Strategic Adversarial Attacks in AI-assisted Decision Making to Reduce Human Trust and Reliance

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

Abstract

With the increased integration of AI technologies in human decision making processes, adversarial attacks on AI models become a greater concern than ever before as they may significantly hurt humans' trust in AI models and decrease the effectiveness of human-AI collaboration. While many adversarial attack methods have been proposed to decrease the performance of an AI model, limited attention has been paid on understanding how these attacks will impact the human decision makers interacting with the model, and accordingly, how to strategically deploy adversarial attacks to maximize the reduction of human trust and reliance. In this paper, through a human-subject experiment, we first show that in AI-assisted decision making, the *timing* of the attacks largely influences how much humans decrease their trust in and reliance on AI—the decrease is particularly salient when attacks occur on decision making tasks that humans are highly confident themselves. Based on these insights, we next propose an algorithmic framework to infer the human decision maker's hidden trust in the AI model and dynamically decide when the attacker should launch an attack to the model. Our evaluations show that following the proposed approach, attackers deploy more efficient attacks and achieve higher utility than adopting other baseline strategies.

1 Introduction

Artificial Intelligence (AI) technology has undergone tremendous growth recently. As a result, AI-based decision aids have been widely utilized to aid people in decision making in diverse domains, and it is found that AI recommendations can often help human decision makers make more accurate decisions [Lai and Tan, 2019; Bansal *et al.*, 2021b]. However, recent research raises the concerns that AI models can be quite vulnerable to adversarial attacks [Szegedy *et al.*, 2013]. For example, researchers have found that a single piece of black electrical tape placed on a speed limit sign caused a self-driving car's computer vision system to misclassify the sign [Povolny and Trivedi, 2020]. These adversarial attacks pose a significant threat to the performance and security of AI-based decision aids, and they may significantly affect the effectiveness of human-AI collaboration in AI-assisted decision making. Indeed, observing an AI model's failures often makes people lose trust in the model and decrease their reliance on them, even if the AI model has high performance overall [Lee *et al.*, 2021; Dietvorst *et al.*, 2015]. This presents possibilities for attackers to intentionally mislead human decision makers to not trust or rely on an AI model by deploying adversarial attacks to the model during the AI-assisted decision making process.

Recent research has proposed a wide variety of adversarial attack methods that aim to deceive the AI model into producing erroneous outputs through small perturbations that are imperceptible to humans [Eykholt *et al.*, 2018; Xiao *et al.*, 2018; Evtimov *et al.*, 2017]. In contrast, relatively limited attention has been paid to understand how these attacks, when deployed to an AI model, will impact the human decision makers who are assisted by the model, especially in terms of their trust in and reliance on the model. The long line of empirical research in the human-computer interaction community on understanding humans' interactions with AI in AI-assisted decision making [Lu and Yin, 2021; Chong *et al.*, 2022; Nourani *et al.*, 2020], however, suggests that likely, not all attacks are created equal. For example, it is found that the disagreement between AI recommendations and the human decision makers' own judgment on tasks that humans are highly confident in often results in a significant decrease in humans' trust and reliance [Lu and Yin, 2021]. This implies that the operationalizations of adversarial attacks in AI-assisted decision making, such as on what tasks the attacks are deployed, may largely affect the "effectiveness" of the attacks.

Therefore, in this paper, we start by conducting a randomized human-subject experiment to understand that in AI-assisted decision making, how the ways that adversarial attacks are deployed affect decision makers' trust in and reliance on the AI model. Our experiment considers two specific aspects in the deployment of attacks—the *timing* of the attacks (i.e., randomly selected vs. selected based on human confidence), and the *types* of attacks (i.e., deceive the model to produce a random decision vs. the "least likely" decision). Our results show that attacking the tasks that people have high

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

confidence in results in a larger reduction of their trust and reliance on the AI model, while the type of the attack does not appear to have a significant impact.

