Proceedings of the Twelfth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2024)

Mix and Match: Characterizing Heterogeneous Human Behavior in AI-assisted Decision Making

Zhuoran Lu, Syed Hasan Amin Mahmood, Zhuoyan Li, Ming Yin

Purdue University
{lu800, hasanamin, li4178, mingyin}@purdue.edu

Abstract

AI-assisted decision-making systems hold immense potential to enhance human judgment, but their effectiveness is often hindered by a lack of understanding of the diverse ways in which humans take AI recommendations. Current research frequently relies on simplified, "one-size-fits-all" models to characterize an average human decision-maker, thus failing to capture the heterogeneity of people's decision-making behavior when incorporating AI assistance. To address this, we propose Mix and Match (M&M), a novel computational framework that explicitly models the diversity of human decision-makers and their unique patterns of relying on AI assistance. M&M represents the population of decision-makers as a mixture of distinct decision-making processes, with each process corresponding to a specific type of decision-maker. This approach enables us to infer latent behavioral patterns from limited data of human decisions under AI assistance, offering valuable insights into the cognitive processes underlying human-AI collaboration. Using real-world behavioral data, our empirical evaluation demonstrates that M&M consistently outperforms baseline methods in predicting human decision behavior. Furthermore, through a detailed analysis of the decision-maker types identified in our framework, we provide quantitative insights into nuanced patterns of how different individuals adopt AI recommendations. These findings offer implications for designing personalized and effective AI systems based on the diverse landscape of human behavior patterns in AI-assisted decision-making across various domains.

Introduction

The increasing integration of artificial intelligence (AI) into people's decision-making processes across diverse domains, from entertainment to healthcare and to finance (De Mantaras and Arcos 2002; Shaheen 2021; Cao 2022), has initiated a new era of human-AI collaboration. Combining AI's competence and humans' agency, the paradigm of AI-assisted decision-making, where AI models provide recommendations and humans make the final decisions, holds immense potential to enhance human judgment and improve decision outcomes (Lysaght et al. 2019; Lai et al. 2021). However, realizing this potential hinges on a deep understanding of how humans interact with and adopt AI-generated recommendations (Steyvers and Kumar 2023).

Although a growing body of research has focused on quantitatively describing how human decision-makers respond to AI recommendations, these studies suffer from a few limitations. Some approaches, particularly those rooted in deep learning, treat the problem as a mere prediction task without considering the cognitive underpinnings and interpretability of the decision-making process (Hartford, Wright, and Leyton-Brown 2016). Despite the high performance, these models provide little insight into the underlying reasons behind people's decision behavior. While some recent work has attempted to incorporate cognitive processes for characterizing behavioral patterns in AI-assisted decision-making, these works often rely on an "average" human decision-maker representation (Wang, Lu, and Yin 2022; Tejeda et al. 2022). This simplification overlooks the inherent diversity in people's decision-making patterns under AI assistance, potentially resulted from individual preferences, risk tolerances, and cognitive styles (Franken and Muris 2005; Appelt et al. 2011). Neglecting this diversity impedes the development of personalized AI assistance and restricts our ability to fully harness AI's potential in augmenting human decision-making.

To address these gaps, we propose Mix and Match (M&M), a computational framework designed to model the diverse ways in which humans interact with and adopt AI recommendations. M&M explicitly acknowledges the heterogeneity of human decision-makers. The framework operates in two main stages: "Mix" and "Match". In the "Mix" stage, M&M considers K distinct decision-making processes, each representing a different type of decision-maker. Specifically, each decision process captures the cognitive process the corresponding type of decision-maker goes through to generate their AI-assisted decisions—the decision-maker first forms their independent judgments without AI assistance, and then aggregates their independent judgments with the AI model's recommendations to arrive at a final decision after computing the utilities of different possible actions. We assume that each decision made by an individual is influenced by a probability distribution over these K types, with a latent variable indicating the specific types of decision-makers responsible for that decision. Thus, during this stage, given a set of AI-assisted decisions made by a population of decision-makers,

we jointly learn the latent variables and parameters for each decision-maker type. Next, in the "Match" stage, given a new individual decision-maker, we estimate the likelihood that each decision-maker type is responsible for this individual's final decision on a particular decision task, and predict their final decision accordingly. Note that M&M leverages the varied decision-making behavior across the population to uncover underlying patterns. Additionally, M&M acknowledges that the same individual may exhibit different decision-making behaviors depending on the context or task, providing a more nuanced understanding of the dynamic nature of AI-assisted decision-making.

Using real-world behavioral data collected from diverse decision-making scenarios, our empirical evaluation demonstrates that M&M consistently outperforms baseline methods in predicting human decisions under AI assistance. By accurately characterizing the different types of decision-makers and their unique adoption patterns, our framework offers valuable insights into the cognitive processes underlying human-AI collaboration. For example, our analysis reveals that there exists a range of decision-makers with different perceptions of penalties for incorrect decisions and sensitivities to utility differences in accepting or rejecting AI recommendations. In addition, the majority of decision-makers perceive high penalties for incorrect decisions and exhibit high sensitivity to utility differences. Furthermore, perceptions of penalties for incorrect decisions and sensitivity to utility differences tend to be positively correlated in AI-assisted decision-making. These insights can inform the design of more effective and personalized AI assistance, ultimately leading to improved AI-assisted decision-making outcomes in various domains.

Related Work

Empirical Studies in AI-assisted Decision Making

The increasing use of AI-powered decision aids has spurred a wave of experimental studies aimed at understanding how humans interact with and rely on AI models in decision-making scenarios. Researchers have identified a multitude of factors that can influence people's reliance on AI in decision-making—on a population level—including the model's accuracy (Yin, Wortman Vaughan, and Wallach 2019; Lai and Tan 2019), confidence (Zhang, Liao, and Bellamy 2020; Rechkemmer and Yin 2022), the type and presentation of AI explanations (Yang et al. 2020; Bansal et al. 2021b), individuals' mental models of AI (Bansal et al. 2019a,b), the degree of agreement between human judgment and AI recommendations (Lu and Yin 2021), and more.

