

# Modeling Human Trust and Reliance in AI-Assisted Decision Making: A Markovian Approach

Zhuoyan Li\*, Zhuoran Lu\*, Ming Yin

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

## Abstract

The increased integration of artificial intelligence (AI) technologies in human workflows has resulted in a new paradigm of AI-assisted decision making, in which an AI model provides decision recommendations while humans make the final decisions. To best support humans in decision making, it is critical to obtain a quantitative understanding of how humans interact with and rely on AI. Previous studies often model humans' reliance on AI as an analytical process, i.e., reliance decisions are made based on a cost-benefit analysis. However, theoretical models in psychology suggest that the reliance decisions can often be driven by *emotions* like humans' *trust* in AI models. In this paper, we propose a hidden Markov model to capture the affective process underlying the human-AI interaction in AI-assisted decision making, by characterizing how decision makers adjust their trust in AI over time and make reliance decisions based on their trust. Evaluations on real human behavior data collected from human-subject experiments show that the proposed model outperforms various baselines in accurately predicting humans' reliance behavior in AI-assisted decision making. Based on the proposed model, we further provide insights into how humans' trust and reliance dynamics in AI-assisted decision making is influenced by contextual factors like decision stakes and their interaction experiences.

## Introduction

With the rapid development of artificial intelligence (AI) technologies in recent years, AI models have been increasingly adopted to help people make better decisions in various domains ranging from finance to healthcare. The nature of many decisions involving high stakes and the need to maintain human agency in decision making have led to the paradigm of *AI-assisted decision making*, that is, an AI model makes a decision recommendation to humans, who will then make the final decision. Designing an AI model to best support human decision makers in such a paradigm requires thorough understandings of how humans factor AI recommendations into their final decisions, for example, through deciding whether to rely on the AI model or not in each decision making case.

Many important efforts have been made to obtain these understandings. For example, experimental studies have been conducted, through which a wide range of factors that influence decision makers' reliance on AI models in AI-assisted decision making have been identified (Yin, Wortman Vaughan, and Wallach 2019; Zhang, Liao, and Bellamy 2020; Bansal et al. 2019a). More recently, researchers go beyond the empirical understandings and start to explore how to computationally model human decision makers' reliance on AI models in AI-assisted decision making, taking multiple influencing factors into account (Wang, Lu, and Yin 2022; Kumar et al. 2021; Bansal et al. 2021a). Interestingly, in these studies, humans' decisions on whether to rely on an AI model or not are often modeled as an *analytical* process, that is, people make their reliance decision by estimating the "utility" of accepting or rejecting the AI recommendation and completing a cost-benefit analysis. Meanwhile, conceptual models proposed in the psychology literature suggest that people's reliance behavior can be largely controlled by their emotions, such as their *trust* in the AI model. It is thus natural to ask, can we computationally model people's *affective* process underlying their reliance decisions, and how well can these models fit people's real-world reliance behavior in AI-assisted decision making?

To answer these questions, in this paper, we propose a hidden Markov model to characterize the dynamics of decision makers' latent trust in the AI model in AI-assisted decision making, as well as their adjustment of reliance decisions based on the trust. Specifically, building upon a few conceptual models in the psychology literature (Lee and See 2004; Hoff and Bashir 2015), our proposed model consists of three components—an initial trust model, a trust dynamics model, and a decision model. The initial trust model captures how people's trust in AI is influenced by their own characteristics and some contextual factors *prior to* their interactions with the AI model. On the other hand, the trust dynamics model and the decision model specify that *during* the interactions, how decision makers' trust changes over time, or the relationships between trust and reliance, depend on contextual factors and their interaction experiences.

To evaluate the performance of the proposed model in capturing people's reliance behavior in AI-assisted decision making, we collect data on real human subjects' reliance decisions in AI-assisted income prediction tasks through a ran-

---

\*Li and Lu have made equal contributions to this work.  
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

domized experiment. By fitting various computational models to the behavior dataset collected, we find that our proposed model consistently outperforms other baseline models in accurately predicting both the reliance behavior pattern exhibited by a population of decision makers and the reliance decisions made by individual decision makers. Through an ablation study, we further identify some differences in how contextual factors and decision makers’ interaction experiences influence reliance—while decision makers’ interaction experiences (e.g., the feedback on whether their previous reliance decision is appropriate) appear to influence their reliance decisions both directly and indirectly via changing their latent trust, contextual factors like task characterizations and decision stakes seem to mostly influence the reliance decisions indirectly through trust. Finally, a close examination of the learned model enables us to provide a few quantitative explanations on *how* contextual factors and interaction experiences impact decision makers. For example, it is shown that decision makers’ trust levels are more stable when the decision stake becomes higher.

## Related Work

The increased usage of decision aids driven by AI models has inspired a line of experimental studies to understand how humans interact with and rely on AI models in AI-assisted decision making. Researchers have identified a large set of factors that may influence people’s reliance on AI, including the AI 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 of AI explanations and the ways that they are presented (Yang et al. 2020; Bansal et al. 2021b), humans’ mental models about AI (Bansal et al. 2019a,b), and the level of human-model agreement (Lu and Yin 2021). It was shown in many experimental studies that decision makers often can not rely on AI models appropriately, which leads to new studies on designing innovative methods to promote appropriate reliance on AI (Bućinca, Malaya, and Gajos 2021; Park et al. 2019; Liao and Sundar 2022; Chiang and Yin 2022).

