Decoding AI's Nudge: A Unified Framework to Predict Human Behavior in AI-Assisted Decision Making

Zhuoyan Li, Zhuoran Lu, Ming Yin

Purdue University, USA
li4178@purdue.edu, lu800@purdue.edu, mingyin@purdue.edu

Abstract

With the rapid development of AI-based decision aids, different forms of AI assistance have been increasingly integrated into the human decision making processes. To best support humans in decision making, it is essential to quantitatively understand how diverse forms of AI assistance influence humans' decision making behavior. To this end, much of the current research focuses on the end-to-end prediction of human behavior using "black-box" models, often lacking interpretations of the nuanced ways in which AI assistance impacts the human decision making process. Meanwhile, methods that prioritize the interpretability of human behavior predictions are often tailored for one specific form of AI assistance, making adaptations to other forms of assistance difficult. In this paper, we propose a computational framework that can provide an interpretable characterization of the influence of different forms of AI assistance on decision makers in AI-assisted decision making. By conceptualizing AI assistance as the "nudge" in human decision making processes, our approach centers around modelling how different forms of AI assistance modify humans' strategy in weighing different information in making their decisions. Evaluations on behavior data collected from real human decision makers show that the proposed framework outperforms various baselines in accurately predicting human behavior in AI-assisted decision making. Based on the proposed framework, we further provide insights into how individuals with different cognitive styles are nudged by AI assistance differently.

Introduction

As AI technology advances, AI models are increasingly integrated into the human decision making process spanning various domains from healthcare to finance. This has created a new paradigm of human-AI collaboration—Given a decision making task, AI provides assistance to humans while humans make the final decisions. To fully unlock the potential of AI-based decision aids in enhancing human decision making, a growing line of research has been developed in the human-computer interaction community to develop different forms of AI assistance to better support decision makers (Lai et al. 2023). Each form of AI assistance shows different impacts on human decision makers. For instance, the most intuitive type of AI assistance is to directly provide the decision maker with an AI model's decision recommendation on a task (Passi and Vorvoreanu 2022; Wang, Liang, and Yin 2023), though it is found that such assistance sometimes foster a degree of over-reliance on the AI recommendation (Ma et al. 2023). In contrast, the delayed recommendation paradigm, another form of AI assistance where humans are first required to deliberate on a task before receiving the AI model's decision recommendation, has the potential to mitigate this over-reliance, but possibly at the cost of increased under-reliance (Buçinca, Malaya, and Gajos 2021; Fogliato et al. 2022). Thus, to best utilize diverse forms of AI assistance and to determine when and how to present the most suitable type of AI assistance to humans, it is critical to quantitatively understand and predict how these AI assistance influences humans on different decision making tasks.

A few existing studies have worked on modeling and predicting human behavior in AI-assisted decision making (Kumar et al. 2021; Bansal et al. 2021), but they come with several limitations. For example, some research focuses on predicting humans' interaction with AI assistance in an end-to-end manner using black-box models (Subrahmanian and Kumar 2017). While these methods can be effortlessly adapted to various forms of AI assistance, the black-box nature of the model makes it challenging to unpack the cognitive mechanisms driving humans' decision behavior under AI influence. Meanwhile, other studies that aim for interpretability propose computational models based on economics or psychology theories. For instance, Wang, Lu, and Yin (2022) employed the Cumulative Prospect Theory (CPT) to understand how humans decide whether to adopt AI recommendations by analyzing the utility and cost of different decision options. Kumar et al. (2021) captured humans' metacognitive processes of deciding when to rely on themselves and when to solicit AI assistance using the metacognitive bandits model. However, these models are typically tailored for one specific form of AI assistance, making generalizations to other forms of AI assistance a challenging task that requires significant methodology adaptation.

The absence of a unified computational framework to quantitatively characterize how diverse forms of AI assistance influence human decision making processes in an interpretable way impedes the further intelligent utilization of AI assistance. As such, decision makers often have to interact with the default forms of AI assistance instead of benefit-

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ing from personalized and intelligent AI assistance that can best support them. Therefore, in this study, we aim to bridge this gap by proposing such a computational framework.

Specifically, inspired by Callaway, Hardy, and Griffiths (2022) that explores the designs of optimal nudges for cognitively bounded agents, we conceptualize the AI assistance as a "*nudge*" to the human decision making process, which would modify how humans weigh different information in making their decisions. Therefore, in our framework, we first establish an independent decision model that reflects how humans form their independent decisions without any AI assistance. We then model the nudge of AI assistance to humans as the alterations to their decision models. To evaluate the performance of the proposed framework, we collect data on real human subjects' decisions in AI-assisted diabetes prediction tasks with the aids of three common types of AI assistance through a randomized experiment. By fitting various computational models to the behavior dataset collected, we find that our proposed framework consistently outperforms other baseline models in accurately predicting the human decision behavior under different forms of AI assistance. Furthermore, the proposed framework demonstrates robust performance in accurately predicting human behavior in AI-assisted decision making even with limited training data. Lastly, through a detailed analysis of the nudging effects of AI assistance identified by our framework, we offer quantitative insights into how individuals with different cognitive styles are nudged by AI assistance differently. For instance, we observed that AI explanations appear to show a larger effect in redirecting the attention of intuitive decision makers than reflective decision makers.