Beyond factors existing at the population level, recent studies also found that individual differences make significant impacts on how humans take AI recommendations. For instance, it was found that an individual's personality affects their trust in and advice-taking from AI (Sharan and Romano 2020). As another example, Matthews et al. (2019) found that people could activate different mental models when collaborating with AI, thus leading to diverse attitudes towards AI. These studies have revealed a wide array of behavioral patterns exhibited by decision-makers in AI-assisted contexts, highlighting the importance of characterizing the diversity in human behavior.

Modeling Human Decision Behavior

Research on modeling human decision behaviors has been well-established in economy and psychology, as decision-making is an abstract of a wide range of human behaviors (Wang and Ruhe 2007; Montgomery 1983). Pivoting around this, a large amount of theories and models were developed to capture how people make decisions. For instance, the expected utility theory links behavior with the utilities behind decisions (Schoemaker 1982). Research further reveals that factors including task context and individual differences can lead to people's different ways of calculating the utilities of their actions (Schoemaker 2013).

With AI-based decision aids becoming more prevalent, the community started to investigate computational modeling of human behavior in AI-assisted decision-making, with a focus on characterizing and predicting when decision-makers will solicit or rely on AI recommendations (Pynadath, Wang, and Kamireddy 2019; Bansal et al. 2021a; Li, Lu, and Yin 2023; Kumar et al. 2021; Wang, Lu, and Yin 2022; Guo et al. 2024; Strickland et al. 2024). Drawing inspiration from economic theories (e.g., Cumulative Prospect Theory (Tversky and Kahneman 1992; Allais 1953)) or cognitive modeling exemplified by sociocognitive construct (Askarisichani et al. 2022), previous work made efforts in constructing computational models with the capability to explain human decision-making under modern AI systems with uncertainty. These models have also been used to improve AI-assisted decision-making by enabling AI systems to adapt their recommendations based on human behaviors or by designing interfaces that adjust how AI recommendations are presented depending on how people behave (Ma et al. 2023; Amin, Lu, and Yin 2024). However, most of these studies model decision behaviors using an "average" human decision-maker to represent the entire population, overlooking the diversity in decision-making patterns that can result from individual differences.

Problem Setup

We focus on the *AI-assisted decision-making* setting, where a human decision-maker (DM) completes a sequence of T tasks, receiving a decision recommendation from an AI model on each task but making the final decision by themselves. This setting is particularly prevalent in high-stakes domains such as medical diagnosis, where the human retains the ultimate authority to make the final decision. Each decision making task $t \in \{1, \ldots, T\}$ is characterized by features $\mathbf{x}^t \in \mathbb{R}^n$ and an associated correct decision $y^t \in \mathcal{Y}$. For illustrative purposes and without loss of generality, our study centers on binary classification tasks (i.e., $\mathcal{Y} = \{0, 1\}$).

Under this setup, an AI model first provides a decision recommendation $m(\mathbf{x}^t; \theta_m)$ to a human DM, who has their own independent judgment $h(\mathbf{x}^t; \theta_h)$ on the same case. The human DM then aggregates the AI's suggestion with their own assessment to arrive at a final team decision $\hat{y}^t$:

$$\hat{y}^t = f(\mathbf{x}^t, m(\mathbf{x}^t; \theta_m), h(\mathbf{x}^t; \theta_h); \theta_a) \quad (1)$$

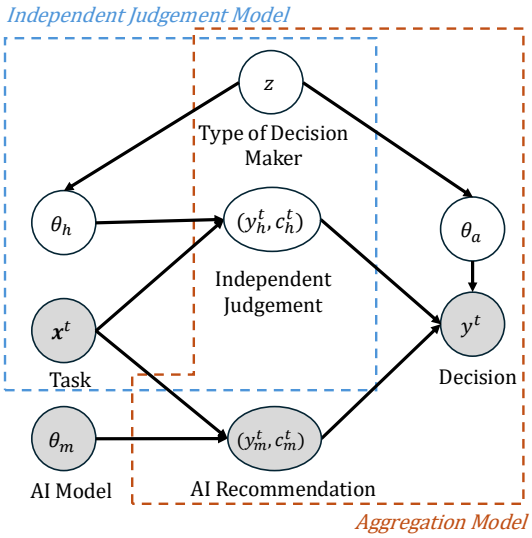


Figure 1: The probabilistic model of the generation process of the decision-maker's final decision in AI-assisted decision-making. The shaded node is observed.

We consider the scenario where the AI decision recommendation comprises two components: a binary decision and the confidence in that decision. Formally, $m(\mathbf{x}^t;\theta_m) = \{1 : \mathbb{P}(y^t = 1 \mid \mathbf{x}^t), 0 : \mathbb{P}(y^t = 0 \mid \mathbf{x}^t)\}$, which can be further used to generate the binary recommendation $\hat{y}_m^t = \arg\max m(\mathbf{x}^t;\theta_m)$ and the confidence in that recommendation $c_m^t = \max m(\mathbf{x}^t;\theta_m) = m(\mathbf{x}^t;\theta_m)[y = \hat{y}_m^t]$. Similarly, the human DM's initial judgment $h(\mathbf{x}^t;\theta_h)$ is characterized by the binary judgement $\hat{y}_h^t = \arg\max h(\mathbf{x}^t;\theta_h)$ and the confidence in that judgment $c_h^t = h(\mathbf{x}^t;\theta_h)[y = \hat{y}_h^t]$. The final team decision $\hat{y}^t = f(\mathbf{x}^t, \hat{y}_m^t, c_m^t, \hat{y}_h^t, c_h^t)$ is influenced by the task features $\mathbf{x}^t$, the AI model's decision recommendation $\hat{y}_m^t$, the AI model's confidence c_m^t, the human DM's initial judgment $\hat{y}_h^t$, and the human DM's confidence c_h^t. Since final team performance is often of paramount interest, it is critical to understand the form of the team decision making model $f(\cdot)$.

While AI model parameters (θ_m) can be accessible, human judgment and aggregation parameters (θ_h, θ_a) are often challenging to characterize. Prior works often assume "average" human behavior models (i.e., each DM shares the same θ_h, θ_a). Our work aims to address this limitation by formally capturing the diversity among decision-making types, recognizing that effective modeling of AI-assisted decision-making must account for individual differences.