Building on these findings as well as the fact that adversarial attacks often come with cost (e.g., risk of exposing the attacker), we then propose an algorithmic framework to enable attackers to make online decisions on *when* to launch attacks on AI models during the AI-assisted decision making process, with the goal of maximizing the attacker's utility in terms of reducing human trust and reliance on the AI model while accounting for the cost of attacks. To do so, we first learn an input-output hidden Markov model to infer the decision maker's latent trust level in the AI model, and then dynamically decide whether the attacker should launch an attack on the next decision making task by comparing the expected maximum utility of different actions. We evaluate the effectiveness of our approach against a few baseline fixed or heuristic attack deployment strategies on several datasets that reflect real-world human behavior. Compared to baselines, our approach consistently enables attackers to obtain higher utility and achieve larger gains from each deployed attack.

2 Related Work

Adversarial attacks in AI/machine learning. Adversarial attacks, in which malicious inputs are crafted to deceive or mislead AI models, have gained significant attention in recent years. Various adversarial attack methods [Yuan *et al.*, 2019] have been proposed, such as Projected Gradient Descent (PGD) [Madry *et al.*, 2017] and Instance Attribute Editing method via generative models [Qiu *et al.*, 2020]. Many studies have highlighted the negative impacts of adversarial attacks on a range of real-world AI applications, including medical pathology analysis [Ghaffari Laleh *et al.*, 2022] and vision-based autonomous driving systems [Jia *et al.*, 2020]. While previous research has largely focused on designing adversarial attack and defense methods [Samangouei *et al.*, 2018] and understanding their impacts on the AI model's performance, our work focuses on understanding the impacts of adversarial attacks on *humans* who interact with the AI model, and we make an initial attempt to examine how the ways that attacks are deployed affect humans' trust in and reliance on the AI model in AI-assisted decision making.

AI-assisted decision making. The wide use of AI-based decision aids has inspired many empirical studies examining how humans interact with and rely on AI models in AI-assisted decision making. Researchers have identified a variety of factors that may influence an individual's trust in AI models, including the perceived model accuracy [Lai and Tan, 2019], the level of human-model agreement [Lu and Yin, 2021], model confidence [Rechkemmer and Yin, 2022], and the perceived transparency and explainability of the model's decision making process [Zhang *et al.*, 2020; Bansal *et al.*, 2021b]. Multiple studies have demonstrated that decision makers may struggle to appropriately trust and rely on AI models, leading to research on designing innovative methods to promote appropriate trust and reliance on AI in AI-assisted decision making [Buçinca *et al.*, 2021; Chiang and Yin, 2022; Ma *et al.*, 2023]. More recently, researchers have started to quantitatively model humans' trust and reliance on AI models, with the goal of designing AI models that are more compatible with decision makers and can better support humans in decision making [Wang *et al.*, 2022; Bansal *et al.*, 2021a; Tejeda *et al.*, 2022; Li *et al.*, 2023]. Different from previous research, our focus in this paper is on identifying strategies for attackers to effectively reduce humans' trust/reliance on AI models—this allows us to better understand the limitations and vulnerabilities of the current AI-assisted decision making process and informs us on possible defense.

3 Empirical Examinations of Impacts of Adversarial Attacks on Humans

We start by investigating that in AI-assisted decision making, how the ways that adversarial attacks on the AI model are operationalized impact the human decision makers who interact with the model, especially on their trust in and reliance on the AI model. To do so, we conducted a randomized human-subject experiment on Amazon Mechanical Turk (MTurk).

3.1 Decision Making Task

In our experiment, we asked human subjects to categorize the species of birds in images [Wah *et al.*, 2011b; Kim *et al.*, 2022], with the assistance of an AI model. Specifically, we focus on 9 categories of birds (e.g., sparrow, gull), and obtained a dataset of bird images of these 9 categories from the Caltech Bird dataset [Wah *et al.*, 2011a] with around 60 images per category. Given this dataset, we randomly generated the training, validation, and test sets, using a $75\%/5\%/20\%$ split. We then trained our AI model M by finetuning a pretrained ResNet model [He *et al.*, 2016] using the training and validation data. The accuracy of M was 96% when evaluated on the test set. Then, in each task, the human subject was presented with a bird image (randomly sampled from the test data) and was asked to identify the bird's species among 9 possible categories, while the AI model M's prediction was provided to the subject for their reference. In other words, during the experiment, the model M will serve as the target for the attacker, who will attempt to deceive the model by injecting adversarial noise into the image in some decision making tasks, thereby decreasing human decision makers' trust in and reliance on the model.

