robust and withstand further investigation, the problem of bias in AI may reach further than previously thought, extending from the aforementioned problem of biased training data into the realm of Human-Centered Explainable AI (HCXAI).

Challenges of human-centered explanations
If people assign human-like traits to artificial agents, they might as well expect explanations from them that are similar to the way in which humans explain their actions [2]. For this reason, researchers have emphasized the importance of incorporating insights from philosophy, social science and psychology into the field of XAI because of their research on how people define, generate, select, evaluate and present explanations [5]. Miller argues that not all explanations are equal, and that some are more valuable for humans than others. He defined certain criteria of what contributes to a meaningful explanation for humans like *selectivity, contrastivity, causality* and *sociality*. This focus on how humans explain decisions to each other is a good start in the endeavour of human-centered AI. The underlying challenge to this approach, however, seems to be that AIs are logic-based systems, whereas we humans are not purely rational agents. People's decision-making processes involve cognitive biases and heuristics, meaning systematic thinking errors and mental shortcuts, which occur when humans process and interpret information. This frequently leads to irrational decisions and non-optimal choices [4]. AI generated explanations must account for such potential cognitive biases and heuristics in order to support humans to make better informed decision and to gain trust. Some researchers argue that humans develop trust solely on the

Sidebar 1: Examples of the different explanation approaches

Control:
"Your guess was $1,000 / month. The AI recommends $1,250 / month."

Feature importance:
"Your guess was $1,000 / month. The AI recommends $1,250 / month. *Next to the main features (size, bedrooms, bathrooms), the second most important reason for this price recommendation was the fact that the apartment has a fitness center.*"

Counterfactual:
"Your guess was $1,000 / month. The AI recommends $1,250 / month. Next to the main features (size, bedrooms, bathrooms), the second most important reason for this price recommendation was the fact that the apartment has a fitness center. *If the apartment did not have a fitness center, the price $1,000 / month would have been recommended.*"

basis of the AI's performance, i.e. its accuracy in a given task over a period of time. We counter that only focusing on this computational aspect without being truly human-centered could lead to what we refer to as a "Cassandra AI", a scenario, inspired by the Trojan priestess of Apollo in Greek mythology. Cassandra made true prophecies but was never believed. Now imagine an AI that always makes accurate predictions but is never trusted. From a computational standpoint, such an AI would be error-free. However, by neglecting the human factor of the collaboration, this quasi optimal AI still would not achieve its objective of helping people to make better decisions, simply because the AI explanations are not tailored to convince people. The existence of cognitive biases and heuristics implies that humans prefer certain types of AI explanations over others, and thus explanations for the most accurate predictions might not be trusted if alternative explanations are available that better fit into existing cognitive schemata. While Miller was the first to emphasize the importance of cognitive biases and heuristics for XAI, we present empirical results of a work-in-progress that suggest the actual existence of such thinking errors and mental shortcuts in a human decision-making task, involving AI.

Initial evidence for biased decision making

We conducted a financial decision-making experiment on Amazon Mechanical Turk (MTurk) with the explainability techniques feature importance (n = 146) and counterfactuals (n = 108) to empirically compare these two with a control condition (n = 133). Participants were asked to imagine a scenario where their goal was to sublease six different apartments on a subleasing website. Based on the features and amenities of the apartment (e.g. number of bedrooms, distance to public transit, etc.), they had to guess an initial subleasing price (T1). After guessing T1, an alleged AI from the website provided a computed price recommendation. In reality, however, a price recommendation based on basic arithmetic with a random number, rather than an actual AI, was given. This random number varied between 10 and 20, meaning that each participant saw a different price recommendation for a given apartment. Apartments were presented in a random order. For three of the six apartments, the recommendation was lower than the initially guessed subleasing price (e.g., if T1 was 1,000 and the random number 20, the AI recommendation was 800) and for the other three apartments higher than the initially guessed subleasing price. After seeing the price recommendation and the accompanied explanation, participants could decide if they wanted to approach the AI recommendation or not, settling for a final subleasing price (T2). Sidebar 1 shows how explanations were presented to the participants. The features and amenities used to form the explanation was different for each apartment.

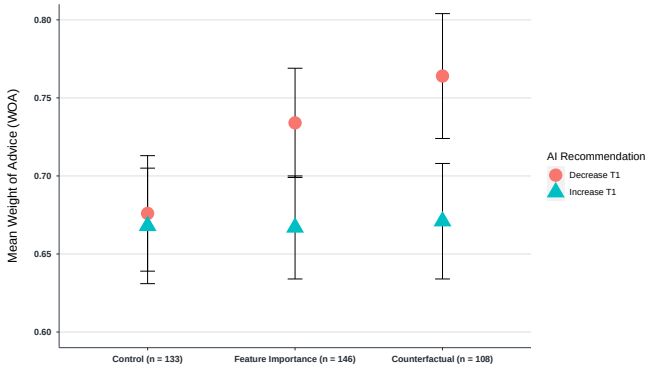

Figure 1: Interaction plot, capturing the interaction between the effects of conditions and recommendations, that were either higher or lower than the initial price guess (T1). Note that the y-axis is scaled to better visualize the effect. The error bars depict 95% confidence intervals.

