# Beyond End Predictions: Stop Putting Machine Learning First and Design Human-Centered AI for Decision Support

**Zana Buçinca**
Harvard University
Boston, MA
zbucinca@seas.harvard.edu

**Alexandra Chouldechova**
Microsoft Research
New York, NY
alexandrac@microsoft.com

**Jennifer Wortman Vaughan**
Microsoft Research
New York, NY
jenn@microsoft.com

**Krzysztof Z. Gajos**
Harvard University
Boston, MA
kgajos@eecs.harvard.edu

## 1 Machine-Learning-First Paradigm Hinders Effective AI-Assisted Decision Making

AI-driven decision-support tools often share a common form: given a decision instance, the decision maker is presented with a prediction from a machine learning model, with or without an explanation. Sometimes, this model predicts a factor thought to be pivotal for the decision, such as a risk score [8]. Other times, the model implicitly or directly predicts a recommended decision, such as whether a patient has a certain disease [7] or which course of treatment to select for a patient [9]. We argue that this "ML-first" paradigm of building decision-support tools around a single machine learning model with readily available data has emerged primarily out of convenience and may be fundamentally limited. We suggest that the community move towards more robust and human-centered ways of supporting decision makers with AI-powered tools.

Mounting empirical evidence from the human-AI decision-making space suggests that end predictions, as the most common output of the ML-first paradigm, fail to deliver on expectations [1, 3]. Human-AI complementarity—where the human and AI team outperforms both the human and AI system alone by harnessing their complementary strengths—remains elusive with the current design of AI decision support [11, 12, 20]. Provided with end predictions, people exhibit inappropriate reliance on them across tasks and settings. Explainable AI, which emerged as a field with the hope of helping the decision makers understand why certain predictions were made, does not seem to help people calibrate reliance either. Numerous studies have shown that people—including experts [5, 7, 19]—are susceptible to erroneous AI recommendations, even when explanations are present [1, 3, 14]. Rather than drawing attention to AI mistakes, explanations seem to serve as a signal of AI's competence and induce further overreliance compared to AI recommendations with no explanations [1, 2].

The ML-first paradigm, in addition to hindering complementary team performance, impedes learning about the domain and may even contribute to the deskilling of the decision maker in the long term. End predictions reduce people's cognitive engagement with the actual decision-making task and the presented information as they shift the focus towards the AI [2, 3, 6]. Part of the reason why people tend to overrely on the provided incorrect AI recommendations is their lack of cognitive engagement with the presented content [2, 3, 6]. Learning is also a measurable outcome of cognitive engagement [16]. Recent work has shown that people learn about the domain and cognitively engage with explanations only when there are no AI predictions present [6]. While our focus is on knowledge work, prior work from decision aids in the context of automation demonstrates that

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

decision recommendations may lead to deskilling of the human operator [15] and that providing information only rather than end predictions is in fact more beneficial for such tasks as well [4].

Given these concerns, why is the ML-first paradigm so prominent? The central reason that has made end predictions ubiquitous as decision-support design rests on the underlying machine learning pipeline rather than on the understanding of human needs, cognition, and behavior when making decisions and when assisted in decision making. While decision-support systems have long been around, it is only in the recent years, with techniques such as deep learning, that providing end predictions has become feasible. However, just because we can now provide such predictions, it does not mean we should in all cases.

Overall, we argue that the field should shift the focus from the ML-first paradigm when building decision-support tools. Instead, we should understand the actual needs of decision makers in context (e.g., via need-finding studies) and build decision aids that support those needs rather than conveniently framing the problem as an end-to-end prediction. We call for deeper reflection on the role of AI in supporting decision making, the design of decision aids, and their long- and short-term impacts on the decision makers.

## 2 Designing AI for Decision-Making Support

Effective decision-support tools can have an immensely positive impact on both decision subjects and decision makers, as stakeholders that have been mostly overlooked by current designs. Countless decision subjects could potentially receive higher quality decisions if human-AI complementarity could be achieved. At the same time, by promoting cognitive engagement and understanding of the task and its underlying causal mechanisms, these tools will improve the skills and capabilities of the decision makers in the long term. Thus, such effective designs will increase the decision maker's agency and independence, in contrast to the current paradigm which renders them co-dependent on AI and susceptible to AI mistakes.