Related Work

Empirical Studies in AI-Assisted Decision Making. The increased usage of decision aids driven by AI models has inspired a line of experimental studies that identify different forms of AI assistance to enhance human-AI collaboration in decision making (Lai et al. 2023). By surveying the literature related to AI-assisted decision making in the ACM Conference on Human Factors in Computing Systems, ACM Conference on Computer-supported Cooperative Work and Social Computing, ACM Conference on Fairness, Accountability, and Transparency, and ACM Conference on Intelligent User Interfaces from 2018 to 2021, we identify three common types of AI assistance:

1. *Immediate assistance*: The AI model's decision recommendation on the decision making task and other indicators of the recommendation are provided to decision makers upfront. Typical indicators of the AI recommendation include the AI model's accuracy (Lai, Liu, and Tan 2020), explanations of the AI recommendation (Poursabzi-Sangdeh et al. 2018; Cheng et al. 2019; Smith-Renner et al. 2020; Liu, Lai, and Tan 2021; Tsai et al. 2021; Bansal et al. 2020; Zhang, Liao, and Bellamy 2020), and confidence levels of the recommendation (Green and Chen 2019; Guo et al. 2019; Zhang, Liao, and Bellamy 2020; Levy et al. 2021), These indicators may help decision makers gauge the credibility of AI recommendation and calibrate their trust in AI. Since various indicators of the AI recommendation serve similar purposes, aligning with prior research (Tejeda et al. 2022; Wang, Lu, and Yin 2022), we focus on modeling how immediate assistance influences human decision makers when the model's prediction confidence is used as the indicator in this study.
2. *Delayed recommendation* (Park et al. 2019; Grgic-Hlaca, Engel, and Gummadi 2019; Lu and Yin 2021; Buçinca, Malaya, and Gajos 2021; Fogliato et al. 2022; Ma et al. 2023): Humans need to first make an initial decision on the task before the AI model's decision recommendation is revealed to them; this type of AI assistance forces humans to engage more thoughtfully with the AI recommendation.
3. *Explanation only* (Lucic, Haned, and de Rijke 2019; Alqaraawi et al. 2020; Rader, Cotter, and Cho 2018; Schuff et al. 2022; van Berkel et al. 2021): Only the AI model's decision explanation but not its decision recommendation is provided to decision makers. The explanation often points out important features of the task that contribute the most to the AI model's unrevealed decision recommendation, aiming to highlight information that AI believes as highly relevant for decision making.

For more details of the literature review, please see the supplementary material. In this study, we focus on building a computational framework to characterize how different forms of AI assistance, such as the three types identified above, impact humans in AI-assisted decision making.

Modeling Human Behavior in AI-assisted Decision Making. Most recently, there has been a surge of interest among researchers in computationally modeling human behavior in AI-assisted decision making (Bansal et al. 2021; Kumar et al. 2021; Tejeda et al. 2022; Pynadath, Wang, and Kamireddy 2019; Li, Lu, and Yin 2023; Lu et al. 2023). Many of these studies build their models on economics frameworks (Wang, Lu, and Yin 2022), which explain human decision making behavior under uncertainty, or on psychological frameworks that describe the relationship between human trust and reliance on automated systems (Ajenaghughrure et al. 2019; Li, Lu, and Yin 2023). However, most existing works are either tailored to one specific form of AI assistance or lack interpretations of how AI assistance influences human decision making processes. Inspired by the recent research in computationally modeling the effects of nudges (Callaway, Hardy, and Griffiths 2022), we take a different approach in this paper and build a framework to characterize diverse forms of AI assistance as nudges in the human decision making process.

Methods

Problem Formulation

We now formally describe the AI-assisted decision making scenario in this study. Suppose a decision task can be characterized by an n-dimensional feature $\boldsymbol{x} \in \mathcal{R}^n$, and y is the correct decision to make in this task. In this study, we focus on decision making tasks with binary choices of decisions, i.e., $y \in \{0, 1\}$, and each feature x_i of the task $\boldsymbol{x}$

is normalized to fall within the interval of $[0, 1]$. We use $\mathcal{M}(\boldsymbol{x}; \boldsymbol{w}_m)$ to denote the AI model's output on the decision task ($\boldsymbol{w}_m$ are model parameters), and it is within the range of $[0, 1]$. Given $\mathcal{M}(\boldsymbol{x}; \boldsymbol{w}_m)$, the AI model can provide a binary decision recommendation to the human decision maker (DM), i.e., $y^m = \mathbb{1}(\mathcal{M}(\boldsymbol{x}; \boldsymbol{w}_m) > 0.5)$. The AI model's confidence in this recommended decision is $c^m = \max\{\mathcal{M}(\boldsymbol{x}; \boldsymbol{w}_m), 1 - \mathcal{M}(\boldsymbol{x}; \boldsymbol{w}_m)\}$. Following explainable AI methods like LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017), the AI model could also provide some "explanations" of its decision recommendation, $\boldsymbol{e} = \mathcal{E}(\mathcal{M}(\boldsymbol{x}; \boldsymbol{w}_m)), \boldsymbol{e} \in \mathcal{R}^n$, by highlighting the "important" features that contribute the most to the decision recommendation. Here, $e_i \in \{0, 1\}$, where $e_i = 1$ means the feature x_i is highlighted as important, while $e_i = 0$ means the feature x_i is not highlighted. In addition, we assume that the human DM also independently forms their own judgment of the decision task, which is characterized by the function $\mathcal{H}(\boldsymbol{x}; \boldsymbol{w}_h)$ whose output is in the range of $[0, 1]$. Thus, $y^h = \mathbb{1}(\mathcal{H}(\boldsymbol{x}; \boldsymbol{w}_h) > 0.5)$ represents the human DM's independent binary decision.