Method

We propose a novel computational framework, `Mix and Match` (`M&M`), to effectively characterize human behavior in AI-assisted decision-making. `M&M` is a Bayesian approach that models the diverse ways in which humans interact with AI recommendations as a generative process involving a mixture of different types of decision-makers. Figure 1 illustrates the structure of the proposed model. The framework consists two main stages:

1. **`Mix`: Modeling decisions as mixture models.** In this stage, instead of a single average model, we use in total K distinct decision-making processes $\{f_1, \ldots, f_K\}$. Each decision is influenced by a probability distribution over these K types, with a latent variable indicating how each specific type of decision-making process is responsible for that decision trial.
2. **`Match`: Inferring DM types.** In this stage, we match a decision trial with a distribution of DM types by inferring how likely each type is to have generated a particular decision given the observed data.

As a particular realization of the `M&M` framework, we now introduce how we model the decision-generation process, as well as how model learning and inference can be done in practice.

Decision Generation

Modeling a Single Type of DM We start with elaborating on how a single type of DM (i.e., DM type k) generates the final decision on task $\mathbf{x}^t$ given $m(\mathbf{x}^t;\theta_m)$. Previous findings in economics and psychology have shown that people's decision-making process consists of multiple steps (Lunenburg 2010). Similarly, it has been suggested that AI-assisted decision-making may also involve multiple steps (Cao, Liu, and Huang 2024). Thus, consistent with previous studies, we divide the each type of DM into three steps: DM's initial judgment, the AI recommendation, and DM's aggregated decision.

Step 1: DM's initial judgment. In the first step, the human DM forms an independent judgment without AI assistance. This judgment is quantified by an independent decision model $h_k(\mathbf{x}^t;\theta_{hk})$, which we assume follows the form of a logistic model:

$$h_k(\mathbf{x}^t;\theta_{hk}) = \text{softmax}(\theta_{hk} \cdot \mathbf{x}^t)$$

This choice is consistent with previous work in well-established decision-making literature from economic research, where Logit models are widely used to model humans' independent decision-making, especially decisions under uncertainty (Chapman 1984; Lovreglio, Fonzone, and Dell'Olio 2016). Logit models and their variations are employed in modeling human behavior in advice-taking and AI-assisted decision-making as well (Tejeda et al. 2022; Li, Lu, and Yin 2024).

The DM's independent judgment on the task is then given by $\hat{y}_{hk}^t = \arg\max h(\mathbf{x}^t;\theta_{hk})$, with the confidence in this judgment being $c_{hk}^t = \max h_k(\mathbf{x}^t;\theta_{hk})$.

Step 2: AI model's recommendation. Given the AI model parameterized by θ_m, we can compute its recommendation consisting of two parts: the prediction $\hat{y}_m^t = \arg\max m(\mathbf{x}^t;\theta_m)$, with its confidence in this prediction being $c_m^t = \max m(\mathbf{x}^t;\theta_m)$.

Step 3: Aggregated decision. In the final step, the DM aggregates their own initial judgment and the AI recommendation to generate the final decision. Previous work in AI-assisted decision-making suggests that the cognitive process for DMs to aggregate their initial judgments and AI recommendations could further involve multiple stages

(Tejeda et al. 2022; Cao, Liu, and Huang 2024). Specifically, to model the aggregation process $g(\cdot)$ while recognizing the decision-maker's goal of maximizing overall utility, we characterize $g(\cdot)$ through the following three stages: confidence estimation, utility calculation, and action selection, as inspired by previous studies (Wang, Lu, and Yin 2022).

First, the DM estimates the likelihood of the AI recommendation being correct by aggregating the confidence of the DM's independent judgment $\hat{y}_h^t$ and the AI's recommendation $\hat{y}_m^t$ together. This is quantified as:

$$c_{h+m,k}^t = \begin{cases} \frac{1}{2}(c_{hk}^t + c_m^t) & \text{if } \hat{y}_{hk}^t = \hat{y}_m^t \\ \frac{1}{2}(1 - c_{hk}^t + c_m^t) & \text{if } \hat{y}_{hk}^t \neq \hat{y}_m^t \end{cases} \tag{2}$$

Intuitively, $c_{h+m,k}^t$ is an average of the DM's confidence in the AI recommendation and the AI's confidence in its recommendation. Higher $c_{h+m,k}^t$ indicates that the DM estimates the AI recommendation to be more likely correct after comparing it with the DM's independent judgment.

Next, in line with the expected utility theory (Schoemaker 1982), we assume the DM estimates the expected utility (EU) of accepting or rejecting the AI recommendation, incorporating a parameter β_k that represents their perceived penalty for making a wrong decision:

$$\begin{aligned} \hat{u}_{\text{accept},k}^t &= EU(\hat{y}^t = \hat{y}_m^t) = (1 + \beta_k)c_{h+m,k}^t - \beta_k \\ \hat{u}_{\text{reject},k}^t &= EU(\hat{y}^t \neq \hat{y}_m^t) = 1 - (1 + \beta_k)c_{h+m,k}^t \end{aligned} \tag{3}$$

After computing the utilities, the human DM needs to select an action to take. We consider the human DM will use a Logit model to compare among actions, assuming that humans are more likely to choose options with higher expected utility. Specifically, the probability for the human DM to accept the AI recommendation is given by a softmax function:

$$r_k^t = \frac{\exp(\delta_k \hat{u}_{\text{accept},k}^t)}{\exp(\delta_k \hat{u}_{\text{accept},k}^t) + \exp(\delta_k \hat{u}_{\text{reject},k}^t)} \tag{4}$$

where the parameter δ_k indicates the DM's sensitivity to utility differences. Such a Logit model is a widely-used model in economics to characterize people's discrete choices (Adeogun et al. 2008; Train 2009).

With the probability of DM accepting the AI recommendation r_k^t, we then model the action a_k^t of DM to accept or reject the AI recommendation with a Bernoulli distribution $a_k^t \sim \text{Bern}(r_k^t)$, i.e., $a_k^t = 1$ ($a_k^t = 0$) means the DM accepts (rejects) the AI recommendation. The final decision $\hat{y}_k^t$ is then determined by $\hat{y}_k^t = \mathbb{I}(a_k^t = 1)\hat{y}_m^t + \mathbb{I}(a_k^t = 0)(1 - \hat{y}_m^t)$.

