Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making

Xinru Wang*
Purdue University
West Lafayette, Indiana, USA
xinruw@purdue.edu

Zhuoran Lu*
Purdue University
West Lafayette, Indiana, USA
lu800@purdue.edu

Ming Yin
Purdue University
West Lafayette, Indiana, USA
mingyin@purdue.edu

ABSTRACT

Internet users make numerous decisions online on a daily basis. With the rapid advances in AI recently, AI-assisted decision making—in which an AI model provides decision recommendations and confidence, while the humans make the final decisions—has emerged as a new paradigm of human-AI collaboration. In this paper, we aim at obtaining a quantitative understanding of whether and when would human decision makers adopt the AI model's recommendations. We define a space of human behavior models by decomposing the human decision maker's cognitive process in each decision-making task into two components: the *utility* component (i.e., evaluate the utility of different actions) and the *selection* component (i.e., select an action to take), and we perform a systematic search in the model space to identify the model that fits real-world human behavior data the best. Our results highlight that in AI-assisted decision making, human decision makers' utility evaluation and action selection are influenced by their own judgement and confidence on the decision-making task. Further, human decision makers exhibit a tendency to distort the decision confidence in utility evaluations. Finally, we also analyze the differences in humans' adoption behavior of AI recommendations as the stakes of the decisions vary.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; **Empirical studies in HCI**.

KEYWORDS

AI-Assisted Human Decision Making, Behavior Model, Human-Subject Experiments

ACM Reference Format:
Xinru Wang, Zhuoran Lu, and Ming Yin. 2022. Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3485447.3512240

*Wang and Lu have made equal contributions to this work.

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04.
https://doi.org/10.1145/3485447.3512240

1 INTRODUCTION

Internet users engage in many decision making activities online everyday, ranging from making investment choices for themselves, to evaluating news veracity for the community, to annotating biomedical images for citizen science projects. Recently, many AI-driven decision aids have been developed to support human decision making, and the widespread usage of these decision aids has created a new paradigm of human-AI collaboration, i.e., *AI-assisted decision making*—that is, given a decision-making task, an AI-based decision aid provides a decision recommendation, while the human decision maker makes the final decision. For example, an investor may be advised by an AI-powered trading tool to buy and sell stocks online, and a citizen scientist may get suggestions from a deep learning model to detect diabetic retinopathy from images of the retina.

To fully unlock the potential of AI-driven decision aids in assisting people to make better decisions, it is critical to obtain a solid understanding of how human decision makers react to the decision recommendations provided by the AI. In particular, how do humans decide whether to trust the AI model and adopt its recommendation or not in a decision-making task? To this end, there is a growing line of experimental studies in the human-computer interaction community which empirically identifies a wide range of factors that can influence people's trust in AI, such as the AI model's accuracy [20, 35, 45], the model's confidence in the decision-making task [35, 47], and the level of agreement between the human and the AI [26]. However, attempts in probing deeper into the mechanisms that govern how these factors interact to influence humans' adoption of AI recommendations is quite limited. This implies a missing opportunity for obtaining more comprehensive and quantitative understandings of human behavior in AI-assisted decision making, which may inform theory development of human cognitive processing. Moreover, due to the limited quantitative understandings of humans' adoption of AI recommendations, existing studies in the AI community on optimizing human-AI joint decision making performance (e.g., [3]) often make simplistic assumptions about how humans interact with the AI model (e.g., assume that humans will always accept the AI's recommendation when the AI model's confidence is above a threshold). Building computational models of *real* humans' adoption behavior of AI recommendations in AI-assisted decision making, thus, can potentially be helpful for redesigning the AI to take the realistic humans' reactions into consideration to enhance human-AI collaborations.

Therefore, in this study, we focus on a basic AI-assisted decision making setting and search for the best computational model to characterize human decision makers' behavior of adopting the AI recommendation in such a setting. Specifically, in this setting, the human decision maker is asked to solve a sequence of binary

decision-making trials with the assistance of an AI model. On each trial, the AI model provides its binary decision recommendation as well as its confidence in this recommendation to the human decision maker. Then, the human decision maker can make her final decision by either accepting the model's recommendation or rejecting it, and depending on whether this final decision is correct or not, the human decision maker will receive some reward or penalty. The human decision maker's objective is to maximize her cumulative utility in all decision-making trials, though she will not receive any feedback on the correctness of her final decisions while she makes those decisions.

To characterize whether the human decision maker will adopt the AI recommendation in each trial, we propose a *space* of human behavior models by decomposing the human's cognitive reasoning process in each trial into two components: the *utility* component for evaluating the utility of different actions (i.e., accept or reject AI), and the *selection* component for stochastically determining an action to take. Within each component, we first define a few basic models to capture how the human decision maker computes the utility of each action based on the AI model's confidence in its recommendation, or how she turns the estimated utility of an action into the probability of taking that action. To reflect the possibility that the human decision maker may take her own judgement on a decision-making trial as well as her confidence in this judgement into account when determining whether to adopt the AI recommendation, we further define a set of *human-adjusted* utility models and *human-adjusted* selection models.

To explore which combinations of the two-component models can best capture real-world human decision makers' behavior of adopting AI recommendations, we collect data on real human subjects' decisions in AI-assisted loan risk assessment tasks through a randomized experiment. We also vary the stakes of the decisions (i.e., reward/penalty associated with correct/incorrect decisions) in different treatments of this experiment to simulate different decision making environments. We find that the two-component models using both human-adjusted utility models and human-adjusted selection models outperform the combinations of basic utility and selection models, in explaining real decision makers' adoption of AI assistance. This indicates that humans tend to aggregate their own opinions with the AI's advice when making AI-assisted decisions. The best-performing two-component behavior model suggests that human decision makers tend to apply a weighting function to interpret probability estimates of the AI model's recommended decision being correct. Moreover, humans also tend to increase their likelihood of accepting the AI recommendation if they agree with the recommendation with high confidence, and decrease if they disagree. Finally, comparing the best-performing models for explaining humans' behavior under different levels of decision stakes, we find that when the decision stakes are higher, people tend to lower their belief in the AI model being correct, and are more inclined to rely on their own judgements on the decision-making trial when choosing whether to accept or reject the AI recommendation.

Together, these results provide a useful starting point for quantitatively characterizing humans' adoption behavior of AI assistance, which can potentially serve as the foundation for developing more effective AI-driven decision aids that are aware of real-world human behavior in the future.

2 RELATED WORK

Empirical Studies in AI-Assisted Decision Making. Many empirical studies have been carried out to explore whether people are willing to trust AI in AI-assisted making, and which factors will influence people's trust. Beyond the exploration into how trust in AI is impacted by some most straight-forward factors such as the performance indicators of the AI model [20, 35, 45, 47], the recent surge of interests in increasing the interpretability of AI (e.g., [16, 18, 36]) has led to an increasing number of evaluations on understanding whether and how AI explanations impact people's trust in the AI model [23, 31, 41, 47]. Most recently, researchers have started to study people's trust in the AI model under some special conditions, such as when the distribution shift occurs [8, 23].