Most recently, some researchers have started to explore computationally modeling humans behavior in AI-assisted decision making, such as characterizing and predicting when decision makers will solicit/rely on the recommendation provided by the AI model (Wang, Lu, and Yin 2022; Bansal et al. 2021a; Kumar et al. 2021; Pynadath, Wang, and Kamireddy 2019). Perhaps inspired by theoretical frameworks in economics that explain human decision making behavior under uncertainty (Tversky and Kahneman 1992; Allais 1953), many of these studies model reliance decisions in AI-assisted decision making as an outcome of humans undertaking bounded rational cost-benefit analyses. We take a different approach in this paper to model the reliance decisions as the outcomes of an *affective* process (i.e., reliance is regulated by emotions like *trust* in AI).

Another related line of works focus on designing algorithmic methods to improve collaborations between humans and AI. One common setting considered in these studies is that humans and AI each work on some tasks *separately*. For example, algorithms have been developed to explicitly assign

tasks to the most suitable party (i.e., humans vs. AI) given their respective strengths and weaknesses (Wilder, Horvitz, and Kamar 2021; Gao et al. 2021), or to teach one party to effectively defer some tasks to the other party (Madras, Pitassi, and Zemel 2018; Mozannar, Satyanarayan, and Sonntag 2022). Alternatively, when both humans and AI have worked on the *same* tasks, meta-decision models have been developed to combine the outputs from both humans and AI to achieve an aggregated, better task performance (Kerrigan, Smyth, and Steyvers 2021; Steyvers et al. 2022). We emphasize that in the AI-assisted decision making setting that we consider in this paper, while both humans and AI “work” on the same decision making tasks, the final decisions are always made by humans instead of a meta-decision model.

## Methods

In this section, we first formally define the AI-assisted decision making setting studied in this paper. Then, we review a few conceptual models proposed in the psychology literature, which describe how human behavior (e.g., *reliance*) is influenced by emotions like *trust* when humans interact with automated systems. Based on these theoretical conceptual models, we introduce our Markov-model-based approach to computationally model humans’ reliance behavior in AI-assisted decision making as an affective process.

## Problem Formulation

We consider the following sequential AI-assisted decision making setting in this paper: Suppose a human decision maker needs to complete a sequence of  $T$  binary decision making tasks with the help of an AI model. In each task  $t$  ( $1 \leq t \leq T$ ), the human decision maker is provided with the task context  $\mathbf{x}_t \in \mathcal{R}^n$ , which may include the features/characterization of the task and situational factors like the stake of the decision. In addition, the decision maker receives the AI model’s binary decision recommendation  $y_t^m \in \{0, 1\}$ , which may or may not be the same as  $y_t$ , the correct decision of this task. With all these information, the human decision maker needs to make a final decision  $y_t^h \in \{0, 1\}$  by forming a reliance decision to either accept or reject the AI model’s recommendation (i.e.,  $d_t \in \{\text{accept}, \text{reject}\}$ ). Once the final decision is made, the decision maker will be informed of its correctness, thus they obtain the feedback  $e_t$  about whether their reliance decision  $d_t$  is appropriate. Depending on the correctness of AI recommendation  $y_t^m$  and the decision maker’s reliance decision  $d_t$ , feedback  $e_t = (c_t = \mathbb{I}(y_t^m = y_t), d_t)$  may take one of the four possible values—appropriate acceptance ( $c_t = 1, d_t = \text{accept}$ ), appropriate rejection ( $c_t = 0, d_t = \text{reject}$ ), inappropriate acceptance ( $c_t = 0, d_t = \text{accept}$ ), and inappropriate rejection ( $c_t = 1, d_t = \text{reject}$ ). The goal of our study is to build computational models to quantitatively characterize how humans adjust their reliance behavior (i.e.,  $d_t$ ) in these sequential AI-assisted decision making settings.

## Conceptual Models of Trust-Reliance Relationships

Conceptual models have been previously proposed in the psychology literature to characterize human behavior

in human-automation interactions. For example, Lee and See (2004) suggested that humans’ risk-taking behavior like relying on the automation is a behavioral expression of *trust*, which is an emotion that varies over time and reflects people’s affective responses to the automation’s violation or confirmation of their implicit expectancies. Hoff and Bashir (2015) further proposed a three-layer model to conceptualize the variability in human-automation trust. According to this model, human-automation trust could be decomposed into *dispositional trust* (i.e., trust decided by humans’ characteristics), *situational trust* (i.e., trust decided by the interaction context), and *learned trust* (i.e., trust decided by the experience relevant to the specific automated system). Moreover, the model stated that people’s initial reliance strategies *prior to* interactions with the automation is decided by their dispositional trust, situational trust, and initial learned trust if they have interacted with the automation before. On the other hand, *during* the interactions, people’s reliance behavior is mostly dependent on situational factors and their dynamic learned trust, which keeps being updated to reflect people’s ongoing experience with the automation.

