Modeling Human Trust and Reliance in AI-Assisted Decision Making: A Markovian Approach (Supplementary Material)

Zhuoyan Li*, Zhuoran Lu*, Ming Yin
Purdue University, USA
li4178@purdue.edu, lu800@purdue.edu, mingyin@purdue.edu

Data and Codes

Codes and the human behavior dataset obtained from our human-subject experiments are available at https://github.com/xfleezy/AAAI23_human_trust.

Interface of the Income Prediction Tasks

Figure S1 shows an example of the task interface that human subjects in our experiment saw in the AI-assisted income prediction tasks. The current balance of the subject's bonus account is shown at the top-right corner. After the subject makes their prediction, we immediately provide them with feedback on whether their prediction is correct and how their account balance is changed (the remaining account balance will be updated once the subject proceeds to the next task).

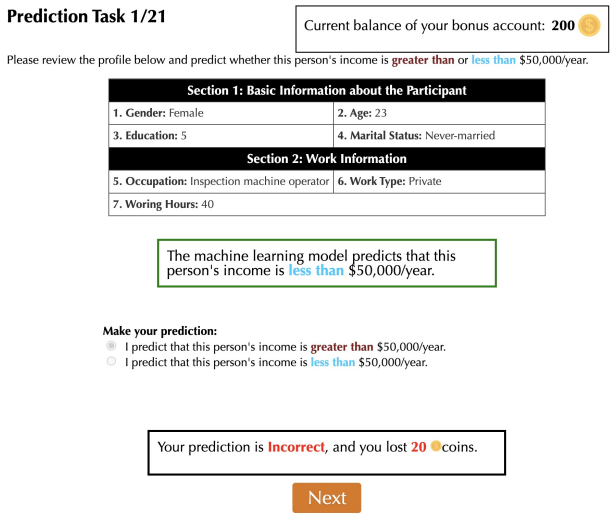


Figure S1: An example of the task interface.

Comparing Model Performance in Capturing Reliance Behavior of the Population

Corresponding to Figure 2 in the main paper, we used root-mean-square deviation (RMSE) to quantify each model's performance in predicting the subject population's reliance on the AI model over time. The results are shown in Table S1 (left two columns) and we find our proposed model achieves the lowest RMSE among all models.

Model	RMSE (×1e-2)		Wasserstein Distance (×1e-2)			
	LP treatment	HP treatment	appropriate		inappropriate	
			acceptance	rejection	acceptance	rejection
LR	16.1	15.8	7.6	12.4	26.0	19.9
XGBoost	14.0	12.6	12.2	11.3	11.7	10.1
LSTM	8.9	7.1	6.5	5.3	5.4	6.3
Analytical	21.6	21.9	22.4	8.8	24.1	6.4
ExpTrust	6.4	8.7	8.7	2.6	15.7	3.9
Ours	5.7	4.6	3.3	1.8	0.9	2.1

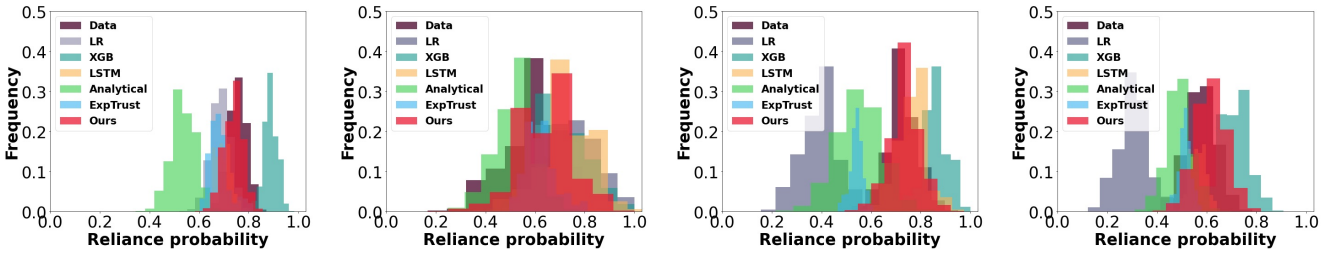
Table S1: The performance of different models in fitting the subject population's likelihood to rely on AI over time (evaluated via RMSE), or their reliance responses to various interaction experiences (evaluated via Wasserstein distance). The best method in each column is colored in blue.

In addition, corresponding to Figure 3 in the main paper, we formally measure the Wasserstein distance between the ground truth distribution and the predicted distribution given by each computational model on subjects' reliance probabilities under the four interaction experiences, and results are reported in Table S1 (right four columns). Again, our proposed model consistently achieves the smallest Wasserstein distance, which indicates that it fits subjects' reliance responses to their interaction experiences the best among all models we have examined.

Finally, we separate the data obtained from the high penalty (HP) treatment and the low penalty (LP) treatment, and again compare different computational model's performance in capturing subjects' reliance responses to their interaction experiences within each treatment.