Modeling Human Decision Making. Understanding how people make decisions is a central problem in psychology and economics. Various theoretical frameworks have been proposed to explain human decision making behavior under uncertainty. One of the earliest frameworks is the expected utility model, which is based on the hypothesis that individuals always choose the options that maximize their expected utilities [40]. However, the recurring observations of people deviating from the optimal decisions lead to the development of many new theories. For example, one of the alternatives is the random utility model [27], which states that the utility of an option is composed of an observable part (e.g., expected utility) and an unobservable stochastic error term. Another alternative is the prospect theory (PT), with the most popular modern variant of it being the cumulative prospect theory (CPT) [11, 24, 39]. Studies have also been carried out to model human behavior in settings where agents have to *repeatedly* make decisions under uncertainty [2, 44]. More recently, data-driven approaches using machine learning to predict human decisions have been explored to complement the theory-driven approaches [12, 29, 30].

Modeling Interactions between Humans and AI/Automation. With the increased use of AI-driven assistive tools in decision making, a growing line of research on designing AI models that can optimize the human-AI team decision-making performance has emerged. Many of these studies aim to exploit the human-machine complementarity by finding an effective "division of labor" between the AI and humans, either by training the algorithmic model in a supervised way to leverage the distinctive strengths of both machines and humans [9, 34], or by explicitly routing the tasks to the appropriate party through formulating the problem in a bandit feedback setting [2, 13]. Most recently, some researchers (e.g., [3]) also looked into the problem of optimizing human-AI team performance in an AI-assisted decision making setting in which humans are always the final decision maker. However, these studies often make overly-simplified assumptions on how humans will interact with the AI (e.g., accept the AI recommendation if that maximizes her expected utility), despite the reasoning processes underlying humans' decisions to trust an AI or not are highly sophisticated [4, 5, 19].

In addition, various works directly model human trust in automation. For example, studies in the field of human-robot interaction provided insights on how a human's trust changes as a human agent interacts with a robotic agent over time, by analyzing the autonomy's performance, the human agent's behavior, and real-time physiological signals [14, 15, 17, 25]. Common approaches

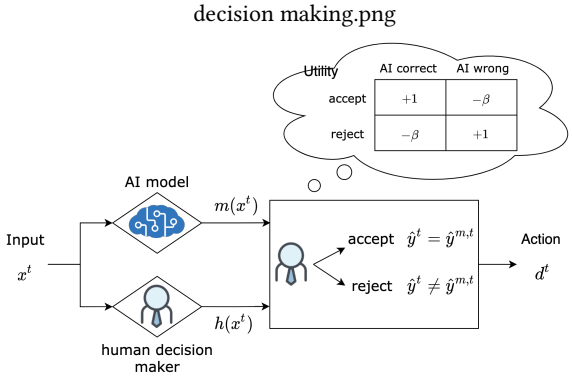


Figure 1: The AI-assisted decision making setting we have studied in this paper.

used in these studies are mostly data-driven, including time series model [22], machine learning techniques [25], and Bayesian inference framework [42], with limited effort spent on providing theory-grounded explanations to the learned models.

Different from the earlier research, in this paper, we focus on computationally exploring how real humans behave when assisted by an AI model, and we borrow theories from behavioral economics to construct elements of human behavior models.

3 PROBLEM DESCRIPTION

We now formally describe the AI-assisted decision making setting that we have studied in this paper (see Figure 1 for an overview diagram). Suppose a decision-making trial can be characterized by an n-dimensional feature vector x (i.e., $x \in R^n$), and y is the correct decision to make in this trial. In this study, we focus on decision-making tasks with binary choices of decisions, i.e., $y \in \{+1, -1\}$. We use $m(x)$ to denote the AI model's output on the decision-making trial, which is a probability distribution over the set of possible decisions, i.e., $m(x) = \{+1 : P(y = +1|x), -1 : P(y = -1|x)\}$. Given $m(x)$, the AI model will make a decision recommendation to the human decision maker, which is composed of two parts—the recommended binary decision $\hat{y}^m = \arg\max m(x)$, and the confidence in its recommended decision $c^m = \max m(x) = m(x)[y = \hat{y}^m]$. We assume that the AI model's confidence is *calibrated*, that is, $c^m = P(y = \hat{y}^m)$. Similarly, we assume that the human decision maker will also form her own judgement on the decision-making trial—$h(x)$ is used to characterize the human's output on the decision-making trial, $\hat{y}^h = \arg\max h(x)$ is the human's binary decision, and $c^h = \max h(x) = h(x)[y = \hat{y}^h]$ is the human's confidence on her decision. Finally, we'll also use $h(x)[y = +1]$ and $h(x)[y = -1]$ later to refer to the human decision maker's confidence on a specific decision.

Now suppose the human decision maker is asked to complete a sequence of T decision-making trials with the help of the AI model. In each trial t ($1 \le t \le T$), the human decision maker is provided with the feature vector x^t, along with the AI model's binary recommendation $\hat{y}^{m,t}$ and confidence $c^{m,t}$. She also develops her own judgement $h(x^t)$ on the trial. With all these information, the human decision maker needs to make a final decision $\hat{y}^t$ by taking an action $d^t \in \{accept, reject\}$ to either accept the AI model's binary decision recommendation or reject it. That is, when $d^t = accept$, $\hat{y}^t = \hat{y}^{m,t}$, otherwise $\hat{y}^t = -\hat{y}^{m,t}$. The human decision maker is informed that based on the correctness of her final decision in each trial, she could get different utility U^t—if her final decision is correct (i.e., $\hat{y}^t = y^t$, y^t is the correct decision for trial t), she will receive a reward of 1 (i.e., $U^t = 1$); otherwise, she will receive a penalty of β (i.e., $U^t = -\beta$). The human decision maker's objective, thus, is to maximize her cumulative utility $\sum_{t=1}^{T} U^t$ in all T trials.

In this study, we focus on the scenario that the human decision maker will not receive immediate feedback after each trial on whether her final decision in trial t is correct or not. The goal of our study, then, is to quantitatively characterize how the human decision maker chooses to adopt the AI model's decision recommendation or not (i.e., determines d^t) in each trial t.

4 MODELS

In this section, we outline the set of computational models for modeling human decision makers' adoption behavior of the AI recommendations in AI-assisted decision making. Specifically, we propose a *space* of models by decomposing the human decision maker's cognitive reasoning processes in each decision-making trial into two components (see Figure A1 in Appendices for a diagram of the model space)—the *utility* component, in which the decision maker evaluates the utility of different actions, and the *selection* component, in which the decision maker stochastically selects an action to take based on the estimated utility of each action.

4.1 Utility Component

The utility component characterizes how the human decision maker estimates the utility u_j^t of each action $j \in \{accept, reject\}$ in a decision-making trial. We first consider a few basic utility models in which the human decision maker infers the utility of each action only based on the AI model's output, that is, $\hat{u}_j^t = f(m(x^t))$. In addition, we conjecture that the human decision maker's own judgement on the decision-making trial may also influence her perceived utility of each action—for example, when the human decision maker's own binary decision is different from that of the model's, she may decrease her estimated utility of the action of accepting the AI model's recommendation. To reflect this possibility, we further propose a few *human-adjusted* utility models to capture humans' possible behavior of aggregating both her own judgement and the AI model's recommendation to evaluate the utility of each action, i.e., $\hat{u}_j^t = f(m(x^t), h(x^t))$.