In summary, the k-th decision process involves two sets of parameters: θ_{hk} captures how the DM forms their independent judgement, and $\theta_{ak} = \{\beta_k, \delta_k\}$ captures how the DM makes the aggregated decision. Together, the k-th decision process is quantified by the set of parameters $\Theta_k = \{\theta_{hk}, \theta_{ak}\}$.

Modeling the Mixture of K Types of DM With each type of DM parameterized by Θ_k and in total K types of DM, we define a parameter set $\Theta = \{\theta_{hk}, \theta_{ak}\}_{k=1}^K$ that characterizes a wide range of different DMs. Since the final decision is considered to be a mixture of the K types of DMs, the conditional probability of the final decision is:

$$\begin{aligned} &\mathbb{P}(\hat{y}^t \mid \mathbf{x}^t, m(\mathbf{x}^t; \theta_m), \Theta, \boldsymbol{Z}) \\ &= \sum_{k=1}^K z_k^t \cdot \mathbb{P}(\hat{y}^t | \mathbf{x}^t, m(\mathbf{x}^t; \theta_m), \Theta_k) \end{aligned} \tag{5}$$

where $\boldsymbol{Z}$ is a latent mixing coefficient matrix with element z_k^t indicating the responsibility of the k-th type of DM in a decision trial t.

Model Learning

Our objective is to learn M&M model given a training dataset of decision trials, and a set of DMs' final decisions $\mathcal{D} = \{\mathbf{d}^t, \hat{y}^t\}_{t=1}^T$ on these trials. Specifically, each decision trial $\mathbf{d}^t$ consists of the decision task $\mathbf{x}^t$ and the AI recommendation on this task $m(\mathbf{x}^t; \theta_m)$. In total, the parameter space of the model has two parts, the parameters of the K types of DMs Θ, and a mixing coefficient matrix $\boldsymbol{Z}$.

With a known $\boldsymbol{Z}$, learning the parameters Θ of the model given involves computing the posterior $P(\Theta \mid \mathcal{D})$. As direct computation is intractable, we leverage variational inference to approximate it using the parameterized distribution $q_\phi(\Theta)$. We aim to minimize the KL divergence between $q_\phi(\Theta)$ and $P(\Theta \mid \mathcal{D})$:

$$\begin{aligned} &\text{KL}(q_\phi(\Theta) \| P(\Theta)) = \int_\Theta q_\phi(\Theta) \log \frac{q_\phi(\Theta)}{P(\Theta \mid \mathcal{D})} \, \mathrm{d}\Theta \\ &= \int_\Theta q_\phi(\Theta) \left(\log \frac{q_\phi(\Theta)}{P(\Theta)} - \log P(\mathcal{D} \mid \Theta) + \log P(\mathcal{D}) \right) \mathrm{d}\Theta \\ &= \text{KL}(q_\phi(\Theta) \| P(\Theta)) - \text{E}_{q_\phi(\Theta)}[\log P(\mathcal{D} \mid \Theta) - \log P(\mathcal{D})] \end{aligned}$$

where $P(\Theta)$ is the prior distribution of Θ and $P(\mathcal{D})$ is a constant. Specifically, $q_\phi(\Theta_k)$ of the k-th type of DM consists of three variational distribution families:

1. For θ_{hk} (DM's independent judgment), we use a multivariate normal distribution: $\mathcal{N}(\theta_{hk}; \mu_\phi, \Sigma_\phi)$.
2. For $\theta_{ak} = \{\beta_k, \delta_k\}$ (DM's aggregation model), we use a Beta distribution for β (reflecting the bounded nature of the penalty parameter) and a normal distribution with a positive constraint for δ (reflecting the sensitivity to utility differences): $q(\beta) = \text{Beta}(A_\beta, B_\beta)$, $q(\delta) = \mathcal{N}(\mu_\delta, \sigma_\delta), \delta > 0$.

We use $\boldsymbol{\lambda}$ to denote all variational parameters in $q_\phi(\Theta)$.

However, the coefficient matrix $\boldsymbol{Z}$ is a latent variable unknown. Therefore, similar to the approximation of posterior of Θ, we again leverage a variational inference to approximate the distribution of $\{z_k^t\}_{k=1}^K$ using a parameterized distribution. Specifically, we use a Dirichlet distribution of order K, $\text{Dir}(\boldsymbol{\alpha^t})$, to model the responsibility of the K types of DMs in each decision trial t. Without further knowledge, we use a $\alpha_k^t = \frac{1}{K}$ as prior.

Due to the presence of the latent variables, we use the expectation maximization algorithm to optimize for the variational parameter space $(\boldsymbol{\lambda}, \boldsymbol{\alpha})$. Firstly, in the E-step, we

Dataset	# of Decision Tasks	# of Task Features	AI Model for Recommendation	Average AI Confidence
Diabetes Prediction	2130	6	Decision Tree Classifier	0.896
Loan-risk Assessment	7600	7	Random Forest Classifier	0.660
Income Prediction	1376	6	Wizard-of-Oz	0.749

Table 1: Summary of the datasets used in the evaluation.

calculate the posterior of each z_k^t based on the current estimate of parameters:

$$\begin{aligned}\gamma_k^t &= P(z_k^t \mid \mathbf{d}^t, \hat{y}^t, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \\ &= \frac{P(\hat{y}^t \mid \mathbf{d}^t, z_k^t, \boldsymbol{\lambda}_k) P(z_k^t \mid \boldsymbol{\alpha}^t)}{\sum_{k=1}^K P(\hat{y}^t \mid \mathbf{d}^t, z_k^t, \boldsymbol{\lambda}_k) P(z_k^t \mid \boldsymbol{\alpha}^t)}\end{aligned}$$

Then for the Maximization step, we search for optimal parameter values to maximize the auxiliary function Q, i.e., the expectation of the complete data log-likelihood:

$$\begin{aligned}Q(\boldsymbol{\lambda}, \boldsymbol{\alpha}) = &\sum_{t=1}^T \sum_{k=1}^K \gamma_k^t \log P(\hat{y}^t \mid \mathbf{d}^t, z_k^t, \boldsymbol{\lambda}_k) + \\ &\sum_{t=1}^T \sum_{k=1}^K \gamma_k^t \log P(z_k^t \mid \boldsymbol{\alpha}^t)\end{aligned}$$

In each M-step, we use gradient descent to update hidden parameters to the values that locally optimize Q.