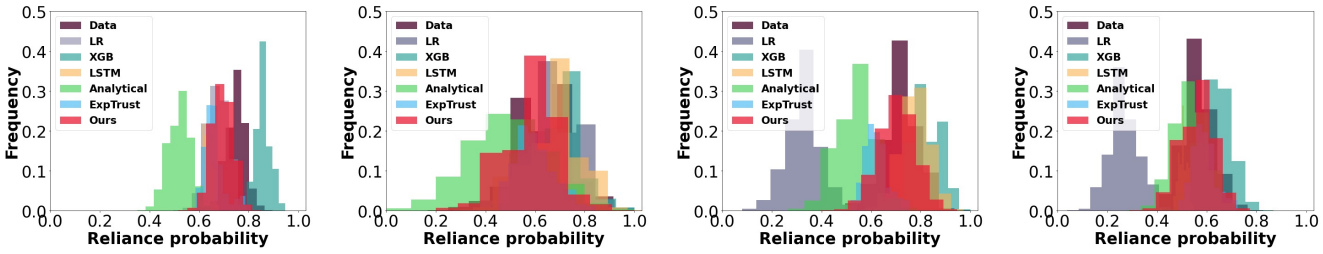
Given a particular treatment (e.g., LP) and a particular type of experience $\boldsymbol{e}$ (e.g., appropriate acceptance), to see what people's reliance response to this experience is, we first obtain the set of subjects' reliance decisions d_t^j in the current task when their experience in the previous task $\boldsymbol{e}_{t-1}^j$ is $\boldsymbol{e}$, from all data collected for this treatment. This enables us to compute the probability for subjects in this treatment to rely on the AI model when their previous experience is $\boldsymbol{e}$. 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

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



(a) Appropriate acceptance (b) Appropriate rejection (c) Inappropriate acceptance (d) Inappropriate rejection

Figure S2: The actual and predicted distributions of subjects' reliance probabilities under four interaction experiences in the low penalty treatment.



(a) Appropriate acceptance (b) Appropriate rejection (c) Inappropriate acceptance (d) Inappropriate rejection

Figure S3: The actual and predicted distributions of subjects' reliance probabilities under four interaction experiences in the high penalty treatment.

Model	Wasserstein Distance (×1e-2)			
	appropriate		inappropriate	
	acceptance	rejection	acceptance	rejection
LR	6.9	11.9	33.3	27.1
XGBoost	13.6	15.4	13.7	13.0
LSTM	8.1	8.9	3.6	10.4
Analytical	21.7	6.1	18.8	8.7
ExpTrust	6.9	7.4	15.8	3.3
Ours	1.4	4.9	1.2	3.1

Table S2: The performance of different models in fitting the subject population's reliance responses to various interaction experiences for subjects in the LP treatment. The best method in each column is colored in blue.

Model	Wasserstein Distance (×1e-2)			
	appropriate		inappropriate	
	acceptance	rejection	acceptance	rejection
LR	7.4	6.3	38.4	29.2
XGBoost	10.7	6.7	10.4	7.3
LSTM	4.8	5.1	6.1	3.2
Analytical	22.8	12.9	18.5	3.1
ExpTrust	8.7	5.2	12.4	2.7
Ours	4.5	3.9	1.3	0.5

Table S3: The performance of different models in fitting the subject population's reliance responses to various interaction experiences for subjects in the HP treatment. The best method in each column is colored in blue.

experience is $\boldsymbol{e}$. Moreover, by replacing the set of actual reliance decisions with the set of *predicted* reliance decisions given by a computational model (e.g., the proposed model), we can also obtain the bootstrapped distribution of subjects' *predicted* reliance probability for that model. The actual and predicted distributions of subjects' reliance probabilities under four interaction experiences for LP and HP treatments are illustrated in Figure S2 and Figure S3, respectively. Correspondingly, Table S2 and Table S3 report the Wasserstein distance comparisons between the ground truth distribution and the predicted distribution given by each computational models for the LP and HP treatments. Again, we find that the proposed model consistently outperforms other baseline models in accurately characterizing subjects' reliance responses to all four different types of possible interaction experiences, regardless of the level of decision stakes involved.

Labeling Trust State

To determine the trust "level" each of the $K = 3$ categorical trust states in the model represents, we estimate the probability for subjects to rely on the AI recommendation when they are in each of the three states. Specifically, given the behavior dataset $\boldsymbol{\mathcal{D}} = \{\boldsymbol{o}^j, \{\boldsymbol{x}_t^j, d_t^j, \boldsymbol{e}_t^j\}_{t=1}^T\}_{j=1}^N$ we collect from our experiment and a particular hidden trust state k, we set all subjects' trust state in all tasks to be k (i.e., let $z_t^j = k$ for all t, j), and we utilize the learned decision model (i.e., $\mathbb{P}(d_t^j = \textit{accept} | z_t^j = k, \boldsymbol{x}_t^j, \boldsymbol{e}_{t-1}^j; \boldsymbol{\theta}_{DM})$) to estimate the probability for subjects to rely on the AI recommendation on these tasks. Figure S4 shows the distributions of subjects' estimated reliance probabilities in all the tasks collected in our experiment, when their trust state is set to be each of the three possible values. For each trust state, we then compute subjects' average reliance probability and sort them in an increasing order. The average estimated reliance probabilities

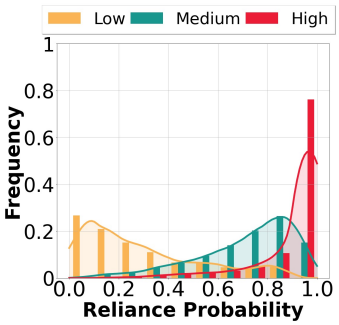


Figure S4: The distribution of subjects' estimated probabilities to rely on the AI recommendation across all tasks in our experiment, when their trust state is set to each of the three possible values. Depending on the average estimated reliance probabilities, we label the three trust states as low, medium, high trust levels.

for the three states are 0.29, 0.71, and 0.91, respectively, and we therefore determine that these three states represent the low, medium, and high trust levels correspondingly.