4.1.1 *Basic Utility Models.*

We consider two basic utility models:

- **Expected Utility (EU)**: In this model, aligning with the expected utility theory, we assume the human decision maker estimates the utility of an action as her *expectation* of the utility that she would receive by taking that action. Since we assume the AI model's confidence is calibrated, the AI model's confidence $c^{m,t}$ on the decision-making trial t effectively reflects the probability that the model's binary decision $\hat{y}^{m,t}$ is correct. Thus, we have:

$$\hat{u}^t_{accept} = EU(\hat{y}^t = \hat{y}^{m,t}) = (1+\beta)c^{m,t} - \beta$$
$$\hat{u}^t_{reject} = EU(\hat{y}^t = -\hat{y}^{m,t}) = 1 - (1+\beta)c^{m,t}$$

- **CPT-Based Utility (CPTU)**: In behavioral economics, the Cumulative Prospect Theory (CPT) [11, 24, 39] provides a generalization of the expected utility theory to explain humans' irrational behavior in decision making. In particular, a key observation of CPT is that people tend to interpret probabilities in a non-linear way such that there exists a *probability weighting function* to transform the objective cumulative probabilities into subjective cumulative probabilities. Further, CPT states that people tend to overweight extreme events, but underweight "average" events (i.e., the probability weighting function has an inverse-S shape), which is consistent with the rank-dependent expected utility theory [24, 33]. In this study, following the earlier literature [21, 32, 38], we adopt a probability weighting function $w(p) = \frac{p^k}{p^k+(1-p)^k}$ where $k > 0$ is a parameter controlling the shape of the function. When $0 < k < 1$, the probability weighting function takes an inverse-S shape, and the smaller k is, the more the probabilities are distorted. In contrast, when $k > 1$, the weighting function takes an S-shape (i.e., underweight the extreme events while overweight the average events). Finally, when $k = 1$, the weighting function is the linear function $w(p) = p$, and the CPT-based utility becomes effectively the same as the expected utility. Note here, we have $w(p) = 1 - w(1-p)$. When human decision makers apply the probability weighting function to interpret the AI model's confidence $c^{m,t}$, we have:

$$\hat{u}^t_{accept} = (1+\beta)w(c^{m,t}) - \beta$$
$$\hat{u}^t_{reject} = 1 - (1+\beta)w(c^{m,t})$$

4.1.2 Human-adjusted Utility Models. In human-adjusted utility models, we attempt to capture the possibility that the human decision maker aggregates her own judgement on the decision-making trial, $h(x^t)$, with the AI model's outputs, $m(x^t)$, when evaluating the utility of different actions. Specifically, given the AI model's binary decision recommendation $\hat{y}^{m,t}$, the human decision maker's belief in how likely this recommended decision is correct is captured by $h(x^t)[y^t = \hat{y}^{m,t}]$—when the decision maker's own binary decision is the same as that of the AI model's (i.e., $\hat{y}^{h,t} = \hat{y}^{m,t}$), we have $h(x^t)[y^t = \hat{y}^{m,t}] = c^{h,t}$; otherwise, $h(x^t)[y^t = \hat{y}^{m,t}] = 1 - c^{h,t}$. Given both her own and the AI model's probabilistic estimates $h(x^t)[y^t = \hat{y}^{m,t}]$ and $c^{m,t}$, the human decision maker next needs to combine them to produce an aggregate estimate $c^{m+h,t}$ to infer the likelihood of the AI model's binary decision recommendation being correct, before she computes the utility of accepting or rejecting the model's recommendation. To do so, two qualitatively different approaches have previously been proposed in the probabilistic forecast aggregation literature [6, 7, 28]—the *compromising* approach which combines estimates by taking the average, and the *naïve Bayesian* approach that tends to push the combined estimate to the extreme (e.g, "60% and 60% is 70%"). Corresponding to these two approaches, we consider 4 ways of aggregating human and model confidence on the AI model's binary decision recommendation in this study:

- **Averaging (AVG)**: Following this rule, the human decision maker simply averages her confidence and the AI model's confidence on each decision as the final, aggregated confidence, i.e., $c^{m+h,t} = \frac{c^{m,t}+h(x^t)[y^t=\hat{y}^{m,t}]}{2}$.
- **Naïve Bayes (NB)**: In this rule, it is assumed that the two probabilistic estimates on the correctness likelihood of the AI model's binary decision recommendation (i.e., $h(x^t)[y^t = \hat{y}^{m,t}]$ and $c^{m,t}$) are conditionally independent with each other. Thus, the human decision maker combines these two estimates according to a naïve implementation of the Bayes rule:

$$c^{m+h,t} = \frac{1}{1 + \frac{(1-c^{m,t})\cdot(1-h(x^t)[y^t=\hat{y}^{m,t}])}{c^{m,t}\cdot h(x^t)[y^t=\hat{y}^{m,t}]}}$$

  Under this rule, the human decision maker tends to produce a more certain final estimate on the correctness likelihood of a decision than the individual estimate of the AI model or her own. In other words, the human decision maker becomes "overconfident" in her final estimate on the correctness likelihood of each decision after the aggregation.
- **Weighted Mean Log-Odds (WMLO)**: This rule is a combination of averaging and naïve Bayes rule [7], which states that the *log-odds* of the aggregated probabilistic estimate is the average value of the log-odds of the individual estimate:

$$c^{m+h,t} = \frac{\exp(\alpha)}{1+\exp(\alpha)},$$
$$\text{where } \alpha = \frac{1}{2}(\ln\frac{c^{m,t}}{1-c^{m,t}} + \ln\frac{h(x^t)[y^t=\hat{y}^{m,t}]}{1-h(x^t)[y^t=\hat{y}^{m,t}]})$$

  As taking the log-odds value of a probability accentuates the differences for extreme probabilities (i.e., probabilities that are close to 0 or 1), the net effect of this rule is thus to stretch out the extreme probabilities before taking the average.
- **Adjusted Naïve Bayes (ANB)**: Following this adjusted naïve Bayes rule [7], the human decision maker will first correct each probabilistic estimate by discounting it and moving it closer to 0.5, and then aggregate them using the naïve Bayes rule:

$$c^{m+h,t} = \frac{1}{1+\frac{(1-a)\cdot(1-b)}{a\cdot b}}, \text{where } a = \frac{(c^{m,t})^\gamma}{(c^{m,t})^\gamma + (1-c^{m,t})^\gamma},$$
$$b = \frac{(h(x^t)[y^t=\hat{y}^{m,t}])^\gamma}{(h(x^t)[y^t=\hat{y}^{m,t}])^\gamma + (1-h(x^t)[y^t=\hat{y}^{m,t}])^\gamma}$$

  The parameter γ varies from 0 to 1 and controls the degree of adjustment before applying the naïve Bayes rule—when $\gamma = 0$, all the probabilities are transformed to 0.5, while when $\gamma = 1$, no adjustment is applied to probabilities.

Finally, after the human decision maker obtains the aggregated confidence $c^{m+h,t}$ on the AI model's binary decision recommendation on a decision-making trial, she can compute the utility of each action following any of the basic utility models as we have discussed in Section 4.1.1. That is, combining the 4 ways of aggregating confidence with the 2 basic utility models, in total, we have 8 possible human-adjusted utility models.