### Characterizing Reliance in AI-Assisted Decision Making with Markov Models

The conceptual models described above provide a theoretical foundation for us to model humans’ reliance behavior in AI-assisted decision making as the outcome of an affective process. To operationalize these conceptual models, we propose a hidden Markov model to capture how human decision makers’ reliance on AI in AI-assisted decision making is influenced by their latent, dynamic trust in the AI model. Figure 1 illustrates the structure of the proposed model.

Specifically, consider a decision maker  $j$  with certain demographic background  $\mathbf{o}^j \in \mathcal{R}^d$  who, for the first time, completes  $T$  binary decision making tasks with the help of an AI model. In each task  $t$ , the decision maker encounters a task context  $\mathbf{x}_t^j$ , and we model the decision maker’s trust in the AI model as a latent categorical variable  $z_t^j \in \mathcal{Z} = \{1, 2, \dots, K\}$  ( $K$  is the total number of trust states). The decision maker’s trust state  $z_1^j \in \mathcal{Z}$  prior to interactions with the AI model is defined by an *initial trust model* (ITM):

$$\mathbb{P}(z_1^j = k | \mathbf{x}_1^j, \mathbf{o}^j; \boldsymbol{\theta}_{ITM}), \forall k \in \mathcal{Z} \quad (1)$$

where  $\boldsymbol{\theta}_{ITM}$  is the model parameter. Aligning with the conceptual models, this initial trust model specifies how the decision maker’s initial trust distribution is influenced by their own characteristics  $\mathbf{o}^j$  (i.e., dispositional trust) and the context of the first task  $\mathbf{x}_1^j$  (i.e., situational trust)<sup>1</sup>.

To reflect the adjustment of the decision maker’s trust in the AI model over time, we then define the *trust dynamics model* (TDM) parameterized by  $\boldsymbol{\theta}_{TDM}$ :

$$\mathbb{P}(z_t^j = k' | z_{t-1}^j = k, \mathbf{x}_t^j, \mathbf{e}_{t-1}^j; \boldsymbol{\theta}_{TDM}), \forall k, k' \in \mathcal{Z} \quad (2)$$

The trust dynamics model specifies the transition probabilities from one latent trust state to another. Consistent with

<sup>1</sup>Since decision maker  $j$  never interacts with the AI model before, we do not consider initial learned trust in this model.

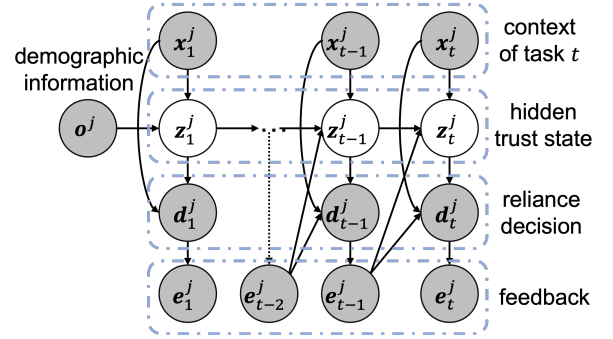


Figure 1: A hidden Markov model characterizing how humans adjust their reliance on AI based on their trust in the AI model. Shaded/unshaded nodes are observed/latent.

the conceptual models, we assume that the decision maker’s trust state in task  $t$  (i.e.,  $z_t^j$ ) evolves from their trust state in the previous task  $z_{t-1}^j$  and depends on situational factors summarized in  $\mathbf{x}_t^j$ , the context of the current task. In addition, to include the dynamic learned trust in the trust dynamics model, we also assume that  $z_t^j$  is influenced by the feedback  $\mathbf{e}_{t-1}^j$  that the decision maker receives from task  $t-1$ , which reflects both the AI model’s correctness and the decision maker’s reliance appropriateness in the previous task.

Finally, to characterize the relationship between the decision maker’s trust  $z_t^j$  and reliance decision  $d_t^j$ , we define a *decision model* (DM) parameterized by  $\boldsymbol{\theta}_{DM}$ :

$$\mathbb{P}(d_t^j = \text{accept} | z_t^j = k, \mathbf{x}_t^j, \mathbf{e}_{t-1}^j; \boldsymbol{\theta}_{DM}), \forall k \in \mathcal{Z} \quad (3)$$

The decision model effectively describes the emission probabilities of different trust states, and it again depends on the current task context  $\mathbf{x}_t^j$  and the feedback  $\mathbf{e}_{t-1}^j$  received in the previous task.

**Model Learning** Given  $N$  human decision makers who each completes  $T$  tasks with the AI assistance, we can collect a set of human behavior data  $\mathcal{D} = \{\mathbf{o}^j, \{\mathbf{x}_t^j, d_t^j, \mathbf{e}_t^j\}_{t=1}^T\}_{j=1}^N$ . The log-likelihood of this dataset  $\mathcal{D}$  under the proposed Markov model  $\mathcal{M}$  (parameterized by  $\{\boldsymbol{\theta}_{ITM}, \boldsymbol{\theta}_{TDM}, \boldsymbol{\theta}_{DM}\}$ ) is:

$$L(\mathcal{M}, \mathcal{D}) = \sum_{j=1}^N \log \left( \sum_z \mathbb{P}(z_1 | \mathbf{x}_1^j, \mathbf{o}^j; \boldsymbol{\theta}_{ITM}) \cdot \prod_{t=2}^T \mathbb{P}(z_t | z_{t-1}, \mathbf{x}_t^j, \mathbf{e}_{t-1}^j; \boldsymbol{\theta}_{TDM}) \cdot \prod_{t=1}^T \mathbb{P}(d_t^j | z_t, \mathbf{x}_t^j, \mathbf{e}_{t-1}^j; \boldsymbol{\theta}_{DM}) \right) \quad (4)$$

We can then use the expectation-maximization (EM) algorithm (McLachlan and Krishnan 2007) to learn the optimal model parameters that maximize this log-likelihood.