Determining the number of DM types. In practice, the number of DM types (K) is not always accessible. Domain knowledge can sometimes provide the prior of K (e.g., known categories of doctors). However, for more general cases, we leverage the Bayesian Information Criterion (BIC) (Kuha 2004) to determine the optimal K:

$$\text{BIC} = -2 \ln \hat{L} + K \ln T$$

where $\hat{L}$ is the maximized likelihood of the model, K is the number of parameters (including those for each DM type), and T is the number of trials in the training dataset. We train models with varying K and select the model with the lowest BIC to balance between fit and complexity.

Model Inference

Given a new decision-making trial $\mathbf{d}^i$ with decision task $\mathbf{x}^i$ and AI recommendation $m(\mathbf{x}^i; \theta_m)$, we calculate the probability of the human DM accepting the AI recommendation and predict the DM's final decision on this trial as follows.

First, with the learned K types of DM, we aim to calculate the latent mixing coefficients of the K types of DMs corresponding to the decision trial. To obtain it, we need to first obtain the parameter set $\boldsymbol{\alpha}^i$ of the Dirichlet distribution parameters that generate the latent coefficients. As the direct computation is intractable, we use a heuristical method to estimate the parameters $\boldsymbol{\alpha}^i$. Intuitively, when two decision trials are similar, DMs are more likely to apply similar decision processes on them. That is, the influence of a training trial on the decision-making process of the current trial increases with its similarity to the current trial. Therefore, we approximate $\boldsymbol{\alpha}^i$ using kernel-weighted parameters:

$$\hat{\boldsymbol{\alpha}}^i = \sum_{t=1}^T K(\mathbf{d}^t, \mathbf{d}^i) \boldsymbol{\alpha}^t P(\boldsymbol{\alpha}^i)$$

where

$$K(\mathbf{d}^t, \mathbf{d}^i) = \frac{\exp(-s(\mathbf{d}^t, \mathbf{d}^i))}{\sum_{j=1}^T \exp(-s(\mathbf{d}^i, \mathbf{d}^j))}$$

$s(\cdot)$ is the Euclidean distance, and $P(\boldsymbol{\alpha}^i)$ is the prior of $\boldsymbol{\alpha}^i$. The mixing coefficient $\mathbf{z}^i$ is then obtained by averaging M samples from the distribution $\text{Dir}(\hat{\boldsymbol{\alpha}}^i)$, with $\mathbf{Z}^i = \frac{1}{M} \sum_{m=1}^M \mathbf{Z}^m$, $\mathbf{Z}^m \sim \text{Dir}(\boldsymbol{\alpha}^i)$. Finally, the DM's final decision in trial i is a weighted mixture calculated by Eq. 5.

Evaluation

In this section, we evaluate the effectiveness and generalizability of our proposed M&M framework.

Decision Tasks

In our evaluation, we consider three distinct real-world datasets encompassing diverse decision-making scenarios collected from previous empirical studies of AI-assisted decision-making (Wang, Lu, and Yin 2022; Li, Lu, and Yin 2024; Vodrahalli et al. 2022):

1. **Loan Risk Assessment:** This dataset focuses on the task of assessing loan default risk. Participants were presented with loan applicant profiles containing seven features: loan amount, interest rate, repayment period, monthly installment, annual income, credit score, and homeownership status. The AI model provided binary recommendations (default or not) along with confidence scores.
2. **Diabetes Prediction:** This dataset involves predicting diabetes in patients based on demographic and medical history data. Patient profiles included six features: gender, age, history of heart disease, Body Mass Index (BMI), HbA1c level, and blood glucose level. The AI model offered binary recommendations (diabetes or not) with confidence scores.
3. **Income Prediction:** The decision task in the dataset is to determine a person's annual income level. Given a profile of a person with seven features—the person's gender, age, education level, marital status, occupation, work type, and working hours per week—people were asked to decide whether this person's annual income is higher or lower than $50k$. The AI model provides its recommendations in the form of binary classification and the confidence score.

Treatment	Loan Risk Assessment			Diabete Prediction			Income Prediction		
	NLL ↓	Accuracy ↑	F1 ↑	NLL ↓	Accuracy ↑	F1 ↑	NLL ↓	Accuracy ↑	F1 ↑
Logistic Regression	0.515	0.601	0.740	0.446	0.713	0.744	0.889	0.815	0.896
Random Forest	0.721	0.602	0.715	0.472	0.692	0.738	0.999	**0.826**	0.903
MLP	0.665	0.599	0.724	0.554	0.734	0.751	0.704	0.757	0.854
SVM	0.558	**0.646**	0.651	0.461	0.754	0.758	0.958	0.652	0.753
CPT Utility	0.542	0.611	0.644	0.546	0.633	0.725	0.784	0.758	0.863
Confidence Threshold	-	0.600	0.656	-	0.629	0.723	-	0.736	0.845
M&M (Ours)	**0.491**	0.632	**0.774**	**0.413**	**0.770**	**0.762**	**0.656**	0.805	**0.913**

Table 2: Comparing the performance of the proposed method with baseline methods on three decision-making tasks 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.

These datasets were preprocessed to ensure consistency and suitability for our analysis. For all datasets, we converted the human decisions and AI recommendations into binary format (0 or 1) and normalized the AI confidence scores to the range of [0, 1] to facilitate comparison. Table 1 provides the summary of the datasets.

Examining the Predictive Performance of M&M

We first examine how well the M&M framework can predict human DMs' final decisions in AI-assisted decision making.

Evaluation Setup For each dataset, we randomly split the data into training (80%) and test (20%) sets. To quantify model performance, we employed three key evaluation metrics: negative log-likelihood (NLL), accuracy, and F1-score. NLL measures the model's ability to predict the probability of observed human decisions, with lower values indicating better performance. Accuracy assesses the proportion of correct predictions, while the F1-score provides a balanced measure of precision and recall, capturing the model's ability to correctly identify both acceptance and rejection of AI recommendations. To ensure the robustness of evaluations, all experiments were repeated 5 times, and the average performance across these repetitions was reported.