We consider the setting where the human DM is asked to complete a set of T decision tasks with the help of the AI model. For each task t ($1 \leq t \leq T$), the human DM is given the feature vector $\boldsymbol{x}^t$ and the AI assistance. As discussed previously, we focus on studying the following three forms of AI assistance:

- *Immediate assistance:* The AI model's binary decision recommendation $y^{m,t}$ and its confidence $c^{m,t}$ are immediately provided to the DM along with the task $\boldsymbol{x}^t$.
- *Delayed recommendation:* The DM is required to first make an initial independent decision $y^{h,t}$ on the task. After that, the AI model's binary decision recommendation $y^{m,t}$ will be revealed to the DM.
- *Explanation only:* The DM is only provided with the AI model's explanation $\boldsymbol{e}^t$, which highlights the important features of the task that contributes the most to the AI model's unrevealed decision recommendation.

The DM's independent judgement on the task is $y^{h,t}$—this is observed as the DM's initial decision when AI assistance comes in the form of *delayed recommendation*, but is unobserved (thus requires inference) when AI assistance comes in the other two forms. Given both their own judgement and the AI assistance, the DM then makes a final decision $\hat{y}^t$ on the task. The goal of our study is to quantitatively characterize how the DM is "nudged" by different forms of AI assistance in making their final decision on each task.

Model Decision Makers' Independent Judgement

To characterize how AI assistance nudges human DMs in AI-assisted decision making, it is necessary to first understand how human DMs form their independent judgement *without* being nudged by AI. That is, we need to quantify human DMs' independent decision model $\mathcal{H}(\boldsymbol{x}; \boldsymbol{w}_h)$. Since each DM may have their own unique independent decision making model with different model parameters $\boldsymbol{w}_h$, given a training dataset of the DM's independent decisions $\mathcal{D} = \{\boldsymbol{x}_i, y_i^h\}_{i=1}^N$, we adopt a Bayesian approach and set out to learn from the training dataset the posterior *distribution* of model parameters for a population of diverse DMs, i.e., $\mathcal{P}(\boldsymbol{w}_h|\mathcal{D})$, instead of learning a point estimate. As directly computing this posterior $\mathcal{P}(\boldsymbol{w}_h|\mathcal{D})$ is intractable, we leverage variational inference to approximate it using the parameterized distribution $q_\phi(\boldsymbol{w}_h) = \mathcal{N}(\boldsymbol{w}_h; \boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi)$ and minimize the KL divergence between $q_\phi(\boldsymbol{w}_h)$ and $\mathcal{P}(\boldsymbol{w}_h|\mathcal{D})$:

$$\begin{aligned}
&\mathrm{KL}(q_\phi(\boldsymbol{w}_h)\|\mathcal{P}(\boldsymbol{w}_h|\mathcal{D})) = \int_{\boldsymbol{w}_h} q_\phi(\boldsymbol{w}_h) \log \frac{q_\phi(\boldsymbol{w}_h)}{\mathcal{P}(\boldsymbol{w}_h|\mathcal{D})} \mathrm{d}\boldsymbol{w}_h \\
&= \int_{\boldsymbol{w}_h} q_\phi(\boldsymbol{w}_h)(\log \frac{q_\phi(\boldsymbol{w}_h)}{\mathcal{P}(\boldsymbol{w}_h)} - \log \mathcal{P}(\mathcal{D}|\boldsymbol{w}_h) + \log \mathcal{P}(\mathcal{D})) \mathrm{d}\boldsymbol{w}_h \\
&= \mathrm{KL}(q_\phi(\boldsymbol{w}_h)\|\mathcal{P}(\boldsymbol{w}_h)) - \mathrm{E}_{q_\phi(\boldsymbol{w}_h)}[\log \mathcal{P}(\mathcal{D}|\boldsymbol{w}_h) - \log \mathcal{P}(\mathcal{D})]
\end{aligned} \tag{1}$$

where $\mathcal{P}(\boldsymbol{w}_h)$ is the prior distribution of $\boldsymbol{w}_h$ and $\mathcal{P}(\mathcal{D})$ is a constant[1]. Given the learned $q_\phi(\boldsymbol{w}_h)$, and without additional knowledge of a human DM's unique independent decision making model, we can only postulate that the DM follows an average model to make their decision:

$$y^{h,t} = \mathbb{1}(\mathrm{E}_{q_\phi(\boldsymbol{w}_h)}[\mathcal{H}(\boldsymbol{x}^t; \boldsymbol{w}_h)] > 0.5) \tag{2}$$

Moreover, after we possess additional observations of the human DM's decision making behavior (e.g., the initial decision $y^{h,t}$ that they make), we can update our belief of the DM's independent decision making model from the general parameter distribution $q_\phi(\boldsymbol{w}_h)$ in order to align with the observed human behavior:

$$\hat{q}_\phi(\boldsymbol{w}_h) \propto q_\phi(\boldsymbol{w}_h) \cdot \mathbb{1}(\mathbb{1}(\mathcal{H}(\boldsymbol{x}^t; \boldsymbol{w}_h) \geq 0.5) = y^{h,t}) \tag{3}$$

Without loss of generality, in this study, we assumed that humans' decision making model $\mathcal{H}(\boldsymbol{x}^t; \boldsymbol{w}_h)$ follows the form of logistic model:

$$\mathcal{H}(\boldsymbol{x}^t; \boldsymbol{w}_h) = \mathrm{sigmoid}(\boldsymbol{w}_h \cdot \boldsymbol{x}^t) \tag{4}$$

Quantify the Nudging Effects of AI Assistance

Inspired by a recent computational framework for understanding and predicting the effects of nudges (Callaway, Hardy, and Griffiths 2022), in this study, we introduce a computational framework to provide an interpretable and quantitative characterization of the influence of diverse forms of AI assistance on human decision makers, which enables us to predict human behavior in AI-assisted decision making. The core idea of this framework is to conceptualize the AI assistance as a "nudge" to the human decision making process, such that it can modify how the human DM weighs different information in their decision making and alter their independent decision model accordingly. Depending on the type of AI assistance used, this alternation could be operationalized as the human DM changing their belief in the relevance of certain task feature to their decisions, or as the human DM redirecting their attention to certain task feature when making their decisions.