**Model Inference** Once the model  $\mathcal{M}$  is learned, given a new decision maker  $j^*$ , we may make online inference of their trust state and predict their reliance decision. Specifically, suppose the new decision maker  $j^*$  has completed  $L$  tasks, leading to a behavior data dataset  $\mathcal{D}_{j^*} =$

$\{\mathbf{o}^{j^*}, \{\mathbf{x}_t^{j^*}, d_t^{j^*}, \mathbf{e}_t^{j^*}\}_{t=1}^L\}$ . Then, the maximum a posteriori estimate of the most likely hidden trust state sequence of decision maker  $j^*$ , i.e.,  $\mathbf{z}^{j^*} = \{z_1^{j^*}, \dots, z_L^{j^*}\}$ , can be estimated using the Viterbi algorithm (Forney 1973). The inferred hidden trust state sequence will further enable us to predict the decision maker’s trust  $z_{L+1}^{j^*}$  and reliance decision  $d_{L+1}^{j^*}$  in the next task (i.e., task  $L + 1$ ) given the task context  $\mathbf{x}_{L+1}^{j^*}$ :

$$\begin{aligned} z_{L+1}^{j^*} &= \operatorname{argmax}_{k \in \mathcal{Z}} \mathbb{P}(z_{L+1}^{j^*} = k | z_L^{j^*}, \mathbf{x}_{L+1}^{j^*}, \mathbf{e}_L^{j^*}; \boldsymbol{\theta}_{TDM}) \\ d_{L+1}^{j^*} &= \operatorname{argmax}_{d \in \{\text{accept}, \text{reject}\}} \mathbb{P}(d_{L+1}^{j^*} = d | z_{L+1}^{j^*}, \mathbf{x}_{L+1}^{j^*}, \mathbf{e}_L^{j^*}; \boldsymbol{\theta}_{DM}) \end{aligned} \quad (5)$$

## Human-Subject Experiment

To evaluate the performance of our proposed model in capturing humans’ reliance behavior in AI-assisted decision making, we conducted an experiment to collect real human behavior data in AI-assisted decision making.

**Decision Making Task** The decision making task we used in our experiment was to determine a person’s annual income level, which was a common task used in previous AI-assisted decision making studies (Zhang, Liao, and Bellamy 2020; Ribeiro, Singh, and Guestrin 2018). Specifically, in each task, the human subject was presented with a profile of a person with 7 features, including the person’s gender, age, education level, marital status, occupation, work type, and working hours per week. The subject was asked to decide whether this person’s annual income is higher or lower than \$50k. The profiles that we showed to subjects were taken from the UCI Income dataset. Based on this dataset, we trained a gradient boosted trees model to make the income prediction, and we presented to subjects the binary prediction made by this model on each task as the AI recommendation.

**Experimental Treatments** To explore how humans trust and rely on the AI models differently under different contexts, and whether our proposed model can capture these differences, we created two treatments in this experiment. The context-related factor that we varied across treatments was the *stake* of the decisions. In particular, in one treatment, subjects would receive high penalty (**HP**) from their incorrect decisions in comparison with the reward that they would get from correct decisions, while in the other treatment, incorrect decisions lead to relatively low penalty (**LP**).

**Experimental Procedure** We posted our experiment on Amazon Mechanical Turk (MTurk) as a human intelligence task (HIT) and recruited MTurk workers as our subjects. Upon arrival, we randomly assigned each subject to one of the two treatments. Subjects started the HIT by completing a tutorial which described the income prediction task that they needed to work on in the HIT and the meaning of each feature they would see in a person’s profile. To help subjects get familiar with the task, we first asked subjects to complete 5 training tasks, in which they made income predictions *without* the AI advice, and we showed them the correct answer of each task immediately after they made their predictions.

The real experiment began after subjects completed the training tasks. Specifically, subjects were told that they would start with an account of 200 virtual points in this HIT, and they needed to determine income levels in a sequence of 20 tasks. For each task, if their decision was correct, they would earn 10 points. However, if their decision was wrong, subjects of the LP treatment would lose 5 points while subjects of the HP treatment would lose 20 points. After subjects made their decision on a task, we immediately provided the feedback to them indicating the correctness of their decision and updated their account balance. At the end of the HIT, the points left in a subject’s account was converted to the subject’s bonus payment using a ratio of 40 points to \$0.1. The 20 tasks a subject worked on in the HIT was randomly sampled from a pool of 500 task instances, and we ensured that among these 20 tasks, the AI model was correct on 15 tasks and wrong on 5 tasks (i.e., the AI model’s accuracy on the 20 tasks was 75%). After subjects completed all 20 tasks, they were asked to complete a survey to report their demographic information, such as their age, gender, education level, and their familiarity with AI and programming.