Our M&M model is trained using a Bayesian approach with variational inference, as outlined in the previous section. We experiment with different numbers of DM types ($K \in \{2, 3, \ldots, 6\}$) and select the K that achieves the minimum BIC score for each task. To provide a robust benchmark for evaluating the performance of our proposed M&M framework, we consider three distinct classes of baseline models, each capturing different aspects of human decision-making behavior in AI-assisted scenarios:

1. *Standard Supervised Learning Models:* We employ four widely-used supervised learning models: Logistic Regression, Random Forest, Multi-Layer Perceptron (MLP), and Support Vector Machine (SVM). These models directly predict the human DM's final decision $\hat{y}^t$ in a decision task based on task features $\mathbf{x}^t$, AI recommendations $\hat{y}_m^t$, and AI confidence scores c_m^t. These models serve as a baseline for predictive accuracy, allowing us to assess whether incorporating explicit modeling of human-AI interaction patterns can improve upon standard machine learning approaches.
2. *Utility-Based Model:* We adapt the model proposed by (Wang, Lu, and Yin 2022), which is grounded in Cumulative Prospect Theory (CPT). This model assumes that DMs assess the utility of accepting or rejecting AI recommendations based on a distorted perception of probabilities, as captured by CPT's probability weighting function $w(p) = \frac{p^k}{p^k+(1-p)^k}$ $(k > 0)$. Based on this distorted estimate, the DM computes the utility of accepting or rejecting the AI recommendation as $U = w(p) \cdot \text{gain} + w(1-p) \cdot \text{loss}$. With calculated utility, the DM then use a Logit model to select the action to accept or reject the AI recommendation.
3. *Confidence Threshold Model:* We include the confidence-based model used in (Amin, Lu, and Yin 2024), which posits that human DMs have an internal confidence threshold τ drawn from a distribution $f(\tau)$. If their confidence in their own judgment exceeds this threshold, they reject the AI recommendation; otherwise, they accept it. Practically, we use a Beta distribution $q(\tau) = \text{Beta}(A_\tau, B_\tau)$ to approximate the distribution $f(\tau)$, with the constraint $\tau \in (0, 1)$. This model serves as a simple yet effective baseline that captures the role of self-confidence in AI-assisted decision-making.

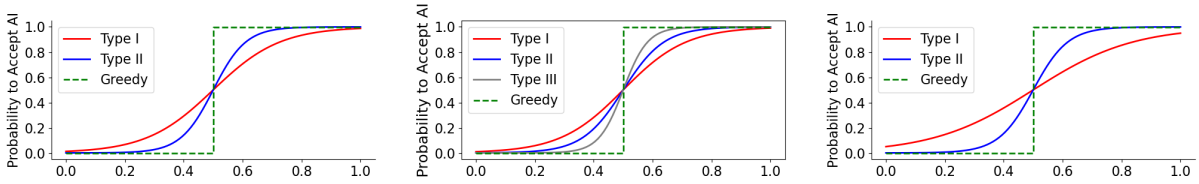
Evaluation Results. Table 2 presents the performance comparison of multiple models in predicting DM's decisions across varied datasets. Overall, our proposed M&M framework consistently emerges as the best-performing model with respect to NLL and F1 score. In terms of accuracy, our method performs the best in diabetes predictions and is comparable to the top-performing models in the other two decision tasks.

Quantifying Heterogeneity in Human DMs

Beyond prediction, the M&M framework offers a nuanced understanding of the heterogeneous nature of human decision-making in AI-assisted contexts. By explicitly modeling diverse DM types and their associated parameters, the proposed framework provides insights into the underlying fac-

Task	k-th Type of DM	Parameters		
		Perceived Penalty (β)	Sensitivity (δ)	Percentage in Population (α)
Diabetes Prediction	Type I	0.81	2.41	0.26
	Type II	0.92	4.72	0.74
Loan Risk Assessment	Type I	0.44	3.24	0.29
	Type II	0.52	4.73	0.38
	Type III	0.66	7.25	0.33
AI-assisted Income Prediction	Type I	0.61	1.83	0.13
	Type II	0.93	4.40	0.87

Table 3: Comparisons between model parameters learned for the identified types of DMs across three datasets.



(a) Diabetes Prediction (b) Loan Risk Assessment (c) Income Prediction

Figure 2: Comparing decision behavior across identified DM types in the three AI-assisted decision-making tasks. The plots show the probability of a DM accepting the AI recommendation given the aggregated confidence. The green dashed line represents a "greedy" decision maker who always accepts AI recommendations with confidence levels above 0.5.

tors influencing human decisions. There are three key parameters learned by the model in our current setup. Perceived penalty (β) reflects the DM's aversion to incorrect decisions, with higher values indicating greater risk aversion. Sensitivity (δ) captures the DM's responsiveness to changes in the utility of accepting or rejecting AI recommendations. Population proportion ($\boldsymbol{\alpha}$) represents the prevalence of each DM type within the overall population. Table 3 presents the learned parameters for the different types of DMs identified in varied AI-assisted decision-making scenarios. Analysis of these parameters yields several important insights.

Decision context significantly influences DM behavior. We find that the decision context significantly influences the distribution of DM types. The proportion of each DM type, as indicated by the α values, varies significantly across tasks. For example, in the diabetes prediction and income prediction tasks, Type II DMs, characterized by higher perceived penalty aversion and sensitivity to utility change, constitute the majority (α = 0.74 and α = 0.87 for the two tasks respectively) of the DMs. However, differences in proportions across DM types are less prominent in the loan risk assessment task, showing that the risk attitude tends to be uniformly distributed among DMs in this decision task. This suggests that the specific nature of decision context can influence the decision-making style adopted by individuals, highlighting the importance of considering context when designing AI systems that aim to assist human DMs.

Task complexity and risk perception shape diversity in DM types. The number of identified DM types and the absolute values of the perceived penalty (β) parameter vary across tasks, reflecting differences in task complexity and the granularity of risk perception. The loan risk assessment task, with its three distinct DM types and relatively lower β values (ranging from 0.44 to 0.66), suggests a more nuanced understanding of risk among DMs due to the availability of detailed information. In contrast, diabetes and income prediction tasks, with only two DM types each and higher β values, may reflect simpler risk assessments or less available information to mitigate potential losses. This observation underscores the need for flexible and adaptable AI systems that can cater to varying levels of task complexity and individual risk perceptions.

