Strategic Adversarial Attacks in AI-assisted Decision Making to Reduce Human Trust and Reliance (Supplementary Materials)

Zhuoran Lu*, **Zhuoyan Li***, **Chun-Wei Chiang**, **Ming Yin**
Purdue University
{lu800, li4178, chiang80, mingyin}@purdue.edu

1 Empirical Examinations of Impacts of Adversarial Attacks on Humans: Additional Results

1.1 Task Interface

Figure 1 shows an example of the task interface that human subjects in our experiment saw in the AI-assisted bird species categorization tasks.

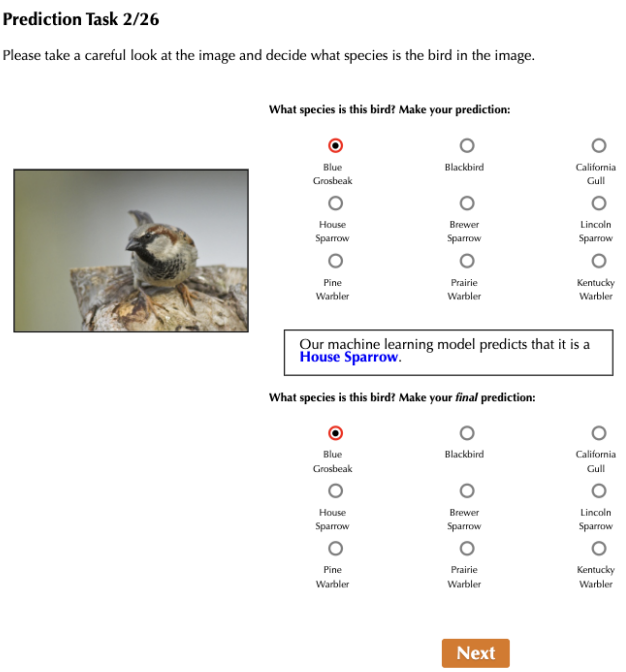
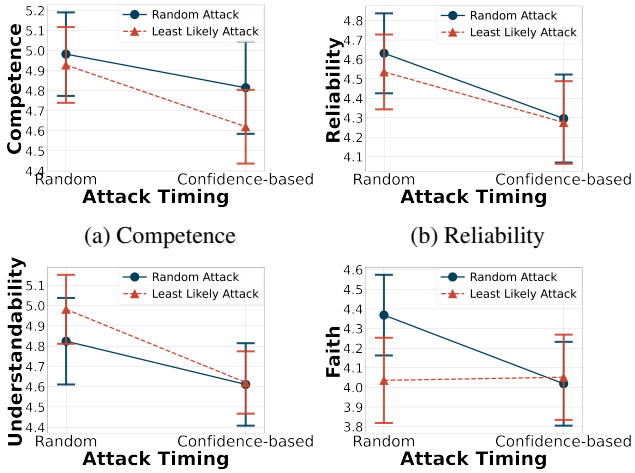


Figure 1: An example of the task interface.

1.2 Impact of attack timing and attack type on perceptions of the model

In our experiment, we asked subjects to rate their perceptions of the following statements on a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree):

- **(Competence)**: The recommendation that the AI model provides to me is as good as that which a highly competent person could provide.
- **(Reliability)**: The AI model provides the reliable recommendation to me in each task.
- **(Understandability)** I understand how the AI model will assist me with decisions I have to make.
- **(Faith)** If I am not sure about my decision in a task, I have faith that the AI model will provide the best solution.



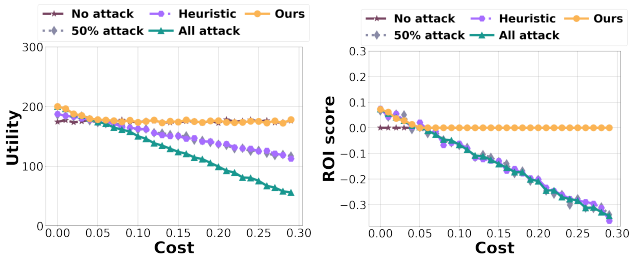
(a) Competence (b) Reliability

(c) Understandability (d) Faith

Figure 2: Average values of subject's assessment on model competence, reliability, understandability, as well as their faith in the AI model. Error bars represent the standard errors of the mean.

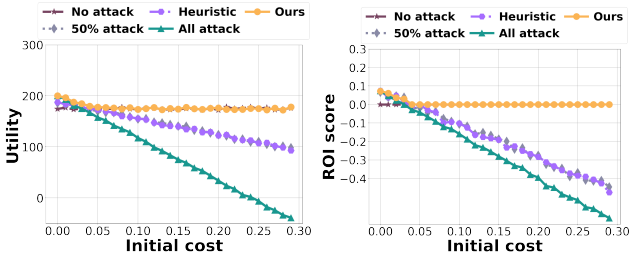
Comparisons on subjects' evaluations of the AI model's competence, reliability, understandability, and their faith in the AI model are shown in Figure 2a-2d. We again find a consistent trend that compared to attacks that are deployed on randomly selected tasks, confidence-based attacks appear to make subjects perceive the AI model as less competent, less reliable, and less understandable, and subjects also report lower levels of faith in the model, although our statistical tests suggest the differences are not significant.

