Decoding AI’s Nudge: A Unified Framework to Predict Human Behavior in AI-assisted Decision Making (Supplementary Material)

Zhuoyan Li, Zhuoran Lu, Ming Yin
Purdue University, USA
li4178@purdue.edu, lu800@purdue.edu, mingyin@purdue.edu

Literature Review
We screened research papers related to AI-assisted decision making that are published between 2018 and 2021 in the ACM CHI Conference on Human Factors in Computing Systems (CHI), ACM Conference on Computer-supported Cooperative Work and Social Computing (CSCW), ACM Conference on Fairness, Accountability, and Transparency (FAccT), and ACM Conference on Intelligent User Interfaces (IUI) to identify different forms of AI assistance developed in the literature. We grouped different forms of AI assistance into a few categories:


2. **Delayed recommendation** (Zhang, Liao, and Bellamy 2020; Dodge et al. 2019; Wang and Yin 2021; Yin, Vaughan, and Wallach 2019; Buccinca, Malaya, and Gajos 2021; Lu and Yin 2021; Poursabzi-Sangdeh et al. 2018; Grgić-Hlača, Engel, and Gummadi 2019; Park et al. 2019)

3. **Explanation only** (Lai and Tan 2018; Alqarawi et al. 2020; Lucic, Haned, and de Rijke 2019; Rader, Cotter, and Cho 2018; van Berkel et al. 2021; Buccinca et al. 2020; Gajos and Mamykina 2022; Anik and Bunt 2021; Lucic, Haned, and de Rijke 2019; Rader, Cotter, and Cho 2018)

4. **Interaction between human and AI**: Different from the three “static” types of AI assistance, this form of AI assistance emphasizes the interaction between human decision maker (DM) and the AI assistant. For example, during the collaboration with AI, AI can provide the accuracy feedback to help DMs recalibrate their trust in AI (Bansal et al. 2020; Yu et al. 2019). In addition, DMs may actively explore the decision space of AI assistants (Cai et al. 2019a; Levy et al. 2021a), or they can be provided with interactive explanations to gain a deeper understanding of how AI models arrive at their decisions (Cai et al. 2019a; Yang et al. 2020; Smith-Renner et al. 2020; Liu, Lai, and Tan 2021; Cai et al. 2019b), thereby enhancing their appropriate trust in AI assistants.

Given the limited number of papers in the Interaction between Human and AI category, and their unique interaction designs, in this study, we focus on building computational framework to model the first three types of AI assistance influence human DMs.

Additional Details of Human-Subject Experiment

Data Validity Check. To verify the engagement of subjects in our study, an attention check question was included in which subjects were instructed to select a pre-specified option. Among the 285 workers participated in our study, 202 passed the attention check question. Only the data from them were considered as valid and used to train/evaluate our models. Also, as an evidence of “consistency”, across all decision making tasks, the average fraction of subjects who agreed with the majority decision on the task was 82% (though decision makers did not need to agree with others’ decisions).

Working Time. The mean completion times for a decision making task and their standard deviations in different treatments are: Independent: 4.61s ± 3.27s, Immediate assistance: 5.03s ± 3.42s, Delayed recommendation: 9.89s ± 6.07s, Explanation only: 5.45s ± 3.54s.

Ablation Study

In our approach, we adopt a probabilistic framework to learn a distribution of the independent human decision model $q_h(w_h)$. In this study, we conducted an ablation study by replacing the distribution of the decision model $q_h(w_h)$ with a deterministic logistic regression model that can be learned in the Delayed recommendation scenario (because human DMs need to first provide their initial decision before the AI
The Potential Influence of the LLM-Powered Decision Aids on Humans

The AI model we used in our study was a supervised learning model that was trained independently without human feedback. However, with the rapid development of large language models (LLMs), one may envision that future AI-based decision aids can be powered by LLMs. It is known that LLMs may learn from human feedback and may have the tendency to provide affirmative responses to humans, which could reinforce human DMs’ beliefs and biases in the long run. This could be particularly concerning if the DM is intentionally providing feedback to LLMs in a way that seeks approval for a decision that is flawed or biased. As the LLM keeps internalizing the human DM’s biases through their feedback and learns to provide affirmative response to DMs, the DM might perceive the AI’s affirmative response as an endorsement from an expert, leading to an increased likelihood of confirmation bias. Moreover, the consistent affirmative feedback from LLMs could subtly alter the human cognitive decision making process. For example, if LLMs continually affirm DMs’ decisions or ideas, it may lead to DMs’ overconfidence in their decisions. Developing computational frameworks to characterize the dynamics between the influence of AI assistance to human DMs and the influence of human DMs’ feedback to AI assistance for future AI-based decision aids that are powered by LLMs can be a very interesting future direction.

References


Desmond, M.; Ashktorab, Z.; Brachman, M.; Brimijoin, K.; Duesterwald, E.; Dugan, C.; Finegan-Dollak, C.; Muller,
M. J.; Joshi, N. N.; Pan, Q.; and Sharma, A. 2021. Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface. *26th International Conference on Intelligent User Interfaces*.


