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:

1. *Immediate assistance* (Lai and Tan 2018; Liu, Lai, and Tan 2021; Nourani et al. 2021; Green and Chen 2019b; Tsai et al. 2021; Bansal et al. 2020; Buccinca, Malaya, and Gajos 2021; Feng and Boyd-Graber 2018; Guo et al. 2019; Lee et al. 2020, 2021; Levy et al. 2021b; Cheng et al. 2019; Lai, Liu, and Tan 2020; Poursabzi-Sangdeh et al. 2018; Chromik et al. 2021; Jacobs et al. 2021; Smith-Renner et al. 2020; Desmond et al. 2021; Buccinca et al. 2020; Gajos and Mamykina 2022; Gomez et al. 2020; Abdul et al. 2020; Brown et al. 2019; Cai, Jongejan, and Holbrook 2019; Buccinca et al. 2020; Szymanski, Millecamp, and Verbert 2021; Green and Chen 2019a; De-Arteaga, Fogliato, and Chouldechova 2020; Yang et al. 2020; Kunkel et al. 2019; Das and Chernova 2020; Yu et al. 2019; Lee et al. 2019; Harrison et al. 2020; Kocielnik, Amershi, and Bennett 2019)
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; Alqaraawi 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 how 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 $\pm$ 3.27s, *Immediate assistance*: 5.03s $\pm$ 3.42s, *Delayed recommendation*: 9.89s $\pm$ 6.07s, *Explanation only*: 5.45s $\pm$ 3.54s.

Ablation Study

In our approach, we adopt a probabilistic framework to learn a distribution of the independent human decision model $q_\phi(\boldsymbol{w}_h)$. In this study, we conducted an ablation study by replacing the distribution of the decision model $q_\phi(\boldsymbol{w}_h)$ 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

Number of Training Instances	5	10	15	20	25
Deterministic Decision Model	0.514	0.469	0.454	0.434	0.416
Ours	0.430	0.422	0.413	0.402	0.394

Table 1: Comparing the performance of our method against an alternative that substitutes the distribution of decision model $q_\phi(\boldsymbol{w}_h)$ of our method with a deterministic logistic regression model in the *Delayed recommendation* scenario. NLL is adopted as the evaluation metric, with a lower NLL denoting superior performance.

recommendation is revealed). As shown in Table 1, we observed that our approach consistently outperforms the counterpart using the deterministic decision model as we vary the number of training instances.

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

Abdul, A.; von der Weth, C.; Kankanhalli, M. S.; and Lim, B. Y. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Alqaraawi, A.; Schuessler, M.; Weiß, P.; Costanza, E.; and Bianchi-Berthouze, N. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. *Proceedings of the 25th International Conference on Intelligent User Interfaces*.

Anik, A. I.; and Bunt, A. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Bansal, G.; Wu, T. S.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. S. 2020. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Brown, A.; Chouldechova, A.; Putnam-Hornstein, E.; Tobin, A.; and Vaithianathan, R. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Buccinca, Z.; Lin, P.; Gajos, K. Z.; and Glassman, E. L. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces*.

Buccinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To Trust or to Think. *Proceedings of the ACM on Human-Computer Interaction*, 5: 1 – 21.

Cai, C. J.; Jongejan, J.; and Holbrook, J. 2019. The effects of example-based explanations in a machine learning interface. *Proceedings of the 24th International Conference on Intelligent User Interfaces*.

Cai, C. J.; Reif, E.; Hegde, N.; Hipp, J.; Kim, B.; Smilkov, D.; Wattenberg, M.; Viegas, F.; Corrado, G. S.; Stumpe, M. C.; and Terry, M. 2019a. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359702.

Cai, C. J.; Reif, E.; Hegde, N.; Hipp, J. D.; Kim, B.; Smilkov, D.; Wattenberg, M.; Viégas, F. B.; Corrado, G. S.; Stumpe, M. C.; and Terry, M. 2019b. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Cheng, H. F.; Wang, R.; Zhang, Z.; O'Connell, F.; Gray, T.; Harper, F. M.; and 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*.

Chromik, M.; Eiband, M.; Buchner, F.; Krüger, A.; and Butz, A. M. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. *26th International Conference on Intelligent User Interfaces*.

Das, D.; and Chernova, S. 2020. Leveraging rationales to improve human task performance. *Proceedings of the 25th International Conference on Intelligent User Interfaces*.

De-Arteaga, M.; Fogliato, R.; and Chouldechova, A. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

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*.