The base payment of our HIT was \$1.2. The HIT was only open to U.S. workers, and each worker can complete the HIT once. We further included an attention check question in the HIT, in which subjects were asked to choose a randomly specified choice. Only the data of subjects who passed the attention check was considered valid (see supplemental materials for information of the data repository).

## Evaluations

After filtering the inattentive subjects, we obtained valid data from 245 subjects in our experiment (HP:118, LP:127). Below, we conduct our evaluation using the behavior data collected from these valid subjects.

### Model Training and Baselines

For training the proposed hidden Markov model, all information that a subject provides in the exit-survey is included in the subject’s demographic background (i.e.,  $\mathbf{o}^j$ ), while the task context  $\mathbf{x}_t$  includes both the 7 features of the income prediction task and the decision stake (i.e., the treatment-dependent penalty for wrong decisions). We use multinomial logistic regression models as our initial trust and trust dynamic models, while logistic regression models are used as the decision models. We also experiment with a range of values for the number of hidden trust states (i.e.,  $K = 2 \sim 6$ ) and find the model with  $K = 3$  achieves the maximum Bayesian information criterion (BIC) score (Schwarz 1978), thus we set  $K = 3$  throughout our evaluation.

We consider a few baseline computational models in our evaluation. First, we include three supervised learning models—logistic regression, XGBoost (Chen and Guestrin 2016), and LSTM (Hochreiter and Schmidhuber 1997)—as our baselines. These models directly predict decision maker  $j$ ’s reliance decision  $d_t^j$  using their demographic background  $\mathbf{o}^j$ , the current task context  $\mathbf{x}_t^j$ , and the feedback received in the previous task  $\mathbf{e}_{t-1}^j$ , *without* characterizing the analytical or affective processes underlying the reliance decision.

As a second type of baseline, we adapt a model from Wang, Lu, and Yin (2022) to characterize humans’ reliance decision in AI-assisted decision making as an *analytical* process, i.e., subjects make reliance decisions based on a cost-benefit analysis. Specifically, in task  $t$ , we assume a decision maker estimates the correctness likelihood of the AI recommendation  $y_t^m$  using the AI model’s past accuracies weighted by a time discounting factor  $\epsilon \in [0, 1]$ , i.e.,  $\hat{c}_t = \frac{1}{C} \sum_{i=1}^{t-1} \epsilon^{t-1-i} \mathbb{I}(y_i^m = y_i)$  and  $C = \sum_{i=0}^{t-2} \epsilon^i$  is the normalizing factor. Based on this correctness likelihood estimate, the subject can compute the utility of accepting or rejecting the AI recommendation:

$$\begin{aligned} u_t(\text{accept}) &= (1 + \lambda)w(\hat{c}_t) - \lambda \\ u_t(\text{reject}) &= 1 - (1 + \lambda)w(\hat{c}_t) \end{aligned} \quad (6)$$

where  $\lambda$  is the ratio between the wrong decision penalty and the correct decision reward, i.e.,  $\lambda = 0.5$  or  $2$  in the LP or HP treatments, respectively. Consistent with the Cumulative Prospect Theory (CPT), we assume people tend to distort probabilities via a weighting function  $w(p) = \frac{p^k}{p^k + (1-p)^k}$  ( $k > 0$ ). The decision maker’s stochastic reliance decision is then made based on a softmax function of the computed utilities, i.e.,  $\mathbb{P}(d_t = \text{accept}) = \frac{\exp(\delta u_t(\text{accept}))}{\sum_d \exp(\delta u_t(d))}$  ( $\delta$  is a parameter reflecting people’s sensitivity to utilities).

Finally, while our proposed method learns the dynamics of a decision maker’s trust in the AI model through a hidden Markov model, previous research has proposed quantitative models to explicitly specify the trust dynamics in human-automation interaction. Thus, as a final baseline, we consider a model which still characterizes the *affective* process underlying reliance decisions, but adopts an explicit trust dynamics model that is adapted from Hu et al. (2018):

$$\begin{aligned} z_t^j &= z_{t-1}^j + \alpha_e(\beta_e e_{t-1}^j - z_{t-1}^j) + \alpha_a(A_{t-1}^j - z_{t-1}^j) \\ &\quad + \alpha_o(\beta_o o^j - z_{t-1}^j), \\ A_t^j &= \gamma z_{t-1}^j + (1 - \gamma)A_{t-1}^j \end{aligned} \quad (7)$$

where  $\alpha_e, \alpha_a, \alpha_o, \beta_e, \beta_o$ , and  $\gamma$  are learnable model parameters. According to this model, the decision maker  $j$ ’s change of trust level from task  $t-1$  ( $z_{t-1}^j$ ) to task  $t$  ( $z_t^j$ ) is affected by (1) their most recent experience with the AI model (i.e., reflected in the feedback  $e_{t-1}^j$  received in task  $t-1$ ), (2) their accumulated trust in the AI model (i.e., reflected in  $A_{t-1}^j$ , the exponentially weighted moving average of the past trust levels), and (3) their expectation bias determined by their demographics  $o^j$ . Logistic regression models are again used as the decision model for this baseline model, and we ensure that a higher trust level  $z_t^j$  leads to a higher probability for the decision maker to rely on the AI recommendation.