Immediate Assistance. As shown in Figure 1, in this scenario, human DMs are directly presented with the AI model's decision recommendation $y^{m,t}$ and confidence $c^{m,t}$

[1] In this study, $\mathcal{P}(\boldsymbol{w}_h)$ is set to be $\mathcal{N}(\boldsymbol{w}_h; \boldsymbol{0}, \boldsymbol{I}_n)$.

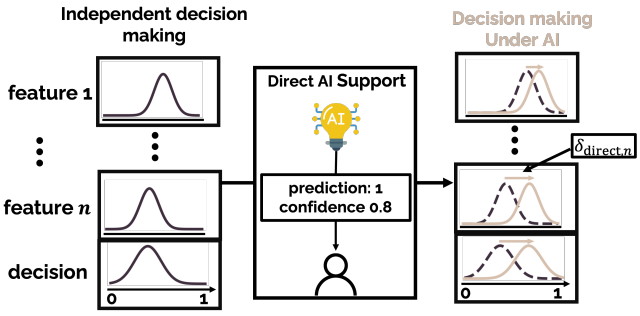


Figure 1: The illustration of how *immediate assistance* nudges human decision makers.

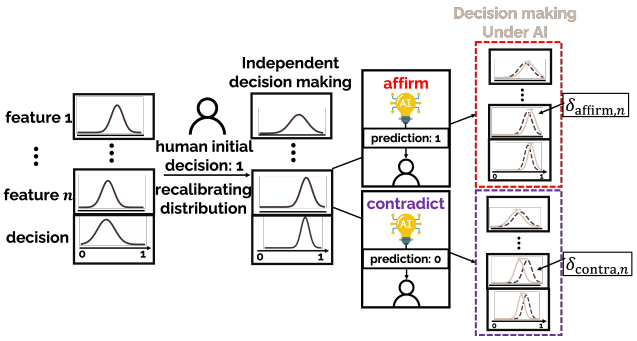


Figure 2: The illustration of how *delayed recommendation* nudges human decision makers.

before they deliberate on the task trial $\boldsymbol{x}^t$. Human DMs may consciously or unconsciously incorporate the AI recommendation into their final decisions. The specific influence of AI on a human DM may largely vary with the DM's inherent attitudes towards AI. For example, DMs with a high tendency to trust AI may simply adopt AI's recommendation while skeptical DMs may simply adopt the opposite decision than what AI suggests. In less extreme cases, DMs are not simply trust or distrust the AI recommendation, but the AI recommendation may change their belief of the relevance/importance of different task features so that they can either align more with the AI recommendation $\hat{y}^t$, or deviate more from the AI recommendation $\hat{y}^t$. The magnitude of this adjustment may be controlled by the AI model's confidence level $c^{m,t}$. Therefore, given any human DM whose independent decision making model is decided by $\boldsymbol{w}_h \sim q_\phi(\boldsymbol{w}_h)$, with the information $y^{m,t}$ and $c^{m,t}$, the DM's final decision $\hat{y}^t$ would be nudged by $y^{m,t}$ and $c^{m,t}$ as:

$$\hat{y}^t = \mathbb{1}(\mathrm{E}_{q_\phi(\boldsymbol{w}_h)}[\mathcal{H}(\boldsymbol{x}^t; \boldsymbol{w}_h + (2y^{m,t}-1)c^{m,t}\boldsymbol{\delta}_{\text{direct}})] > 0.5) \tag{5}$$

where $\boldsymbol{\delta}_{\text{direct}} \in \mathcal{R}^n$ ($\delta_{\text{direct},i} \cdot \delta_{\text{direct},j} \geq 0, \forall i,j \in \{1,\ldots,n\}$) represents the updates in the DM's belief of the relevance of different task features after receiving the immediate AI assistance. Note that $\delta_{\text{direct},i} > 0$ indicates that the DM has a disposition to trust the AI (hence they update their belief of task features' relevance to align more with the AI recommendation), whereas $\delta_{\text{direct},i} < 0$ suggests that the DM has a tendency to distrust AI (hence they update their belief of task features' relevance to deviate more from the AI recommendation). $c^{m,t}$ moderates the magnitude of the update, and $y^{m,t}$ controls the direction of the update. For example, if $y^{m,t} = 1$, $\delta_{\text{direct},i} > 0$ (or $\delta_{\text{direct},i} < 0$) increases (or decreases) the chance of the final decision $\hat{y}^t$ being 1 compared to that of the DM's independent decision $y^{h,t}$. Conversely, If $y^{m,t} = -1$, $\delta_{\text{direct},i} > 0$ (or $\delta_{\text{direct},i} < 0$) increases (or decreases) the chance of the final decision $\hat{y}^t$ being -1 compared to that of the DM's independent decision $y^{h,t}$.