4.2 Selection Component

The selection component describes how the human decision maker stochastically decides which action to take. Again, we first consider a few basic models in which this selection process is only influenced by the estimated utility of each action, i.e., $r^t_j = g(\hat{u}^t_{accept}, \hat{u}^t_{reject})$, where r^t_j is the probability of choosing the action $j \in \{accept, reject\}$

in trial t and $\sum_j r_j^t = 1$. Then, we consider the possibility that the selection process is affected by the human decision maker's own judgement on the decision-making trial, especially in terms of the agreement between the human's and the model's decisions. We thus look into the *human-adjusted* selection models to capture such behavior, i.e., $r_j^t = g(\hat{u}_{accept}^t, \hat{u}_{reject}^t, m(x^t), h(x^t))$.

4.2.1 Basic Selection Models. We consider three basic selection models in this study:

- ϵ-**Greedy**: Suppose the action with the highest estimated utility $\hat{u}_j^t$ is j^*. Then, $r_j^t = \epsilon$ when $j = j^*$; otherwise, $r_j^t = 1 - \epsilon$. In other words, the action with the maximal utility will be chosen, but there is a constant chance (i.e., $1 - \epsilon$) for humans to make errors.
- **Logit**: The probability for the human decision maker to take action j is given by a softmax function $r_j^t = \frac{\exp(\delta \hat{u}_j^t)}{\sum_{j'} \exp(\delta \hat{u}_{j'}^t)}$. The Logit model is a widely-used discrete choice model in economics [1, 37], which assumes that humans choose a suboptimal option more often when it is associated with a larger utility. The parameter δ in the model reflects the human decision maker's sensitivity to utilities: when $\delta \to 0$, the human decision maker takes actions randomly; when $\delta \to \infty$, the human decision maker almost always takes the action with the optimal estimated utility.
- **Double Hurdle (DH)**: The previous two models assume that humans select actions in each trial independently. In the Double Hurdle model [10], we assume that in each trial, with a probability of π ($\pi \geq 0$), the human decision maker takes the same action as that in the last trial, regardless of the estimated utility of that action. This could reflect humans' inherent trust or distrust in the AI model. In addition, conditioned on that the human decision maker takes utility into account in a trial, the probability for her to select each action follows the Logit model:

$$r_j^t = \begin{cases} \pi + \dfrac{(1-\pi)\exp(\delta \hat{u}_j^t)}{\sum_{j'} \exp(\delta \hat{u}_{j'}^t)}, j = d^{t-1} \\ \dfrac{(1-\pi)\exp(\delta \hat{u}_j^t)}{\sum_{j'} \exp(\delta \hat{u}_{j'}^t)}, j \neq d^{t-1} \end{cases}$$

4.2.2 Human-adjusted Selection Model. In human-adjusted selection models, we attempt to characterize how the human decision maker's judgement on the decision-making trial changes the ways that she takes actions. Intuitively, if the human decision maker's own binary decision $\hat{y}^{h,t}$ is the same as (different from) the AI model's decision $\hat{y}^{m,t}$, she may increase (decrease) the probability of accepting the model's decision recommendation and decrease (increase) the probability of rejecting the model's recommendation; moreover, the more confident she is about her own decision (i.e., the larger $c^{h,t}$ is), the more she would increase (decrease) r_{accept}^t and decrease (increase) r_{reject}^t. To reflect this intuition, we propose the following adjustment method. In particular, the human decision maker first computes the probability of choosing each action using a basic selection model. Then, she will adjust these probabilities based on the agreement between her decision and the AI model's decision, as well as her own decision confidence:

$$r_{accept}^t \propto r_{accept}^t \cdot \exp(\eta \cdot \hat{y}^{m,t} \cdot \hat{y}^{h,t} \cdot (c^{h,t} - 0.5))$$
$$r_{reject}^t \propto r_{reject}^t \cdot \exp(-\theta \cdot \hat{y}^{m,t} \cdot \hat{y}^{h,t} \cdot (c^{h,t} - 0.5))$$

After the adjustment, the human decision maker normalizes the values of r_j^t to ensure that $\sum_j r_j^t = 1$. In this model, parameters η and θ ($\eta, \theta > 0$) describe the extent to which the human decision maker adjusts the selection probabilities of each action—the larger η (or θ) is, the more the human decision maker will change the probability of accepting (or rejecting) the AI model's decision recommendation based on her own judgement.

As the proposed adjustment method can be applied on top of the probabilities produced by any basic selection model, in this study, we have a total of 3 human-adjusted selection model corresponding to the 3 basic selection models.

5 EXPERIMENTAL DESIGN

To explore which combinations of the two-component models can best capture human decision makers' behavior in AI-assisted decision-making, we conduct a human-subject experiment to collect the real human behavior data in AI-assisted decision making.

5.1 Decision-Making Task

The decision-making task that we asked subjects to complete in our experiment was to evaluate loan default risks. Specifically, in each task, the subject was presented with the profile of a loan applicant consisting of 7 features—the amount and interest rate of the loan, the number of months to pay off the loan, the value of each monthly installment, as well as the applicant's annual income, credit score, and homeownership status. After reviewing this information, the subject was asked to predict whether this applicant would default on the loan or not. Loan applicant profiles that we showed to subjects in the experiment were taken from a public dataset that records the loan information of a peer-to-peer lending platform, LendingClub [43]. To simplify the problem as a binary prediction, we restricted our attention only to those cases where the loan applicant either fully paid back the loan or defaulted on the loan. Further, the loan risk assessment tasks that subjects worked on in the experiment were randomly sampled from a pool of 1300 candidate loans that we selected from the dataset—We conjectured that humans may consider 4 factors to be highly predictive of the default risk of a loan, i.e., the loan amount, interest rate, installment to income ratio, and the applicant's credit score. Thus, we maximized the representativeness of our selected pool of decision-making tasks by ensuring a *balanced* distribution of values for each of these four factors as well as the intersection of these factors.

5.2 Pilot Study: Human Decisions Only

We first conducted a pilot study to collect data on how humans make loan default risk predictions *without* the assistance of an AI model. In this pilot study, each subject was asked to complete a sequence of 40 loan risk assessment tasks that were randomly sampled from a subset of 300 tasks in our task pool. In each task, the subject was asked to provide her binary prediction (i.e., "will default" or "will not default"). Further, she also needed to report her confidence in her prediction by indicating the probability that

she believed her prediction would be correct as a value between 50% to 100%[1]. In total, 211 subjects participated in this pilot study. Data collected from this pilot study later enables us to learn the human decision-making function $h(x)$ to infer the human decision maker's prediction and confidence on a decision-making trial. In practice, given a decision-making task, any historic data on how humans make decisions on this task can effectively serve as the "pilot study data" to allow us to model $h(x)$.

5.3 AI-Assisted Decision Making Experiment

Our real experiment was conducted to collect human behavior data under the AI-assisted decision-making setting.

5.3.1 AI model. First, we trained a gradient boosted trees model based on the LendingClub dataset to predict whether loan applicants would default on their loans. We then used the histogram binning method [46] to calibrate this model's confidence so that the model's confidence score on a decision-making trial could accurately reflect the correctness likelihood of the model's prediction on that trial. We evaluated the performance of this model on a hold-out test dataset and found an AUC score of 0.731, suggesting a reasonable predictive validity. Thus, the outputs of this model (including both the binary prediction and the prediction confidence) were provided as the decision recommendation to subjects in each of the decision-making tasks in the experiment.