## Comparing Model Performance

We first compare the performance of various computational models in capturing subjects’ reliance behavior in AI-assisted decision making, on both the population level and the individual level. To do so, we conduct a 5-fold cross-validation—We randomly divide all subjects into five groups

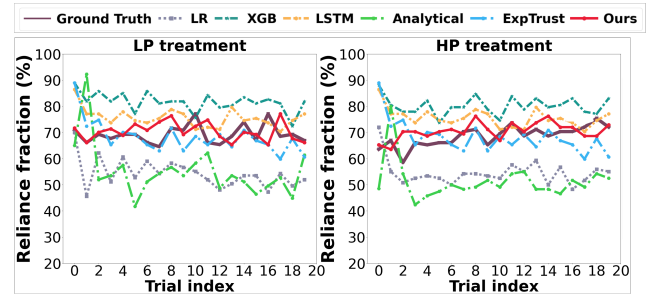


Figure 2: The actual (i.e., “data”) and predicted fractions of subjects who rely on the AI model over a period of 20 tasks.

and split the behavior dataset into five folds accordingly. Within each cross validation iteration, we use four folds to train the models with grid search being taken to find the best (hyper-)parameters for baseline models, while the learned model’s performance is evaluated on the remaining fold.

To begin with, we look into how the entire population of subjects adjust their reliance on the AI model *over time* in AI-assisted decision making, and we aim to examine how well different models can capture these patterns. In Figure 2, we plot the fraction of subjects who rely on the AI model (i.e., accept the AI recommendation) on each task over the course of the 20 decision making tasks in our experiment, as well as the same fractions computed from different models’ predictions<sup>2</sup>. Visually, it appears that in both the LP and HP treatments, the two models that characterize decision maker’s reliance decisions as affective processes (i.e., ExpTrust and our proposed Markov model) outperform other models in capturing subjects’ reliance dynamics over time (see Table S1 in the supplemental materials for the root-mean-square deviation computed for each model).

Since reliance on AI models reflects people’s affective responses to their experience with the AI model (e.g., the feedback  $e_{t-1}$  received), we next move on to examine how well different models can capture people’s reliance responses *to their interaction experience*. As discussed earlier, a decision maker can have one of the four experiences  $e_{t-1}$  in their previous task—appropriate acceptance, appropriate rejection, inappropriate acceptance, and inappropriate rejection. Given a particular type of experience  $e$  (e.g., appropriate acceptance), to see what people’s reliance response to this experience is, we first obtain the set of subjects’ reliance decisions  $d_t^j$  in the current task when their experience in the previous task  $e_{t-1}^j$  is  $e$ . This enables us to compute the probability for subjects to rely on the AI model when their previous experience is  $e$ . Using bootstrapping ( $R = 100000$ ) to re-sample this set of reliance decisions, we further obtain a bootstrapped distribution of subjects’ reliance probability when their previous experience is  $e$ . Moreover, by replacing the set of actual reliance decisions with the set of *predicted* reliance decisions given by a computational model (e.g., the

<sup>2</sup>In the 5-fold cross validation, predictions for subjects in one fold are made using the models trained on the other 4 folds, and the predicted reliance fraction is then averaged across all subjects.



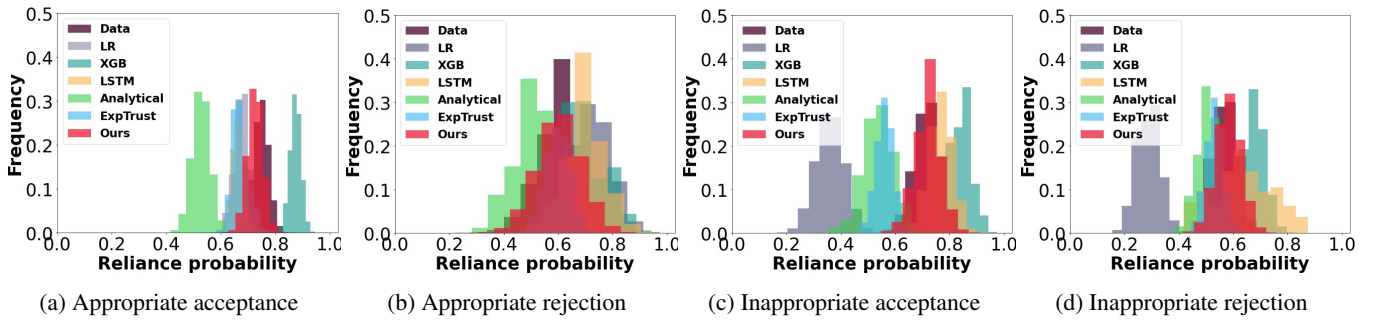


Figure 3: The actual and predicted distributions of subjects’ reliance probabilities under four interaction experiences.