Dodge, J.; Liao, Q. V.; Zhang, Y.; Bellamy, R. K. E.; and Dugan, C. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces*.

Feng, S.; and Boyd-Graber, J. L. 2018. What can AI do for me?: evaluating machine learning interpretations in cooperative play. *Proceedings of the 24th International Conference on Intelligent User Interfaces*.

Gajos, K. Z.; and Mamykina, L. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. *27th International Conference on Intelligent User Interfaces*.

Gomez, O.; Holter, S.; Yuan, J.; and Bertini, E. 2020. ViCE: visual counterfactual explanations for machine learning models. *Proceedings of the 25th International Conference on Intelligent User Interfaces*.

Green, B.; and Chen, Y. 2019a. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Green, B.; and Chen, Y. 2019b. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 3: 1 – 24.

Grgić-Hlača, N.; Engel, C.; and Gummadi, K. P. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Guo, S.; Du, F.; Malik, S.; Koh, E.; Kim, S.; Liu, Z.; Kim, D.; Zha, H.; and Cao, N. 2019. Visualizing Uncertainty and Alternatives in Event Sequence Predictions. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Harrison, G.; Hanson, J.; Jacinto, C.; Ramirez, J.; and Ur, B. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Jacobs, M. L.; He, J.; Pradier, M. F.; Lam, B.; Ahn, A. C.; McCoy, T. H.; Perlis, R. H.; Doshi-Velez, F.; and Gajos, K. Z. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Kocielnik, R.; Amershi, S.; and Bennett, P. N. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Kunkel, J.; Donkers, T.; Michael, L.; Barbu, C.-M.; and Ziegler, J. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Lai, V.; Liu, H.; and Tan, C. 2020. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Lai, V.; and Tan, C. 2018. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Lee, M. H.; Siewiorek, D. P.; Smailagic, A.; Bernardino, A.; and i Badia, S. B. 2020. Co-Design and Evaluation of an Intelligent Decision Support System for Stroke Rehabilitation Assessment. *Proceedings of the ACM on Human-Computer Interaction*, 4: 1 – 27.

Lee, M. H.; Siewiorek, D. P.; Smailagic, A.; Bernardino, A.; and i Badia, S. B. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Lee, M. K.; Jain, A.; Cha, H. J.; Ojha, S.; and Kusbit, D. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Levy, A.; Agrawal, M.; Satyanarayan, A.; and Sontag, D. 2021a. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.

Levy, A.; Agrawal, M.; Satyanarayan, A.; and Sontag, D. A. 2021b. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Liu, H.; Lai, V.; and Tan, C. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 5: 1 – 45.

Lu, Z.; and Yin, M. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Lucic, A.; Haned, H.; and de Rijke, M. 2019. Why does my model fail?: contrastive local explanations for retail forecasting. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Nourani, M.; Roy, C.; Block, J. E.; Honeycutt, D. R.; Rahman, T.; Ragan, E. D.; and Gogate, V. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. *26th International Conference on Intelligent User Interfaces*.

Park, J. S.; Berlin, R. B.; Kirlik, A.; and Karahalios, K. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 3: 1 – 15.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Vaughan, J. W.; and Wallach, H. M. 2018. Manipulating and Measuring Model Interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Rader, E. J.; Cotter, K.; and Cho, J. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

Smith-Renner, A.; Fan, R.; Birchfield, M. K.; Wu, T. S.; Boyd-Graber, J. L.; Weld, D. S.; and Findlater, L. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Szymanski, M.; Millecamp, M.; and Verbert, K. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. *26th International Conference on Intelligent User Interfaces*.

Tsai, C.-H.; You, Y.; Gui, X.; Kou, Y.; and Carroll, J. M. 2021. Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

van Berkel, N.; Gonçalves, J.; Russo, D.; Hosio, S. J.; and Skov, M. B. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Wang, X.; and Yin, M. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. *26th International Conference on Intelligent User Interfaces*.

Yang, F.; Huang, Z.; Scholtz, J.; and Arendt, D. L. 2020. How do visual explanations foster end users' appropriate trust in machine learning? *Proceedings of the 25th International Conference on Intelligent User Interfaces*.

Yin, M.; Vaughan, J. W.; and Wallach, H. M. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Yu, K.; Berkovsky, S.; Taib, R.; Zhou, J.; and Chen, F. 2019. Do I trust my machine teammate?: an investigation from perception to decision. *Proceedings of the 24th International Conference on Intelligent User Interfaces*.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. E. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.