For trust behavior, the metric weight of advice (WOA) from the advice-taking literature was used [3]. WOA measures the degree to which people update their beliefs and quantifies how much people weigh the received advice (i.e. the AI recommendation). Since WOA was a repeated measure for each apartment, a mixed-effects model was used with participants as the random within-subject factor and conditions as the fixed between-subject factor. Our results show that regardless of the different conditions, participants on average displayed high AI trust behavior with an overall WOA of 0.7. (M = 0.69, SD = 0.36). A WOA of 0.70 implies that participants adopted 70% of the AI recommendations when updating their prior beliefs to form T2. The effect of explanations, however, depended on the nature of the decision that participants had to make. When the AI recommended to *increase* the initial guess (T1), explanation techniques had no effect on WOA. Contrarily, if the AI recommended to *decrease* the price, there was a significant effect of explanations on WOA (β = 0.07, t = 2.17, CI = [0,02, 0.13], p = .03). In the latter, participants in the experimental condition updated their initial guesses and approached the AI recommendations up to 9% more than participants in the control group (See Figure 1). This finding seems counterintuitive at first glance, since one might expect that participants would always choose to embrace the prospect of obtaining a higher subleasing price. We argue, however, that the two types of recommendations can be thought of as two distinct decision-making processes. The well-studied concept of loss aversion [9] could account for this discrepancy and serve as an explanation attempt for these findings. When participants were advised by the AI to increase their initial guess, it is likely that they were concerned that this potential price raise would cause an unsuccessful sublease. The prospect of getting more money (gain) mattered less in this decision-making process than the possibility of not being able to sublease at all (loss). When faced with loss aversion, the explanations from the pseudo AI seems not to be convincing enough to overcome the participants' higher assigned utility to losses. When *not* being faced with loss aversion, human-centered AI explanations seem to convince people to adjust their initial sublease price, compared to the control where no additional explanation was present. As of now, the interpretation under consideration of loss aversion is tentative and we work on replicating our findings with a more elaborate research design.

Discussion

These preliminary results suggest that increased trust behavior through human-centered post-hoc explanations occurs only in certain decision-making processes. Humans may exhibit cognitive biases and apply heuristics when exposed to AI explanations. In a simple subleasing task that potentially induced loss aversion, feature importance and counterfactuals did not appear to persuade participants to change their behavior and demonstrate increased trust. It is possible that inherent biases and heuristics are so hardwired that AI explanations are not convincing enough to disprove non-optimal human decision-making. If that is the case, AI may not help us to reach better decisions in circumstances where human intuition becomes too tempting for our judgment. We suggest that the XAI community should account for potential biases and heuristics in order to design for truly human-centered explanations that help optimizing decision-making. Biased decisions will not simply disappear because AI is involved and while heuristics are useful in some situations, we may not want them to influence us in others. The better we understand biases, the more likely we are to overcome them. Future research should focus on different types of cognitive biases and heuristics that could potentially undermine AI explanations, such as loss aversion [9], framing [8] or confirmation bias [10]. If Cassandra had known why people did not believe

her, she could have addressed their doubts. By knowing that such irrational tendencies exist in humans, perhaps an AI could likewise address them and help us moderate these tendencies. We believe that the collaboration between humans and AI works best when the weaknesses of one party are balanced by the strengths of the other.

REFERENCES

[1] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. (2018). Retrieved February 11, 2021 from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[2] Maartje De Graaf and Bertram Malle. 2018. People's Judgments of Human and Robot Behaviors: A Robust Set of Behaviors and Some Discrepancies. *Companion of the International Conference on Human-Robot Interaction* (2018), 97–98. DOI: http://dx.doi.org/10.1145/3173386.3177051

[3] Nigel Harvey and Ilan Fischer. 1997. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes* 70, 2 (1997), 117–133. DOI: http://dx.doi.org/10.1006/obhd.1997.2697

[4] Daniel Kahneman. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, NY, USA.

[5] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. DOI: http://dx.doi.org/10.1016/j.artint.2018.07.007

[6] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016), 1–21. DOI: http://dx.doi.org/10.1177/2053951716679679

[7] Constantine Stephanidis, Gavriel Salvendy, Margherita Antona, Jessie Y. C. Chen, Jianming Dong, Vincent G. Duffy, Xiaowen Fang, Cali Fidopiastis, Gino Fragomeni, Limin Paul Fu, Yinni Guo, Don Harris, Andri Ioannou, Kyeong-ah Jeong, Shin'ichi Konomi, Heidi Krömker, Masaaki Kurosu, James R. Lewis, Aaron Marcus, Gabriele Meiselwitz, Abbas Moallem, Hirohiko Mori, Fiona Fui-Hoon Nah, Stavroula Ntoa, Pei-Luen Patrick Rau, Dylan Schmorrow, Keng Siau, Norbert Streitz, Wentao Wang, Sakae Yamamoto, Panayiotis Zaphiris, and Jia Zhou. 2019. Seven HCI Grand Challenges. *International Journal of Human–Computer Interaction* 35, 14 (2019), 1229–1269. DOI: http://dx.doi.org/10.1080/10447318.2019.1619259

[8] Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211, 4481 (1981), 453–458. DOI: http://dx.doi.org/10.1126/science.7455683

[9] Amos Tversky and Daniel Kahneman. 1991. Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics* 106, 4 (1991), 1039–1061. DOI: http://dx.doi.org/10.2307/2937956

[10] Peter C Wason. 1960. On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology* 12, 3 (1960), 129–140. DOI: http://dx.doi.org/10.1080/17470216008416717