5.3.2 Experimental treatments. We included two treatments in our experiment to simulate different types of decision-making environments. In particular, we suspected that human decision makers' behavior in AI-assisted decision making may vary with the *stakes* of the decision. As discussed in Section 3, we could use a parameter β, which is the ratio between the penalty triggered by a wrong decision and the reward associated with a correct decision, to characterize the relative stakes of the decision. We thus created two treatments by varying the value of β—in the high penalty (**HP**) treatment, we set $\beta = 2$, while in the low penalty (**LP**) treatment, we had $\beta = 0.5$. The behavior data that we obtained from these two treatments, thus, allows us to explore whether human decision makers utilize different processes to decide whether to adopt the AI model's recommendations when the stakes of the decisions vary.

5.3.3 Experimental Procedure. We conducted our experiment by posting human intelligence tasks (HITs) on Amazon Mechanical Turk (MTurk) and recruiting MTurk workers as our subjects. Upon arrival, subjects were randomly assigned to one of the two experiment treatments. Subjects were told that they need to complete a sequence of loan default risk assessment tasks in the HIT. We also told subjects that they were given a bonus account with an initial balance of 200 virtual points, and their bonus account balance would be updated based on whether her prediction in each task was correct (though the updated account balance would not shown to subjects in the real time)—a correct prediction would bring an addition of 10 points, while a wrong prediction would result in a deduction of 5 points for subjects in the LP treatment and 20 points for subjects in the HP treatment. Next, subjects received a tutorial about the meaning of each feature in a loan applicant's profile. At the end of the tutorial, we tested whether subjects understand the information presented in loan applicants' profiles via qualification questions, and subjects can only proceed after answering them correctly. After completing the tutorial, the subject started to work on a set of 40 decision-making tasks that were randomly selected from our task pool (we excluded the subset of 300 tasks used in the pilot study from the pool), with the assistance of our AI model. Specifically, in each task, the subject first saw the loan applicant's profile as well as the AI model's prediction and confidence on this profile, then she needed to make a final prediction by either accepting or rejecting the model's recommendation (see Figure A2 in Appendices for our task interface). The subject was explicitly told that the model's confidence scores were calibrated. We did not provide any immediate feedback to the subject on whether her or the model's prediction was correct on any of the tasks.

Our experiment was posted on weekdays 8am–6pm EST, and was open to U.S. workers only. Workers who had participated in our pilot study were not allowed to take the real experiment, and each worker can participate only once. Further, we included three common-sense questions (e.g., "what is 2+3?") in our HIT as the attention check questions, which later helped us to filter out the data from inattentive subjects. The base payment of the experiment was $1.80. We converted the subject's bonus account balance to the actual bonus payment using a rate of 400 points to $1, which leads to a maximum bonus of $1.50. The median time a subject spent on our HIT was 8.1 minutes, leading to a median hourly wage of $15.9.

[1]On the task interface, we told the subject that if she believed the probability for her prediction to be correct was lower than 50%, she might want to flip her prediction.

6 RESULTS

After filtering data from inattentive subjects, we obtained data from 404 subjects in our AI-assisted decision making experiment (HP: 214, LP: 190). In this section, we first examine the performance of various computational models in explaining an *average* human decision maker's behavior in adopting the AI model's recommendation in AI-assisted decision making (i.e., one model is learned to predict all human subjects' behavior). Then, we explore how changes in decision stakes impact human behavior.

6.1 Model Training

6.1.1 Human decision-making function $h(x)$. We first learned the human's decision-making function $h(x)$ by utilizing the data collected in our pilot study as the training data. Specifically, we processed the data by transforming the combination of each subject's self-reported binary prediction and prediction confidence in a task into the subject's confidence in the positive prediction for that task (i.e., $h(x)[y = +1]$). Next, for each task, we used the average value of $h(x)[y = +1]$ across all subjects who worked on that task to represent the average decision maker's binary decision $\overline{y^h}$ and confidence $\overline{c^h} = \overline{h(x)}[y = \overline{y^h}]$ on that task. We next trained a multi-output neural network to predict $\overline{y^h}$ and $\overline{c^h}$ based on x. Through 5-fold cross-validation, we found the average accuracy of this model in predicting $\overline{y^h}$ was 0.783, and its mean absolute error in predicting $\overline{c^h}$ was 0.056. In the following, we used this model's outputs to estimate the average decision maker's binary decision and decision confidence in each task.

6.1.2 Human behavior models. To evaluate the performance of different models in fitting the average human decision maker's behavior in AI-assisted decision making, we conducted a 5-fold cross-validation within each experimental treatment. That is, we randomly split subjects of that treatment into 5 groups and then created five folds of the behavior dataset based on the partition of the subjects. Within each iteration of cross-validation, given a particular type of two-component model (e.g., EU+Logit), we trained the model based on the training folds, using grid search to identify the best hyper-parameters with learning rate scheduler and early-stopping being implemented to avoid over-fitting. We then evaluated the performance of the trained model on the testing fold by computing the average negative log-likelihood (NLL) value of a subject's decisions to adopt the AI recommendations in the experiment across all subjects in the testing fold. We compared the performance of various models in fitting the average human decision maker's behavior by reporting the mean values of the average NLL across 5 cross-validation iterations. Intuitively, the *lower* the mean NLL, the better the model.

Finally, in addition to the two-component models, we also trained a few standard supervised learning models (e.g., SVM, logisitic regression, XGBoost) that directly predict d^t based on the task features x^t, the AI model's recommendation $\hat{y}^{m,t}$ and confidence $c^{m,t}$ in a decision-making trial. Among these models, the logistic regression model achieved the lowest mean NLL, thus we used it as our baseline in the analysis below.

6.2 Comparison in Model Performance

6.2.1 Basic Utility + Basic Selection. We first examine the performance of the two-component models that is composed of the basic utility models and the basic selection models in predicting the adoption behavior of the average human decision maker in AI-assisted decision making. We find here that the average decision maker's selection of different actions is best explained by the ϵ-Greedy selection model, which effectively makes the choice of utility model irrelevant (see Figure A3 in Appendices). However, all combinations of basic utility models and basic selection models are shown to perform worse than the baseline logistic regression model.

In addition, we notice that for models containing the Double Hurdle selection model, the optimal parameter value for π is always estimated to be 0, effectively making the Double Hurdle model degenerate to the Logit model. This implies that the average human decision maker's decision of adopting the AI model's recommendation or not in each trial is likely made in a case-by-case manner rather than being heavily influenced by their inherent trust disposition. In light of this, we exclude the Double Hurdle model from the set of selection models in our further analysis.

6.2.2 Human-adjusted Models. We next explore whether taking human's own judgement in a decision-making trial into consideration could improve the predictive performance of the two-component models. To do so, we first fix the utility model to be the basic ones, while replacing the selection model with the human-adjusted selection models. For example, Figure 2 reports the comparison between the human-adjusted selection models and the basic selection models in fitting the behavior of the average decision maker in the HP treatment. We find that for every combination of basic utility and selection model, applying human-based adjustment on the selection component of it significantly increases the model's performance in fitting the human behavior data, and we observe a similar trend in the LP treatment as well (see Figure A4 in Appendices). Further, we experiment with fixing the selection model to be the basic ones while replacing the utility model to be the human-adjusted versions. In this case, we found that when the selection model was the Logit model, applying human-based adjustment on the utility component of the two-component model always helps to increase the model's predictive performance (see Figure A5 in Appendices).