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



(a) Overall utility (b) ROI score

Figure 3: Comparison of attack deployment strategies with the fixed attack cost for the Type 1 decision makers: Decision maker's behavior model is learned from the synthetic data of two decision maker types.



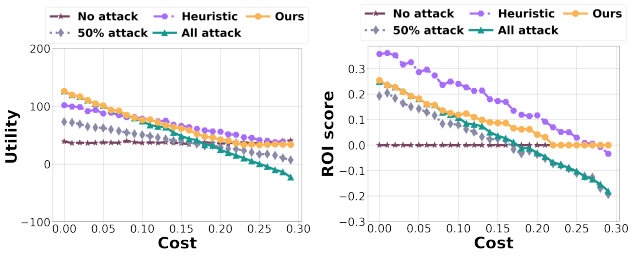
(a) Overall utility (b) ROI score

Figure 4: Comparison of attack deployment strategies with the increasing attack cost for the Type 1 decision makers: Decision maker's behavior model is learned from the synthetic data of two decision maker types.

2 Algorithmic Control of Attack Deployments: Additional Results

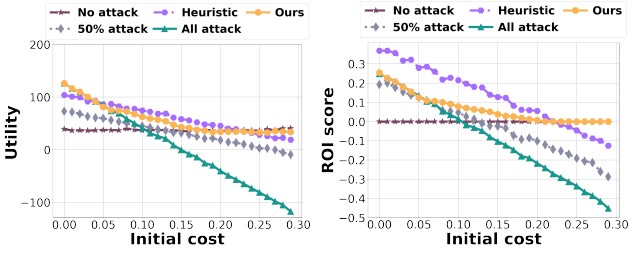
Consider the Behavior Model II discussed in the main paper for characterizing human decision makers' reliance behavior on the AI model in AI-assisted decision making under adversarial attacks, which includes two types of decision makers—Type 1 is skeptical of AI and has low reliance on AI in general, while Type 2 is quite willing to rely on AI except for if they observe the AI to be obviously "wrong". In the main paper, we report the performance of different attack deployment strategies when the attacker needs to deploy attacks when Type 1 and 2 decision makers each accounts for half of the decision maker population. Here, we take a zoomed-in look at the performance of different attack deployment strategies only for Type 1 (Figures 3 and 4) and Type 2 decision makers (Figures 5 and 6) in the population, separately.

For both types of decision makers, we can see that when adversarial attacks are deployed based on our proposed strategy, the attacker could almost always achieve the highest utility when compared to using the best baseline strategy, regardless of the level of the cost of the attack. In fact, as shown in Figure 7, as the cost of attack increases, our proposed strategy quickly learns to reduce the number of attacks deployed for Type 1 decision makers, since Type 1 decision makers have low trust/reliance on AI anyway, and the small reduction in their trust/reliance brought up by the adversarial attacks may not even compensate for the cost. In contrast, the number



(a) Overall utility (b) ROI score

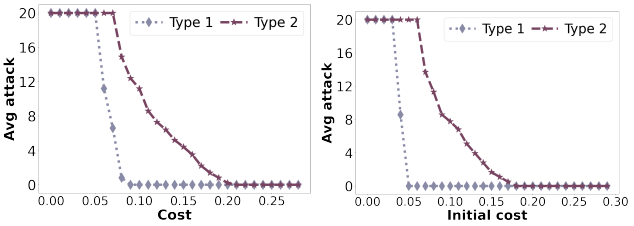
Figure 5: Comparison of attack deployment strategies with the fixed attack cost for the Type 2 decision makers: Decision maker's behavior model is learned from the synthetic data of two decision maker types.



(a) Overall utility (b) ROI score

Figure 6: Comparison of attack deployment strategies with the increasing attack cost for the Type 2 decision makers: Decision maker's behavior model is learned from the synthetic data of two decision maker types.

of attacks deployed for Type 2 decision makers decreases at a much slower rate as the cost of the attack increases. We also note that for Type 2 decision makers, when the cost of the attack is relatively low, as shown in Figures 5b and 6b, the heuristic attack deployment strategy (i.e., only deploy attacks on high confidence tasks) appear to achieve higher ROI scores and be more efficient than our proposed strategy. This is because our strategy is designed to maximize the attacker's *utility*, so following our strategy, beyond attacking on high-confidence tasks, the attacker would deploy attacks on some low-confidence tasks even if the "marginal returns" of these attacks may not be as high. However, when the attack cost becomes very high, we again find our strategy outperforms the heuristic strategy on the efficiency of the attacks as it guides the attacker to conduct no attacks to avoid the high cost.



(a) Fixed cost (b) Changeable cost

Figure 7: Comparison of the average number of attacks deployed by our strategy for the two types of decision makers under both fixed and changeable cost scenarios.