Risk preference could act as a moderator of decision sensitivity. A consistent positive correlation is observed between perceived penalty and sensitivity across DM types within each task. For example, in the loan risk assessment task, Type III DMs, with the highest perceived penalty (β = 0.66), also demonstrate the highest sensitivity (δ = 7.25). This finding implies that people who are more averse to incorrect outcomes (higher β) are more likely to engage in analytical decision-making processes, carefully considering the potential consequences of their choices (higher δ). This aligns with economic theories that highlight the role of perceived risk in decision strategies under uncertainty(Kim, Menzefricke, and Feinberg 2007; Train 2009).

Figure 2 further illustrates these behavioral differences by plotting the probability of accepting AI recommendations against the aggregated confidence level for each DM type. For instance, the curves for Type II DMs consistently rise more steeply than those for Type I DMs in the interval close to 0.5, suggesting changes in aggregated confidence around 0.5 will lead to a higher change chance of decision change for Type II DMs. Furthermore, the acceptance probability of Type I DM is consistently higher than that of Type II DM when confidence is lower than 0.5, indicating that Type I DMs are generally more trusting of AI recommendations and require lower confidence levels to accept them. This observation aligns with their lower perceived penalty and sen-

sitivity values, suggesting a less risk-averse decision making style. In contrast, Type II DMs exhibit a more cautious approach, requiring higher levels of confidence before accepting AI recommendations. In the loan risk assessment task, the presence of a third DM type with even higher perceived penalty and sensitivity further emphasizes the diversity of human behavior in AI-assisted decision-making scenarios.

Inferring Human DM's Independent Judgment

An additional benefit of our proposed framework is its ability to infer independent human judgment without relying on explicitly labeled data. This can address a crucial limitation in prior research that employs a separate model trained exclusively to characterize the human DM's independent judgment, a process that can be resource-intensive and may not be feasible in all scenarios. To validate the efficacy of M&M in inferring human DM's independent judgment, we leveraged the pilot study data from the loan risk assessment and diabetes prediction datasets. These previously conducted studies used pilot studies to collect data for human DMs reviewing tasks and making judgments without AI assistance, providing a ground truth for initial DM judgment.

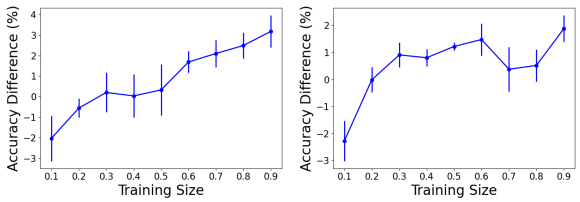
Evaluation Setup. This analysis involved the pilot study data collected in the loan risk assessment and diabetes prediction tasks. Specifically, we split the pilot data into training and test sets, gradually increasing the training size from 10% to 90% of the entire pilot data. For each split, we trained an independent human decision model, as done in previous studies, and evaluated its accuracy on the test set (p_{idp}). Specifically, we use a random forest classifier for predictions in loan risk assessment tasks and logistic regression with diabetes data, following the approach taken by the work that collected the data.

We then used the independent human decision models $\{\theta_{hk}\}_{k=1}^{K}$ inferred by M&M on AI-assisted data to predict DMs' initial judgments on the test set of pilot data. Similar to Equation 5, we use a weighted mixture of independent human decision models to predict the initial DM judgment on a specific data instance $\mathbf{x}^i$:

$$\hat{y}^i = \text{argmax}(\sum_{k=1}^{K} z_k^i h_k(\mathbf{x}^i; \theta_{hk}))$$

and we evaluated its accuracy (p_{inf}). Finally, we compared p_{inf} with p_{idp} across different splits, to evaluate M&M's ability to accurately capture independent human judgment in scenarios with varying amounts of training data.

Evaluation Results. Figure 3 presents the comparison between the accuracy of independent human decision model trained on actual data of initial DM judgment without AI assistance (p_{idp}), and the accuracy of the initial human judgments inferred by M&M (p_{inf}). We generally find the accuracy difference ($p_{idp} - p_{inf}$) between the model trained on independent judgment data and our proposed model inferring independent judgments to be small, indicating comparable usefulness of the M&M model without the need for additional data collection. Interestingly, when the training set size is smaller, indicating a scarcity of training data for the independent model, the accuracy difference is actually negative, i.e., the independent human judgment model inferred from data of AI-assisted prediction ends up outperforming the model trained on the independent human judgment dataset. These findings underscore M&M's potential to unlock insights into human decision making in situations where independent judgment data is limited or unavailable.



(a) Loan Risk Assessment (b) Diabetes Prediction

Figure 3: Accuracy difference between the human decision model trained on the training set and the model inferred from AI-assisted decision-making data on the test set of human judgment data. Error bars depict standard error of the mean.

Conclusion

In this work, we present Mix and Match (M&M), a novel computational framework that models the heterogeneous nature of human decision-making under AI assistance as a mixture of distinct decision-making processes. M&M acknowledges variations across different individuals and recognizes that the same individual may adopt different decision-making processes for different tasks. Our empirical evaluation on real-world data across three distinct scenarios demonstrates that the M&M framework consistently outperforms baseline methods in predicting human decisions under AI assistance. Notably, the framework infers independent human judgment without the need for additional training data. Moreover, by analyzing the learned parameters of different DM types, we uncover nuanced behavioral patterns that align with established psychological theories and reveal context-dependent variations in decision-making styles. By unfolding the interplay between human intuition and AI recommendations, the M&M framework paves the way for the development of more effective, personalized, and trustworthy AI systems that can more effectively empower human DMs.

It is still important to acknowledge that this study has limitations. The human behavior data used for evaluation were collected from laypeople on predictive tasks based on tabular data with relatively few features. Whether the proposed model generalizes to tasks with higher-dimensional feature spaces or greater complexity remains to be investigated. Furthermore, the AI-assisted scenarios examined explicitly provided confidence values to human DMs. Future research should explore the applicability of the M&M framework to scenarios where AI models communicate confidence implicitly, such as through verbal descriptions in large language models. Finally, we assumed a logistic regression model for independent human judgment, and exploring alternative models could further enhance the framework's flexibility.