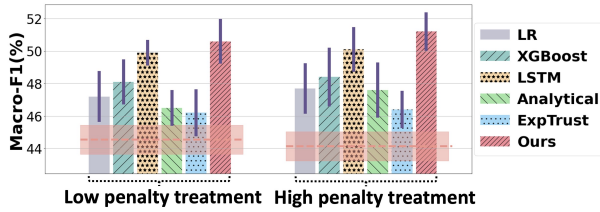


Figure 4: Model performance in predicting reliance decisions of individual subjects. The red dotted line represents the performance of the *heuristic* method, which predicts a subject’s reliance decision stochastically using the overall reliance probability observed in the training data. Error bars (shades) represent the standard errors of the mean.

proposed model), we can also obtain the bootstrapped distribution of subjects’ *predicted* reliance probability for that model. Figure 3 presents the bootstrapped distributions of subjects’ reliance probabilities under all four interaction experiences, computed using subjects’ reliance decisions obtained from both the actual experiment and different models’ predictions<sup>3</sup>. It is clear from the figures that for all types of interaction experiences, the predicted distributions of subjects’ reliance probabilities matches the ground truth distribution the best when the predictions are generated by our proposed hidden Markov model (see Table S1 in the supplemental materials for results on the Wasserstein distance between the ground truth distribution and the predicted distribution given by each computational model).

Finally, we explore how well different models can predict *individual* decision makers’ reliance decisions as they interact with the AI model in AI-assisted decision making. We use Macro-F1 score to measure the performance of each model, and the average Macro-F1 scores across the 5-fold cross validations for different models are shown in Figure 4. We again observe that our proposed model outperforms all baseline models in predicting individual decision maker’s reliance decisions. In particular, while ExpTrust (i.e., the baseline model in which the trust dynamics is explicitly specified) is sometimes on par with our proposed model in

predicting the reliance behavior of a *population*, its performance in predicting *individual*’s reliance decisions is significantly worse than the proposed model.

### The Importance of Accounting for Dependencies on Context and Feedback

As discussed earlier, conceptual models in the psychology literature suggest that during the interactions, people’s reliance on automated systems depends on both situational factors and their interaction experience. To reflect this, in our proposed hidden Markov model, we assume that both the task context  $x_t$  (i.e., reflecting situational factors) and the feedback received  $e_{t-1}$  (i.e., reflecting interaction experience) affect not only the transition of the trust state  $z_t$  in the trust dynamics model (TDM), but also the reliance decision  $d_t$  in the decision model (DM) directly. To gain a deeper understanding of how important it is to account for the dependencies on context and feedback in both TDM and DM, we conduct an ablation study. In particular, starting from the original full model structure (Model A), we construct four additional model structures by removing the dependencies on previous feedback  $e_{t-1}$  from TDM (Model B), removing the dependencies on  $e_{t-1}$  from DM (Model C), removing the dependencies on task context  $x_t$  from TDM (Model D), or removing the dependencies on  $x_t$  from DM (Model E). Figure 5 presents the performance of these 5 model structures in predicting individual decision maker’s reliance decisions in a 5-fold cross validation. Overall, as the performance decreases in Models B and C are larger than those in Models D and E compared to the full model, it implies that people’s interaction experience feedback  $e_{t-1}$  plays a larger role in determining their reliance behavior than the task context  $x_t$ . The significant drop that we see in the performance of both Models B and C suggest that the feedback that people get from their interaction experience may indeed influence their reliance decisions both directly and indirectly (i.e., through changing their trust). On the other hand, the observation that the predictive performance of the model does *not* significantly decrease after dependencies on  $x_t$  is removed from DM (i.e., Model E) suggests that the situational factors may mainly influence reliance decisions indirectly through affecting trust, rather than directly.

<sup>3</sup>See the supplemental materials for the results when distinguishing data obtained from HP and LP treatments.

Start \ End	HP-LP			Approp Acc-Inapprop Acc			Inapprop Rej-Approp Rej		
	low	medium	high	low	medium	high	low	medium	high
low	1.62	-0.88	-0.74	19.27	-23.37	4.10	-12.29	9.93	2.35
medium	-3.01	4.26	-1.24	-11.54	12.44	-0.90	-9.03	17.61	-8.57
high	-0.45	-2.17	2.62	-7.35	-7.51	14.86	-0.54	-0.37	0.91

Table 1: Trust state transition probability difference (%) for HP vs. LP, previous experience  $e_{t-1}$  being appropriate vs. inappropriate acceptance, or inappropriate vs. appropriate rejection. *Start* is  $z_{t-1}$  (i.e., the trust state in task  $t-1$ ), and *End* is  $z_t$ .

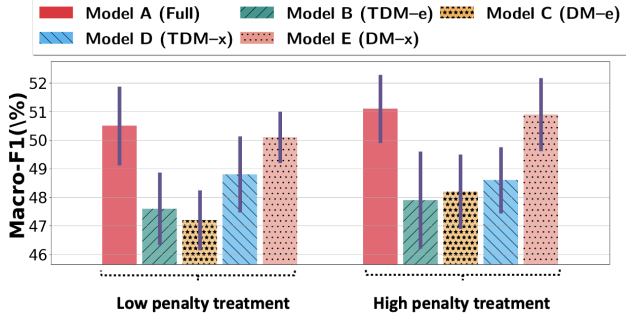


Figure 5: The predictive performance of five model structures. Error bars represent the standard errors of the mean.