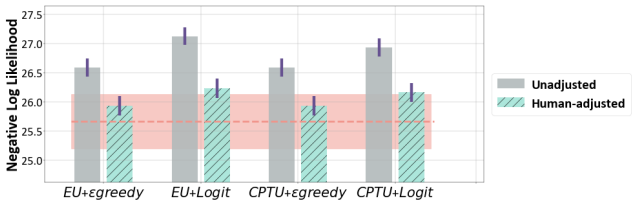


Figure 2: Performance of the two-component models in the HP treatment when using a basic utility model, and the selection model is either basic (gray bars) or human-adjusted (green bars). The red line represents the performance of the baseline logistic regression model. Error bars (shade) represent the standard errors of the mean.

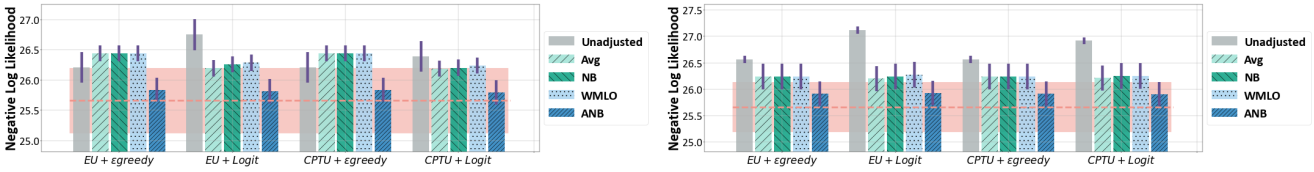
Finally, we apply the adjustment to both the utility and the selection model simultaneously. Figure 3 compares the predictive performance of the two-component models that are composed of both human-adjusted utility model and human-adjusted selection model, against that of the two-component models that are composed of the basic utility and basic selection model (gray bars). In both treatments, we find the best predictive performance we could get in predicting the average decision maker's behavior is achieved by the combination of the human-adjusted utility model (following the Adjusted Naïve Bayes rule) and the human-adjusted selection model[2]. Moreover, the best-performing two-component models can match the average performance of the baseline logistic regression model, while their performance variance decreases.

Taken together, our results here provide clear evidence supporting the conjecture that the average human decision maker incorporates her own judgement in a decision-making trial to decide whether to accept an AI model's recommendation. Further, such judgement may influence the decision makers' behavior through multiple steps in their cognitive reasoning processes.

6.3 Comparing Behavior Across Treatments

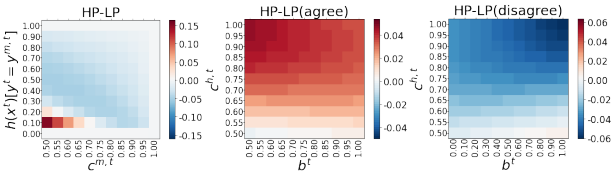
Lastly, we explore how the average decision makers' adoption behavior of the AI assistance is similar or different in varying decision making environments where the decision stakes vary. To do so, we compare the learned model parameters between the two experimental treatments for the combination of the adjusted CPT-based utility model (following the Adjusted Naïve Bayes rule) and the adjusted Logit selection model, as this two-component model achieves the best predictive performance in both treatments. In total, there are 5 model parameters for this two-component model: $k, \gamma, \delta, \eta, \theta$.

[2]The choice of the basic utility/selection model before applying adjustment does not seem to affect the two-component model's predictive performance significantly.



(a) Low Penalty Treatment (b) High Penalty Treatment

Figure 3: Performance of the two-component models when both utility and selection models are the basic models (gray bars), and when both the utility and selection models are human-adjusted models (the other four bars). The red lines represent the performance of the baseline models. Error bars (shades) represent the standard errors of the mean.



(a) Difference in b^t (b) Difference in r^t_{accept} (c) Difference in r^t_{accept}

Figure 4: Comparing human behavior in the two treatments with different decision stakes. Colors reflect the average value of the variable (b^t or r^t_{accept}) in the HP treatment minus the average value of that variable in the LP treatment.

We first note that based on the 5 models that we've obtained through the 5-fold cross-validation, the average estimates of the parameter k in the CPT-based utility model are $\overline{k}_{LP} = 1.99(\sigma = 0.07)$ for the LP treatment and $\overline{k}_{HP} = 0.63(\sigma = 0.01)$ for the HP treatment. This implies that in the LP treatment, the average decision maker tends to apply an S-shape weighting function to interpret probabilities, while the probability weighting function takes an inverse-S shape in the HP treatment. In other words, humans tend to overweight probabilities that are close to 0.5 when the decision stake is low and underweight them when the decision stake is high.

Next, we focus on the two model parameters that belong to the utility component of the model, i.e., k and γ. Recall that the cumulative effect of k and γ is to transform the AI model's confidence in its decision recommendation $c^{m,t}$ and the human decision maker's confidence in the AI model's decision $h(x^t)[y^t = \hat{y}^{m,t}]$ into a distorted probability $b^t = w(c^{m+h,t})$, which is the human decision maker's final "belief" in the AI model's correctness after the confidence aggregation and probability weighting. It is therefore interesting to explore that given the same $c^{m,t}$ and $h(x^t)[y^t = \hat{y}^{m,t}]$, how would the value of b^t differ when the decision stake is different. Given the 5 behavior models we've learned in our 5-fold cross-validation, we computed the *difference* in the average value of b^t between the HP treatment and the LP treatment, for all combinations of $c^{m,t}$ and $h(x^t)[y^t = \hat{y}^{m,t}]$, and the result is shown as a heatmap in Figure 4(a). We find that as the decision stakes become larger, human decision makers tend to increasingly underweight the likelihood for the AI model's decision recommendation being correct, as their final belief in AI correctness b^t is smaller in the HP treatment than that in the LP treatment for most combinations of AI confidence and human confidence. Such underweighting is particularly significant when the AI model's confidence (i.e., $c^{m,t}$) is relatively low while the human decision maker is highly uncertain in her decision (i.e., $(h(x^t)[y^t = \hat{y}^{m,t}])$ is around 0.5).

Similarly, we then look into the three model parameters in the selection component of the model, i.e., δ, η, and θ—they cumulatively transform the human decision maker's final belief in AI correctness b^t, together with her confidence in her own decision $c^{h,t}$, into a probability of accepting the AI model's recommendation r^t_{accept}, and this transformation follows different formulas depending on whether the human decision maker's own decision $\hat{y}^{h,t}$ agrees with the AI's decision $\hat{y}^{m,t}$ or not. Again, Figures 4(b) and 4(c) show the difference in the average value of r^t_{accept} between the HP and LP treatments for different combinations of b^t and $c^{h,t}$, when the human decision maker agrees or disagrees with the AI model's decision recommendation, respectively. We find that when the human's own decision is the same (different) as that of the AI's, such agreement (disagreement) leads to a consistently higher (lower) probability for the average decision maker to accept the AI's recommendation when the decision stake is larger. This is particularly salient when a human's confidence in her own decision $c^{h,t}$ is high.