References

Adeogun, O.; Ajana, A.; Ayinla, O.; Yarhere, M.; and Adeogun, M. 2008. Application of logit model in adoption decision: A study of hybrid clarias in Lagos State, Nigeria. *American-Eurasian Journal of Agriculture and Environmental Sciences*, 4(4): 468–472.

Allais, M. 1953. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, 503–546.

Amin, H.; Lu, Z.; and Yin, M. 2024. Designing Behavior-Aware AI to Improve the Human-AI Team Performance in AI-Assisted Decision Making. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*.

Appelt, K. C.; Milch, K. F.; Handgraaf, M. J.; and Weber, E. U. 2011. The Decision Making Individual Differences Inventory and guidelines for the study of individual differences in judgment and decision-making research. *Judgment and Decision making*, 6(3): 252–262.

Askarisichani, O.; Bullo, F.; Friedkin, N. E.; and Singh, A. K. 2022. Predictive models for human–AI nexus in group decision making. *Annals of the New York Academy of Sciences*, 1514(1): 70–81.

Bansal, G.; Nushi, B.; Kamar, E.; Horvitz, E.; and Weld, D. S. 2021a. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11405–11414.

Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W. S.; Weld, D. S.; and Horvitz, E. 2019a. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, 2–11.

Bansal, G.; Nushi, B.; Kamar, E.; Weld, D. S.; Lasecki, W. S.; and Horvitz, E. 2019b. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2429–2437.

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 2021b. 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*, 1–16.

Cao, L. 2022. Ai in finance: challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)*, 55(3): 1–38.

Cao, S.; Liu, A.; and Huang, C.-M. 2024. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–32.

Chapman, R. G. 1984. An approach to estimating logit models of a single decision maker's choice behavior. *Advances in Consumer Research*, 11(1).

De Mantaras, R. L.; and Arcos, J. L. 2002. AI and music: From composition to expressive performance. *AI magazine*, 23(3): 43–43.

Franken, I. H.; and Muris, P. 2005. Individual differences in decision-making. *Personality and Individual Differences*, 39(5): 991–998.

Guo, Z.; Wu, Y.; Hartline, J. D.; and Hullman, J. 2024. A Decision Theoretic Framework for Measuring AI Reliance. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 221–236.

Hartford, J. S.; Wright, J. R.; and Leyton-Brown, K. 2016. Deep learning for predicting human strategic behavior. *Advances in neural information processing systems*, 29.

Kim, J. G.; Menzefricke, U.; and Feinberg, F. M. 2007. Capturing flexible heterogeneous utility curves: A Bayesian spline approach. *Management Science*, 53(2): 340–354.

Kuha, J. 2004. AIC and BIC: Comparisons of assumptions and performance. *Sociological methods & research*, 33(2): 188–229.

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.; Liao, Q. V.; Smith-Renner, A.; and Tan, C. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*.

Lai, V.; and Tan, C. 2019. 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*, 29–38.

Li, Z.; Lu, Z.; and Yin, M. 2023. Modeling human trust and reliance in ai-assisted decision making: A markovian approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 6056–6064.

Li, Z.; Lu, Z.; and Yin, M. 2024. Decoding AI's Nudge: A Unified Framework to Predict Human Behavior in AI-assisted Decision Making. *arXiv preprint arXiv:2401.05840*.

Lovreglio, R.; Fonzone, A.; and Dell'Olio, L. 2016. A mixed logit model for predicting exit choice during building evacuations. *Transportation Research Part A: Policy and Practice*, 92: 59–75.

Lu, Z.; and Yin, M. 2021. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Lunenburg, F. C. 2010. The decision making process. In *National Forum of Educational Administration & Supervision Journal*, volume 27.

Lysaght, T.; Lim, H. Y.; Xafis, V.; and Ngiam, K. Y. 2019. AI-assisted decision-making in healthcare: the application of an ethics framework for big data in health and research. *Asian Bioethics Review*, 11: 299–314.

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. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19.

Matthews, G.; Lin, J.; Panganiban, A. R.; and Long, M. D. 2019. Individual differences in trust in autonomous robots: Implications for transparency. *IEEE Transactions on Human-Machine Systems*, 50(3): 234–244.

Montgomery, H. 1983. Decision rules and the search for a dominance structure: Towards a process model of decision making. In *Advances in psychology*, volume 14, 343–369. Elsevier.

Pynadath, D. V.; Wang, N.; and Kamireddy, S. 2019. A Markovian Method for Predicting Trust Behavior in Human-Agent Interaction. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, 171–178.

Rechkemmer, A.; and Yin, M. 2022. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–14.

Schoemaker, P. J. 1982. The expected utility model: Its variants, purposes, evidence and limitations. *Journal of economic literature*, 529–563.

Schoemaker, P. J. 2013. *Experiments on decisions under risk: The expected utility hypothesis*. Springer Science & Business Media.

Shaheen, M. Y. 2021. Applications of Artificial Intelligence (AI) in healthcare: A review. *ScienceOpen Preprints*.

Sharan, N. N.; and Romano, D. M. 2020. The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, 6(8).

Steyvers, M.; and Kumar, A. 2023. Three challenges for ai-assisted decision-making. *Perspectives on Psychological Science*, 17456916231181102.

Strickland, L.; Farrell, S.; Wilson, M. K.; Hutchinson, J.; and Loft, S. 2024. How do humans learn about the reliability of automation? *Cognitive Research: Principles and Implications*, 9(1): 8.

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.

Train, K. E. 2009. *Discrete choice methods with simulation*. Cambridge university press.

Tversky, A.; and Kahneman, D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4): 297–323.

Vodrahalli, K.; Daneshjou, R.; Gerstenberg, T.; and Zou, J. 2022. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 763–777.

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.

Wang, Y.; and Ruhe, G. 2007. The cognitive process of decision making. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 1(2): 73–85.

Yang, F.; Huang, Z.; Scholtz, J.; and Arendt, D. L. 2020. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 189–201.

Yin, M.; Wortman Vaughan, J.; and Wallach, H. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–12.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. 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*, 295–305.