Delayed Recommendation. As shown in Figure 2, in this scenario, human DMs are required to deliberate and make their initial decision $y^{h,t}$ on task trial $\boldsymbol{x}^t$ before the AI model's decision recommendation is provided. The observed human DM's initial decision $y^{h,t}$ can be used to update our belief of the DM's independent decision making model. Specifically, after observing the human DM's initial decision $y^{h,t}$, an adjustment is made to the distribution of the human DM's independent decision model $q_\phi(\boldsymbol{w}_h)$ to filter out decision models that are inconsistent with the observed decision, yielding $\hat{q}_\phi(\boldsymbol{w}_h)$ as given by Eq. 3. Then, as the DM compares their initial decision $y^{h,t}$ with the AI model's decision recommendation $y^{m,t}$, two scenarios may arise:

1. *AI affirms human decision* ($y^{h,t} = y^{m,t}$): For DMs who trust AI, this agreement can boost their confidence in their initial decision $y^{h,t}$. Conversely, for DMs who are skeptical of AI, they may become less confident in their own judgement due to this agreement.
2. *AI contradicts human decision* ($y^{h,t} \neq y^{m,t}$): DMs who tend to trust AI might reflect on their initial decision and could be inclined towards switching to $y^{m,t}$. On the other hand, DMs who are skeptical of AI may be more inclined to stand by their own judgement $y^{h,t}$.

Depending on which scenario that the DM encounters, we model the DM's final decision $\hat{y}^t$ as follows:

$$\begin{aligned}\hat{y}^t = \mathbb{1}(\mathrm{E}_{\hat{q}_\phi(\boldsymbol{w}_h)}[\mathcal{H}(\boldsymbol{x}^t; \boldsymbol{w}_h + (2y^{m,t}-1)(\mathbb{1}(y^{m,t}=y^{h,t})\boldsymbol{\delta}_{\text{affirm}} \\ + \mathbb{1}(y^{m,t}\neq y^{h,t})\boldsymbol{\delta}_{\text{contra}}))] > 0.5)\end{aligned} \tag{6}$$

Here, $\boldsymbol{\delta}_{\text{affirm}}, \boldsymbol{\delta}_{\text{contra}} \in \mathcal{R}^n$ ($\delta_{\text{affirm},i} \cdot \delta_{\text{affirm},j} \geq 0$, $\delta_{\text{contra},i} \cdot \delta_{\text{contra},j} \geq 0$, $\forall i,j \in \{1,\ldots,n\}$) represent the updates in the DM's belief of the relevance of different task features after seeing AI confirms their judgement ($y^{m,t} = y^{h,t}$) and AI contradicts their judgement ($y^{m,t} \neq y^{h,t}$), respectively.

Explanation Only. As shown in Figure 3, in this scenario, human DMs are presented with the AI explanation $\boldsymbol{e}^t$, which highlights the critical features of the task trial $\boldsymbol{x}^t$. These explanations may nudge human DMs to redirect their attention to these highlighted features when forming their final decision $\hat{y}^t$. Intuitively, information that is marked as important might be prioritized by DMs, while other information may be overlooked. As such, the highlighted task features may exert a more salient influence on the DM's final decision $\hat{y}^t$:

$$\begin{aligned}\hat{y}^t = \mathbb{1}(\mathrm{E}_{q_\phi(\boldsymbol{w}_h)}[\delta_{\text{exp}}\mathcal{H}(\boldsymbol{e}^t \odot \boldsymbol{x}^t; \boldsymbol{w}_h) + \\ (1-\delta_{\text{exp}})\mathcal{H}((\mathbf{1}-\boldsymbol{e}^t)\odot \boldsymbol{x}^t; \boldsymbol{w}_h)] > 0.5)\end{aligned} \tag{7}$$

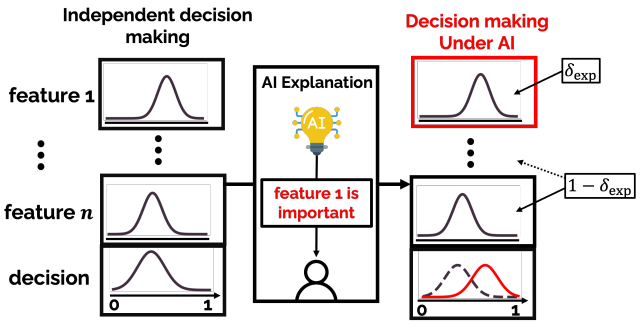


Figure 3: The illustration of how *explanation only* nudges human decision makers.

where $\odot$ is the element-wise product. $\delta_{\text{exp}} \in [0, 1]$ quantifies the degree to which the human DMs redirect their attention to the highlighted features after seeing the AI explanation, with larger values indicating that the DM puts a greater emphasis on the highlighted information.

Human-Subject Experiment

To evaluate the proposed framework, we conducted a human-subject experiment to collect behavior data from real human decision makers in AI-assisted decision making under different forms of AI assistance.

Decision Making Tasks. The decision making task we used in our experiment was to predict diabetes in patients based on demographic information and medical history. Specifically, in each task, the subject is presented with a patient profile encompassing six features: gender, age, history of heart disease, Body Mass Index (BMI), HbA1c level, and blood glucose level at a given time interval. The subject was then asked to determine whether this patient has diabetes. The patient profiles were randomly sampled from the diabetes prediction dataset (Mustafatz 2023).

Experimental Treatments. We created four treatments in this experiment. One of these is an *independent* treatment where human subjects complete decision making tasks without any AI assistance. In the other three treatments, human subjects receive one of the three forms of AI assistance as what we introduced earlier. The AI assistant used in the experiment was based on a boosting tree model trained on the diabetes prediction dataset. The accuracy of the AI model was 87%. In the *Explanation only* treatment, we used SHAP (Lundberg and Lee 2017) to explain the predictions of the boosting tree model. The two most influential features are highlighted as the AI model's explanation.