Building on prior work [4, 6], we argue that a promising path forward for effective decision-support tools may be providing relevant information or synthesis either about the data or the model that will help human decision makers make informed decisions and expand their knowledge about the task. This information may or may not be in the form of explanations. In contrast to the goal of explainable and interpretable approaches, however, its main purpose would be assisting the decision maker with the decision-making *task*, rather than solely helping them build a mental model of how the AI makes decisions.

Different types of tasks, however, will necessitate different types of AI support. For each task, careful investigation of the task challenges, underlying cognitive processes it requires, and the best intervention point(s) for the decision aid will be necessary. Current taxonomies of human-AI decision-making tasks that group the tasks along machine-centered dimensions such as types of data (e.g., images, tabular, textual) or types of machine learning problems (e.g., classification, regression) [13] may be insufficient when designing for actual decision support.

Recently, other voices in the AI-assisted decision community have also called for rethinking the design of decision-support tools [10, 17, 18]. For example, in the context of child welfare, Kawakami et al. [10] probed social workers' challenges in integrating AI tools in their decision making and explored ways of designing more useful tools for their needs. In the context of aviation, Storath et al. [18] suggest that shifting the design goal of decision-support tools toward situation awareness rather than decision itself may increase pilots' trust in these tools. These present great examples of studying decision makers in context and designing tools for their actual needs.

Designing AI for decision-making support will introduce novel research challenges across fields – from machine learning to human-computer interaction (HCI), their intersection, and beyond. Rigorous work and human studies are necessary to understand human needs when making decisions in different tasks and settings, to introduce appropriate information synthesis for those tasks, and evaluate them in context. From dataset collection to model building and explanations, appropriate decision-support design will open up new challenges in each step of the machine learning pipeline. For example, we may need new datasets with intermediary labels/annotations rather than raw input to output labels, new ways of building models that predict the intermediary steps as well, and explanations or other interpretability techniques that target decision support rather than other goals like debugging.

Ultimately, we strongly encourage the community to shift from the current ML-first paradigm towards a human-centered approach to building decision-support tools that will amplify the strengths of both human and the AI.

## References


[1] Gagan Bansal, Tongshuang Wu, Joyce Zhu, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, page 1–16, New York, NY, USA, 2021. Association for Computing Machinery.

[2] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 454–464, 2020.

[3] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[4] William M Crocoll and Bruce G Coury. Status or recommendation: Selecting the type of information for decision aiding. In *Proceedings of the human factors society annual meeting*, volume 34, pages 1524–1528. SAGE Publications Sage CA: Los Angeles, CA, 1990.

[5] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. Who goes first? influences of human-ai workflow on decision making in clinical imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1362–1374, 2022.

[6] Krzysztof Z Gajos and Lena Mamykina. Do people engage cognitively with ai? impact of ai assistance on incidental learning. In *27th International Conference on Intelligent User Interfaces*, pages 794–806, 2022.

[7] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. Do as ai say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4(1):1–8, 2021.

[8] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[9] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, 11(1):1–9, 2021.

[10] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. "why do i care what's similar?" probing challenges in ai-assisted child welfare decision-making through worker-ai interface design concepts. In *Designing Interactive Systems Conference*, pages 454–470, 2022.

[11] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 29–38, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287590. URL `https://doi.org/10.1145/3287560.3287590`.

[12] Vivian Lai, Han Liu, and Chenhao Tan. "why is 'chicago' deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376873. URL `https://doi.org/10.1145/3313831.3376873`.

[13] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

[14] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[15] Tapani Rinta-Kahila, Esko Penttinen, Antti Salovaara, and Wael Soliman. Consequences of discontinuing knowledge work automation-surfacing of deskilling effects and methods of recovery. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[16] Jerome I Rotgans and Henk G Schmidt. Cognitive engagement in the problem-based learning classroom. *Advances in health sciences education*, 16(4):465–479, 2011.

[17] Max Schemmer, Niklas Kühl, and Gerhard Satzger. Intelligent decision assistance versus automated decision-making: Enhancing knowledge work through explainable artificial intelligence. *arXiv preprint arXiv:2109.13827*, 2021.

[18] Cara Storath, Zelun Tony Zhang, Yuanting Liu, and Heinrich Hussmann. Building trust by supporting situation awareness: Exploring pilots' design requirements for decision support tools. *TRAIT Workshop at CHI'22*, 2022.

[19] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020.

[20] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 295–305, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372852. URL `https://doi.org/10.1145/3351095.3372852`.