7 CONCLUSION AND DISCUSSION

In this paper, we propose a two-component human behavior model space, including the utility component and the selection component, to describe humans' decisions to adopt the AI recommendation in AI-assisted decision making. We evaluate the performance of a variety of computational models in this space in fitting the real human behavior data collected through a large-scale randomized experiment. Our results show that the human-adjusted models outperform models that are only based on the AI model's outputs, suggesting that humans are prone to make use of their own judgement in a decision-making trial to gauge whether to adopt the AI recommendation. Moreover, the comparison of model parameters suggests that when the stakes of the decisions become larger, people tend to lower their belief in AI recommendation's correctness and rely more on their own judgement in AI-assisted decision making.

There are a few limitations in our current study. For example, the performance of our behavior models could be limited by the accuracy of the human decision-making function $h(x)$. Also, whether people will accept an AI recommendation may be affected by many more factors in real-world AI-assisted decision making settings, such as the presence of AI explanations and people's inherent preference to one decision. We believe our work still provides a general framework for modeling more realistic human-AI interaction in the future. For example, inherent preference among decisions can be captured by selection models with a "default" decision, while the presence of AI explanations may indicate human adjustment should also consider whether the humans' and AI's rationale matches. We

hope our work can inspire more studies in modeling human-AI interaction and in integrating realistic human behavior models into the optimization of AI-driven decision aids.

ACKNOWLEDGMENTS

We are grateful to Alexandros Psomas and all anonymous reviewers who provided many helpful comments. We thank the support of the National Science Foundation under grant IIS-1850335 on this work. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

REFERENCES

[1] OA Adeogun, AM Ajana, OA Ayinla, MT Yarhere, and MO Adeogun. 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 (2008), 468–472.

[2] Amos Azaria, Ya'akov Gal, Sarit Kraus, and Claudia V Goldman. 2016. Strategic advice provision in repeated human-agent interactions. *Autonomous Agents and Multi-Agent Systems* 30, 1 (2016), 4–29.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. (2021).

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[6] Jonathan Baron, Barbara A Mellers, Philip E Tetlock, Eric Stone, and Lyle H Ungar. 2014. Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis* 11, 2 (2014), 133–145.

[7] David V Budescu and Hsiu-Ting Yu. 2006. To Bayes or not to Bayes? A comparison of two classes of models of information aggregation. *Decision analysis* 3, 3 (2006), 145–162.

[8] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *13th ACM Web Science Conference 2021*. 120–129.

[9] Abir De, Paramita Koley, Niloy Ganguly, and Manuel Gomez-Rodriguez. 2020. Regression under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 2611–2620.

[10] Diansheng Dong, Chanjin Chung, and Harry M Kaiser. 2004. Modelling milk purchasing behaviour with a panel data double-hurdle model. *Applied Economics* 36, 8 (2004), 769–779.

[11] Sebastian Ebert and Philipp Strack. 2015. Until the bitter end: on prospect theory in a dynamic context. *American Economic Review* 105, 4 (2015), 1618–33.

[12] Drew Fudenberg, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2019. Measuring the completeness of theories. (2019).

[13] Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. 2021. Human-AI Collaboration with Bandit Feedback. *arXiv preprint arXiv:2105.10614* (2021).

[14] Yaohui Guo and X Jessie Yang. 2020. Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach. *International Journal of Social Robotics* (2020), 1–11.

[15] Yaohui Guo, Chongjie Zhang, and X Jessie Yang. 2020. Modeling Trust Dynamics in Human-robot Teaming: A Bayesian Inference Approach. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–7.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[17] Wan-Lin Hu, Kumar Akash, Neera Jain, and Tahira Reid. 2016. Real-time sensing of trust in human-machine interactions. *IFAC-PapersOnLine* 49, 32 (2016), 48–53.

[18] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems* 29 (2016).

[19] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2018. Human-in-the-loop interpretability prior. *Advances in neural information processing systems* 31 (2018).

[20] Vivian Lai and Chenhao Tan. 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.

[21] Pamela K Lattimore, Joanna R Baker, and Ann D Witte. 1992. The influence of probability on risky choice: A parametric examination. *Journal of economic behavior & organization* 17, 3 (1992), 377–400.

[22] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.

[23] Han Liu, Vivian Lai, and Chenhao Tan. 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, CSCW2 (2021), 1–45.

[24] Siqi Liu, J Benjamin Miller, and Alexandros Psomas. 2019. Risk Robust Mechanism Design for a Prospect Theoretic Buyer. In *International Symposium on Algorithmic Game Theory*. Springer, 95–108.

[25] Yidu Lu. 2020. *Detecting and overcoming trust miscalibration in real time using an eye-tracking based technique.* Ph. D. Dissertation.

[26] Zhuoran Lu and Ming Yin. 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.

[27] Daniel McFadden. 2001. Economic choices. *American economic review* 91, 3 (2001), 351–378.

[28] Robert Mislavsky and Celia Gaertig. 2020. Combining Probability Forecasts: 60% and 60% Is 60%, but Likely and Likely is Very Likely. *Forthcoming at Management Science, Johns Hopkins Carey Business School Research Paper* 20-14 (2020).
[29] Gali Noti, Effi Levi, Yoav Kolumbus, and Amit Daniely. 2016. Behavior-based machine-learning: A hybrid approach for predicting human decision making. *arXiv preprint arXiv:1611.10228* (2016).
[30] Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. 2021. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* 372, 6547 (2021), 1209–1214.
[31] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.
[32] Drazen Prelec. 1998. The probability weighting function. *Econometrica* (1998), 497–527.
[33] John Quiggin. 2012. *Generalized expected utility theory: The rank-dependent model*. Springer Science & Business Media.
[34] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. 2019. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220* (2019).
[35] Amy Rechkemmer and Ming Yin. 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*.
[36] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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.
[37] Kenneth E Train. 2009. *Discrete choice methods with simulation*. Cambridge university press.
[38] Amos Tversky and Craig R Fox. 1995. Weighing risk and uncertainty. *Psychological review* 102, 2 (1995), 269.
[39] Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty* 5, 4 (1992), 297–323.
[40] John Von Neumann and Oskar Morgenstern. 1947. Theory of games and economic behavior, 2nd rev. (1947).
[41] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328.
[42] Anqi Xu and Gregory Dudek. 2015. Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. IEEE, 221–228.
[43] Yash. 2020. Lending club 2007-2020q3 | Kaggle. https://www.kaggle.com/ethon0426/lending-club-20072020q1?select=Loan_status_2007-2020Q3.gzip
[44] Ming Yin and Yu-An Sun. 2015. Human behavior models for virtual agents in repeated decision making under uncertainty. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. 581–589.
[45] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 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.
[46] Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, Vol. 1. Citeseer, 609–616.
[47] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 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.