Experimental Procedure. We posted our experiment on Amazon Mechanical Turk (MTurk) as a human intelligence task (HIT) and recruited MTurk workers as our subjects. Upon arrival, we randomly assigned each subject to one of the four treatments. Subjects started the HIT by completing a tutorial that described the diabetes prediction task that they needed to work on in the HIT and the meaning of each feature they would see in a patient's profile. To familiarize subjects with the task, we initially asked them to complete five training tasks. During these training tasks, subjects made diabetes predictions without AI assistance, and we immediately provided them with the correct answers and the end of each task. The real experiment began after the subject completed the training tasks. Specifically, subjects were asked to complete a total of 30 tasks, which were randomly sampled from a pool of 500 task instances. After subjects completed all 30 tasks, subjects were asked to undertake a 3-item Cognitive Reflection Test (CRT) (Frederick 2005), intended to assess the subject's tendency in engaging with intuitve vs. reflective thinking. We offered a base payment of $1.2 for the HIT. The HIT was open to US-based workers only, and each worker can complete the HIT once. We further included an attention check question within the HIT, where subjects were required to select a randomly determined option. Data collected from subjects who successfully passed the attention check were considered valid for our study (see the supplementary materials for more details of the human-subject experiment).

Evaluations

After filtering the inattentive subjects, we obtained valid data from 202 subjects in our experiment (*Independent*: 53, *Immediate assistance*: 50, *Delayed recommendation*: 53, *Explanation only*: 46). Below, we conduct our evaluation using the behavior data collected from these valid subjects.

Model Training and Baselines

We first learned the general parameter distribution $q_\phi(\boldsymbol{w}_h)$ of human DMs' independent decision making model utilizing the data collected in the *Independent* treatment. Through 5-fold cross-validation, we found the average accuracy in predicting an average human DM's independent decision using $q_\phi(\boldsymbol{w}_h)$ and Eq. 2 is 0.673. In the following evaluations, we used $q_\phi(\boldsymbol{w}_h)$ to reflect the human DM's independent decision making model, and used the maximum likelihood estimation to learn how different forms of AI assistance nudge each individual human DM to modify their decision making model.

Specifically, to evaluate the performance of the proposed framework, for each human subject, we randomly split the behavior data collected from them into training (50%) and test (50%) sets. We computed the average negative log-likelihood (NLL) to measure how well different models capture the likelihood of subjects making their final decisions under the influence of AI assistance, and we averaged the NLL values across all subjects in each treatment. A lower mean NLL indicates a better prediction performance. In addition, we also employed F1-score and Accuracy scores to evaluate the performance of different models. For both these metrics, higher scores denote better performance. To ensure the robustness of evaluations, all experiments were repeated 5 times, and the average performance across these repetitions was reported.

We consider utility-based model proposed by Wang, Lu, and Yin (2022) and a few standard supervised learning models as baselines in evaluations, including Logistic Regression, XGBoost, Multi-Layer Perceptron (MLP), and Support

Treatment	Immediate assistance			Delayed recommendation			Explanation only		
	NLL ↓	Accuracy ↑	F1 ↑	NLL ↓	Accuracy ↑	F1 ↑	NLL ↓	Accuracy ↑	F1 ↑
Logistic Regression	0.522	0.753	0.789	0.446	0.809	0.782	**0.549**	**0.728**	0.767
XGBoost	0.533	0.768	0.737	0.472	0.812	0.797	0.617	0.711	0.753
MLP	0.656	0.753	0.729	0.554	0.777	0.751	0.606	0.686	0.778
SVM	0.530	0.754	0.707	0.461	0.791	0.758	0.603	0.721	0.743
Utility	0.573	0.739	0.779	-	-	-	-	-	-
Ours	**0.435**	**0.800**	**0.818**	**0.413**	**0.825**	**0.812**	0.563	0.715	**0.791**

Table 1: Comparing the performance of proposed method with baseline methods on three forms of AI assistance, in terms of NLL, Accuracy, and F1-score. "↓" denotes the lower the better, "↑" denotes the higher the better. Best result in each column is highlighted in bold. All results are averaged over 5 runs. "-" means the method can not be applied in this scenario.

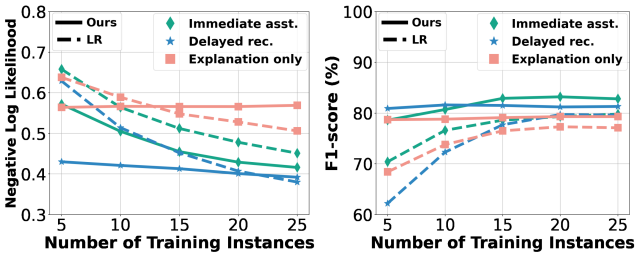
Vector Machines (SVM). These supervised learning models directly predicts human DMs' final decisions $\hat{y}^{h,t}$ in a decision task based on various features:

1. *Immediate assistance:* task trial features $\boldsymbol{x}^t$, as well as the AI model's decision recommendation $y^{m,t}$ and confidence $c^{m,t}$ in the task trial.
2. *Delayed recommendation:* task trial features $\boldsymbol{x}^t$, human DMs' initial decision $y^{h,t}$, and the AI model's decision recommendation $y^{m,t}$ in the task trial.
3. *Explanation only:* task trial features $\boldsymbol{x}^t$ and the AI explanation $\boldsymbol{e}^t$ in the task trial.

Comparing Model Performance

Table 1 presents the comparative performance of various models in predicting human DMs' decisions across three forms of AI assistance. Overall, our proposed method consistently outperforms the baseline methods in the *Immediate assistance* and *Delayed recommendation* scenarios across all metrics by a significant margin. For instance, within the *Immediate assistance* scenario, the NLL for our method stands at a mere 0.435, whereas the best baseline achieves an NLL of 0.522. In the *Explanation only* scenario, the performance of our method is comparable with the best-performing baseline model, logistic regression, in terms of NLL and Accuracy, and outperforms it on the F1-score.