## Explaining the Impacts of Context and Feedback

Finally, we aim to provide some quantitative explanations on how task context and feedback received during interactions may impact people’s reliance decisions. For this analysis, we train a hidden Markov model to fit the reliance behavior of *all* human subjects in our dataset, and we focus on analyzing how context and feedback influence the trust dynamics. To determine the trust “level” each of the  $K = 3$  categorical trust states in the model represents, we first utilize our human-subject dataset to compute the average probability for subjects to rely on the AI recommendation in a task (estimated by the decision models learned) when the subject’s trust state is set to each of the three values. We then sort the average reliance probabilities and label the three trust states as representing low, medium, or high trust levels, correspondingly (see more details in supplemental materials).

To begin with, we explore how situational factors included in task context such as the decision stakes affect trust transitions. For each of the 500 task instances used in our experiment, we use it as the current task (i.e., task  $t$ ) and compute the differences in the transition probabilities for all pairs of previous/current trust levels when varying the decision stake from low (LP) to high (HP), and these differences are averaged across all four possible feedback  $e_{t-1}$  that can be received from the previous task. Table 1 (the left section) reports the trust transition probability differences between low and high stakes, after averaging across all 500 task instances—the probability differences along the diagonal of the transition matrix are all positive, which suggests that people’s trust state are less likely to change (i.e., more stable) when decision stake is higher.

Using similar methods, we then explore the impacts of feedback on trust transitions by computing the trust transi-

tion probability differences when the decision maker accepts the previous AI recommendation while it turns out to be correct or wrong (Table 1 middle section), or when they previously reject the AI and the AI is correct or wrong (Table 1 right section). In both cases, we find the values *below* the diagonal of the transition matrices are consistently negative. This means that when the previous AI recommendation is correct, the chance for people to transit to a lower trust level is decreased compared to when the previous AI recommendation is wrong, regardless of whether the decision maker has accepted or rejected the previous AI recommendation. In contrast, when examining the sums of the values *above* the diagonal of the transition matrices, we find the sum is only positive in the matrix shown in the right section of Table 1. This implies that when an AI recommendation turns out to be correct rather than wrong, it seems to slightly increase the chance for people to transit to a higher trust level only if people previously have not relied on it (i.e., exhibit under-reliance) but not if they have relied on it.

## Conclusion and Discussion

In this paper, we propose a theory-based, hidden Markov model to characterize human’s trust and reliance dynamics in AI-assisted decision making. We evaluate the proposed model’s performance in fitting the real human behavior data collected from a randomized experiment. Our results show that the proposed model consistently outperforms other baselines in accurately predicting humans’ reliance behavior in AI-assisted decision making. Further analyses on the learned model allow us to provide insights into how human’s trust and reliance dynamics are influenced by contextual factors and people’s interaction experiences.

There are a few limitations of this study. For example, the proposed model may not generalize to settings where immediate AI performance feedback is not available. Also, the behavior data is collected from laypeople on the income prediction task, which is representative of common decision making tasks that do not require specialized knowledge. Whether the proposed model can perform well on tasks that require domain knowledge still needs to be explored. There are many interesting future directions for this work as well. For example, we are interested in personalizing the proposed model to capture the trust and reliance dynamics for different “types” of decision makers to better predict individual’s reliance decisions. Incorporating other emotions that may affect people’s reliance behavior into the model, and adjusting it to accommodate settings where other information about the AI recommendation (e.g., confidence, explanations) is presented are also interesting future work.

## Acknowledgements

We are grateful to the 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

- 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.
- 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*, volume 35, 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.
- Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–21.
- Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.
- Chiang, C.-W.; and Yin, M. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *27th International Conference on Intelligent User Interfaces*, 148–161.
- Forney, G. D. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3): 268–278.
- Gao, R.; Saar-Tsechansky, M.; De-Arteaga, M.; Han, L.; Lee, M. K.; and Lease, M. 2021. Human-ai collaboration with bandit feedback. *arXiv preprint arXiv:2105.10614*.
- Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.
- Hoff, K. A.; and Bashir, M. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3): 407–434.
- Hu, W.-L.; Akash, K.; Reid, T.; and Jain, N. 2018. Computational modeling of the dynamics of human trust during human-machine interactions. *IEEE Transactions on Human-Machine Systems*, 49(6): 485–497.
- Kerrigan, G.; Smyth, P.; and Steyvers, M. 2021. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems*, 34: 4421–4434.
- 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.; 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.
- Lee, J. D.; and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1): 50–80.
- Liao, Q. V.; and Sundar, S. S. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. *arXiv preprint arXiv:2204.13828*.
- 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.
- Madras, D.; Pitassi, T.; and Zemel, R. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 31.
- McLachlan, G. J.; and Krishnan, T. 2007. *The EM algorithm and extensions*. John Wiley & Sons.
- Mozannar, H.; Satyanarayan, A.; and Sontag, D. 2022. Teaching Humans When To Defer to a Classifier via Exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5323–5331.
- Park, J. S.; Barber, R.; Kirlik, A.; and Karahalios, K. 2019. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–15.
- 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.
- Rechtemmer, A.; and Yin, M. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*, 1–14.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics*, 461–464.
- Steyvers, M.; Tejeda, H.; Kerrigan, G.; and Smyth, P. 2022. Bayesian modeling of human-AI complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119.



- Tversky, A.; and Kahneman, D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4): 297–323.
- 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.
- Wilder, B.; Horvitz, E.; and Kamar, E. 2021. Learning to complement humans. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 1526–1533.
- 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.