A APPENDICES

A.1 An Overview of the Model Space

components.png

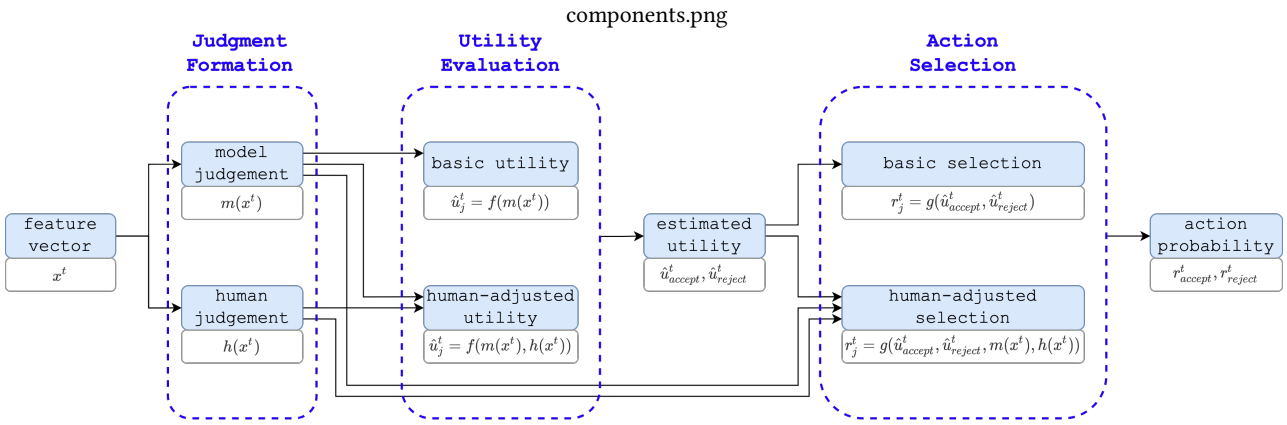
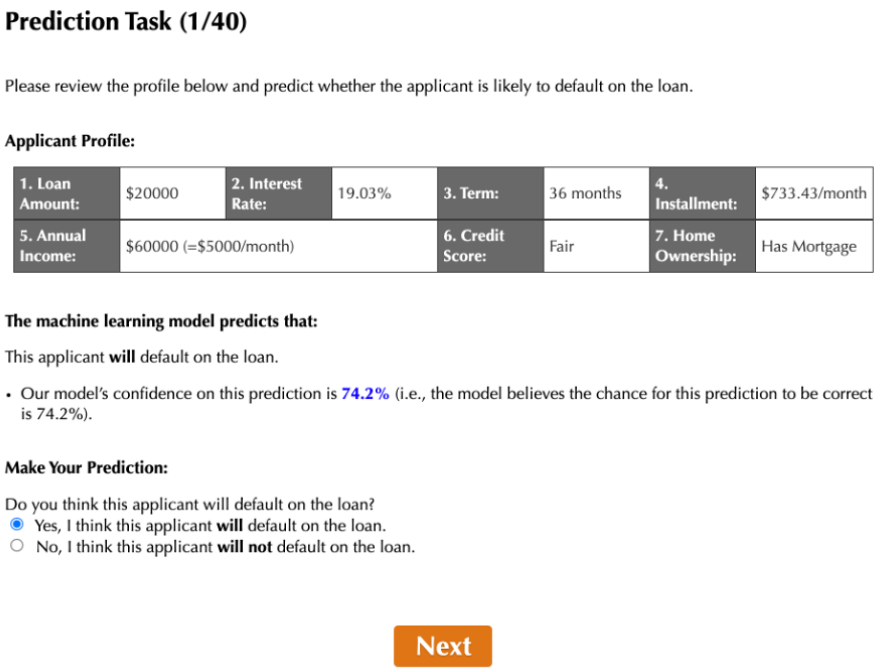


Figure A1: An overview of the two-component model space. The utility component contains the basic utility models and the human-adjusted utility models. Similarly, the selection component contains the basic selection models and the human-adjusted selection models. A two-component model can be composed by any model in the utility component and any model in the selection component.

A.2 Interface of the Loan Default Risk Assessment Tasks



Prediction Task (1/40)

Please review the profile below and predict whether the applicant is likely to default on the loan.

Applicant Profile:

1. Loan Amount:	$20000	**2. Interest Rate:**	19.03%	**3. Term:**	36 months	**4. Installment:**	$733.43/month
5. Annual Income:	$60000 (=$5000/month)			**6. Credit Score:**	Fair	**7. Home Ownership:**	Has Mortgage

The machine learning model predicts that:

This applicant **will** default on the loan.

- Our model's confidence on this prediction is **74.2%** (i.e., the model believes the chance for this prediction to be correct is 74.2%).

Make Your Prediction:

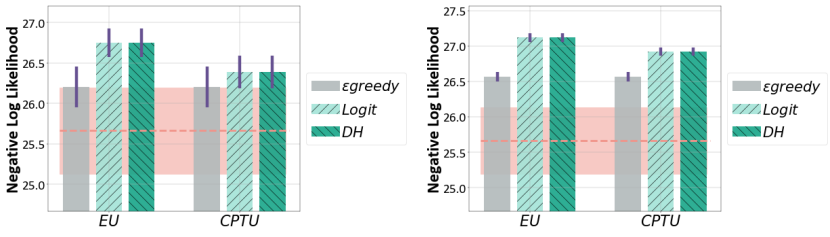
Do you think this applicant will default on the loan?

- (●) Yes, I think this applicant **will** default on the loan.
- ( ) No, I think this applicant **will not** default on the loan.

Next

Figure A2: An example of the task interface that subjects saw in the AI-assisted decision making experiment.

A.3 Additional Experiment Results



(a) Low Penalty Treatment (b) High Penalty Treatment

Figure A3: Performance of the two-component models that are composed of basic utility and selection models, as well as the baseline model (the red line). Error bars (shade) represent the standard errors of the mean.

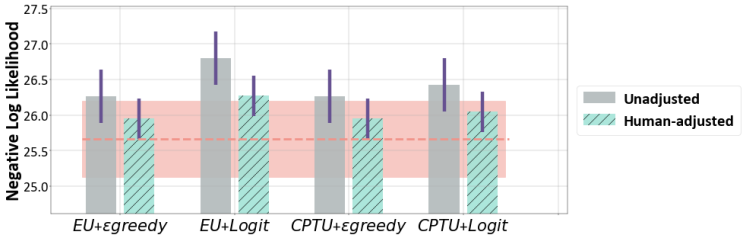
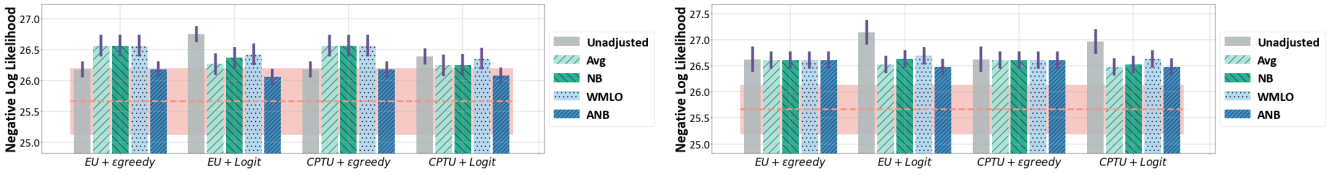


Figure A4: Performance of the two-component models in the LP treatment when using a basic utility model, and the selection model is either basic (gray bars) or human-adjusted (green bars). The red line represents the performance of the baseline logistic regression model. Error bars (shade) represent the standard errors of the mean.



(a) Low Penalty Treatment (b) High Penalty Treatment

Figure A5: Performance of the two-component models when using a basic selection model, and the utility model is either basic (gray bars) or human-adjusted (the other four bars). The red lines represent the performance of the baseline logistic regression models. Error bars (shades) represent the standard errors of the mean.