To assess the robustness of our approach, we varied the proportions of training and testing data and observed how the performance of our method changes with the training data size. Given the high performance of the logistic regression model, we selected it as the baseline model in this evaluation. As shown in Figure 4, our approach demonstrates consistently superior performance compared to logistic regression models across three AI-assisted decision making scenarios particularly when the number of training instances is limited. Specifically, the performance of our model remains robust with respect to variations in the amount of training data; it shows only a slight decrease in performance as the number of training instances reduces. In contrast, logistic regression models are highly sensitive to the size of the training data. As the number of training samples decreases, their performance degrades significantly. In other words, unlike the standard supervised learning models like



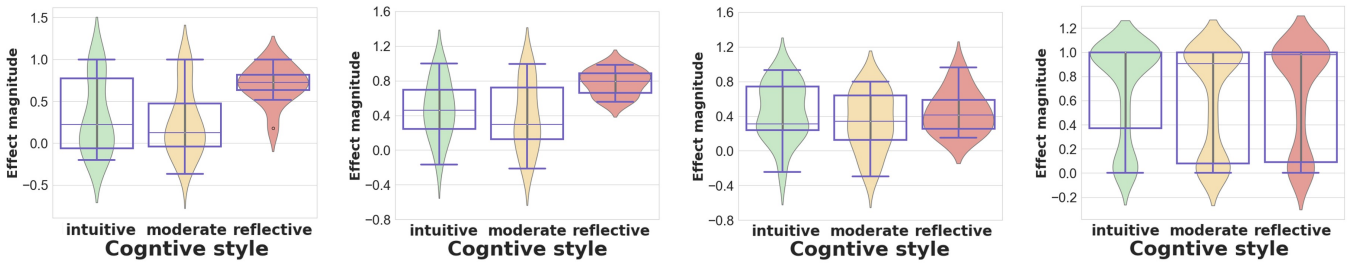
(a) Negative Log Likelihood (b) F1-score

Figure 4: Comparing the performance of our method with logistic regression models when changing the size of training data under three forms of AI assistance.

logistic regression—which requires retraining from scratch for each individual human DM—our approach is endowed with the knowledge of how human DMs in general make decisions. This knowledge makes it possible for us to only tune the parameters of AI's nudging effects $\boldsymbol{\delta}$ on each individual human DM with a few training instances, yet still achieving comparable or even higher performance compared to the supervised learning models.

Examining the Nudging Effect of AI Assistance across Individuals

We now examine how may the AI assistance nudges decision makers with different cognitive styles similarly or differently. To do so, we compared the size of the learned nudging effects on decision makers for three forms of AI assistance. Specifically, we first used all behavior data collected from a human DM to learn the nudging effect of AI assistance on them (i.e., $\boldsymbol{\delta}$). We then used $\text{sign}(\boldsymbol{\delta})||\boldsymbol{\delta}||$ to represent the direction and magnitude of the nudging effects of AI assistance on the human DM ($\text{sign}(\boldsymbol{\delta}) = 1$ when $\forall i, \delta_i > 0$; otherwise $\text{sign}(\boldsymbol{\delta}) = -1$). To categorize the cognitive style of each human DM (i.e., each subject), we utilized their scores in the 3-item Cognitive Reflection Test (CRT) in the experiment—Following previous research (Frederick 2005), subjects scoring 3 were classified as having a reflective thinking style, those with a score



(a) Immediate assistance: $||\boldsymbol{\delta}_{\text{direct}}||$ (b) Delayed recommendation ($y^h = y^m$): $||\boldsymbol{\delta}_{\text{affirm}}||$ (c) Delayed recommendation ($y^h \neq y^m$): $||\boldsymbol{\delta}_{\text{contra}}||$ (d) Explanation only: δ_{exp}

Figure 5: Comparing the nudging effects of AI assistance on decision makers with different cognitive styles across three forms of AI assistance.

of 0 were categorized as having an intuitive thinking style, and those scoring 1 or 2 were categorized as the moderate reflective style.

Figure 5 shows the comparison results of the nudging effects of AI assistance on DMs with different cognitive styles. To examine whether the nudging effects of AI assistance are different across DMs with different cognitive styles, we first used a one-way ANOVA test [2] to determine if there are statistically significant differences in the values of $\text{sign}(\boldsymbol{\delta})||\boldsymbol{\delta}||$ across different groups of DMs. If significant differences are detected, we proceed with post-hoc pairwise comparisons using Tukey's HSD test [3]. Overall, our findings suggest that under the *Immediate assistance* and the *Delayed recommendation* scenarios (when AI affirms human decision), AI assistance exerts a larger influence on DMs with a reflective thinking style than intuitive DMs ($p < 0.05$) and DMs with moderate reflective styles ($p < 0.05$). One potential explanation is that reflective DMs are inclined to deliberate more extensively on tasks and the AI model's recommendations. Thus, through their interactions with the AI model, reflective DMs may have sensed the high performance of the AI model (its accuracy is 87% for the task), making them more willing to align their decisions with the AI recommendation, especially when the AI recommendation affirms their own independent judgement. However, when the human DM's initial decision differs from the AI recommendation in the *Delayed recommendation* scenario, there isn't a statistical difference in the AI's nudging effects across the three types of decision makers. In fact, by comparing the AI's nudging effects on different groups of DMs under the two conditions of the *Delayed recommendation* scenario—AI affirms human decisions or contradicts human decisions—we find that reflective DMs are significantly more likely to align their final decisions with the AI recommendation when AI affirms rather than contradicting their initial judgement ($p < 0.05$). In contrast, the intuitive and moderately reflective DMs do not appear to be impacted by AI significantly differently under these two conditions.

Finally, under the *Explanation only* scenario, we also observe that intuitive DMs tend to place more emphasis on the features highlighted by the AI explanations. The nudging effect of AI explanations on intuitive DMs is found to be significantly greater than that on moderately reflective DMs ($p < 0.05$). While the nudging effect also appears to be slightly larger for intuitive DMs compared to reflective DMs, the difference is not statistically significant.

Conclusion

In this paper, we propose a computational framework to characterize how various AI assistance nudges humans in AI-assisted decision making. We evaluate the proposed model's performance in fitting the real human behavior data collected from a randomized experiment. Our results show that the proposed model consistently outperforms other baselines in accurately predicting humans decisions under diverse forms of AI assistance. Additionally, further analyses based on the proposed framework provided insights into how individuals with varying cognitive styles are impacted by AI assistance differently.

There are a few limitations of this study. For example, the behavior data is collected from laypeople on the diabetes prediction task, which contains a relatively small number of features. It remains to be investigated whether the proposed model can perform well on tasks that involve many more features and thus more complex. Additionally, the AI-assisted decision scenario we examined in this study lacks sequential or temporal feedback regarding AI performance. Further exploration is required to generalize the propose framework to the sequential settings. Lastly, we assumed that the independent human decision model follows the form of logistic regression. Additional research is needed to explore how to adapt the current ways of altering humans' decision models for reflecting the nudging effect of AI assistance on human DMs to other forms of decision models.

[2] Analysis of Variance (ANOVA) is a statistical test for identifying significant differences between group means.

[3] Tukey's HSD (Honestly Significant Difference) is a post-hoc test used to determine specific differences between pairs of group means after a one-way ANOVA test has found significance.

Acknowledgements

We thank the support of the National Science Foundation under grant IIS-2229876 on this work. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

References

Ajenaghughrure, I. B.; Sousa, S. C.; Kosunen, I. J.; and Lamas, D. 2019. Predictive model to assess user trust: a psycho-physiological approach. In *Proceedings of the 10th Indian conference on human-computer interaction*, 1–10.

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*.

Bansal, G.; Nushi, B.; Kamar, E.; Horvitz, E.; and Weld, D. S. 2021. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. In *AAAI Conference on Artificial Intelligence*.

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*.

Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. 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.

Callaway, F.; Hardy, M.; and Griffiths, T. 2022. Optimal nudging for cognitively bounded agents: A framework for modeling, predicting, and controlling the effects of choice architectures.

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*.

Fogliato, R.; Chappidi, S.; Lungren, M. P.; Fitzke, M.; Parkinson, M.; Wilson, D. U.; Fisher, P.; Horvitz, E.; Inkpen, K.; and Nushi, B. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Frederick, S. 2005. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19: 25–42.

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

Grgic-Hlaca, N.; Engel, C.; and Gummadi, K. P. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *DecisionSciRN: Decision-Making in the Legal Field (Topic)*.

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*.

Kumar, A.; Patel, T.; Benjamin, A. S.; and Steyvers, M. 2021. Explaining algorithm aversion with metacognitive bandits. In *Proceedings of the annual meeting of the cognitive science society*, volume 43.

Lai, V.; Chen, C.; Smith-Renner, A.; Liao, Q. V.; and Tan, C. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.

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*.

Levy, A.; Agrawal, M.; Satyanarayan, A.; and Sontag, D. A. 2021. 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*.

Li, Z.; Lu, Z.; and Yin, M. 2023. Modeling Human Trust and Reliance in AI-Assisted Decision Making: A Markovian Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(5): 6056–6064.

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.; Li, Z.; Chiang, C.-W.; and Yin, M. 2023. Strategic adversarial attacks in AI-assisted decision making to reduce human trust and reliance. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 3020–3028.

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*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Ma, S.; Lei, Y.; Wang, X.; Zheng, C.; Shi, C.; Yin, M.; and Ma, X. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Mustafatz. 2023. Diabetes Prediction Dataset. *https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset*.

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.

Passi, S.; and Vorvoreanu, M. 2022. Overreliance on AI Literature Review. *Microsoft Research*.

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*.

Pynadath, D. V.; Wang, N.; and Kamireddy, S. 2019. A Markovian Method for Predicting Trust Behavior in Human-Agent Interaction. *Proceedings of the 7th International Conference on Human-Agent Interaction*.

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*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Schuff, H.; Jacovi, A.; Adel, H.; Goldberg, Y.; and Vu, N. T. 2022. Human Interpretation of Saliency-based Explanation Over Text. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

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*.

Subrahmanian, V.; and Kumar, S. 2017. Predicting human behavior: The next frontiers. *Science*, 355(6324): 489–489.

Tejeda, H.; Kumar, A.; Smyth, P.; and Steyvers, M. 2022. AI-Assisted Decision-making: a Cognitive Modeling Approach to Infer Latent Reliance Strategies. *Computational Brain & Behavior*, 5: 491 – 508.

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.; Liang, C.; and Yin, M. 2023. The effects of AI biases and explanations on human decision fairness: a case study of bidding in rental housing markets. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23, Edith Elkind (Ed.). International Joint Conferences on Artificial Intelligence Organization*, 3076–3084.

Wang, X.; Lu, Z.; and Yin, M. 2022. Will you accept the ai recommendation? predicting human behavior in ai-assisted decision making. In *Proceedings of the ACM Web Conference 2022*, 1697–1708.

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*.