3.2 Experimental Treatments

Earlier empirical studies on humans' trust in and reliance on AI models in AI-assisted decision making suggest that human decision makers tend to use the level of decision agreement between the AI model and themselves on tasks that they have high confidence in as a heuristic to gauge the trustworthiness of the AI model, and adjust their level of reliance on the model accordingly [Lu and Yin, 2021]. Inspired by these findings, in this experiment, we focus on understanding how two aspects in the deployment of adversarial attacks to an AI model impact human decision makers who are assisted by it—*timing of the attacks* (i.e., on what tasks the attacks are deployed?) and *type of the attacks* (i.e., how "wrong" the AI model looks when the attacks are deployed?). Specifically, we created 4 experimental treatments arranged in a 2×2 factorial design varying along two factors:

- **Attack timing:** Given N decision making tasks that humans need to work on with the assistance of the AI model M, the adversarial attacks are either deployed on a *random* subset of n tasks, or on the subset of n tasks that humans have *high confidence* in their own decisions.
- **Attack type:** When an attack is deployed on a task for which the AI model M's original predicted category is y, the attacker could add adversarial noise into the image in the task so that the AI model either predicts another *random* category $y' \neq y$, or predicts the "*least likely*" category y_L for the task.

Implementation of adversarial attacks. In our experiment, we assume the attacker conducts *white box attacks* [Akhtar and Mian, 2018], i.e., the attacker has full knowledge about the deployed AI model including its training dataset, inputs and outputs. Specifically, attacks in our experiment are realized via Projected Gradient Descent (PGD) [Madry *et al.*, 2017], a common gradient-based white box attack method. For PGD, given an input (e.g., the image of a bird), the targeted label needs to be specified, and the algorithm will then perturb the input image to force the AI model into predicting the targeted label. Therefore, to conduct a "random attack" on a decision making task, the targeted label is randomly selected from all categories except for the original one that would have been predicted by the AI model M. On the other hand, to conduct a "least likely attack" on a task, we compare the embeddings of the input image in this task—which is generated by the pre-trained ResNet—with the embeddings for each of the 9 categories, which are obtained by averaging across embeddings of all training samples belonging to the same category. We then identify the category y_L that the image in the current task is *least* similar with using cosine similarity, and use y_L as the targeted label for this task.

Moreover, to deploy attacks on tasks based on human confidence in some treatments, we assume that the attacker can recruit human subjects to make independent decisions for tasks in the training dataset of the AI model M and then learn a predictive model of human confidence based on the human decision data collected. In our experiment, we conducted a pilot study in which 179 human subjects were recruited. Each subject was asked to complete a set of 20 bird species categorization tasks *on their own*, while the bird images in these tasks were again randomly sampled from our dataset. Then, given a decision making task $\boldsymbol{x}_i \in \mathcal{X}$ (i.e., a bird image), we used s_i—the fraction of human subjects whose decision on this task is the same as the majority decision—as a proxy of human decision makers' confidence in this task. That is,

$$s_i = \frac{\max_{y \in \mathcal{Y}} \sum_{j=1}^{J_i} \mathbb{1}(r_i^j = y)}{J_i}$$

where $\mathcal{Y} = \{1, \cdots, Y\}$ is the set of all possible decisions (i.e., bird species), J_i is the total number of human subjects who worked on task $\boldsymbol{x}_i$, and r_i^j is the j-th subject's decision on task $\boldsymbol{x}_i$. Intuitively, a larger value of s_i indicates that people are more likely to reach a consensus for the corresponding task $\boldsymbol{x}_i$, which could imply that the task is relatively "easy" and people are more confident in their decisions. Using human subjects' decisions that we collected in the pilot study on tasks in the training dataset, we again fine-tuned a pre-trained ResNet model to obtain a predictive model $M_{\text{conf}} : \mathcal{X} \rightarrow s \in \{0, 1\}$ to predict humans' confidence on decision making tasks (we simplified it to a binary prediction task by using a median split to convert s_i into low/high confidence). We found that the accuracy of M_{conf} was 87% on the test set, and M_{conf} was then used for determining attack timing when attack deployments were confidence-based.

3.3 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, subjects were randomly assigned to one of the four experimental treatments. The experiment began with a tutorial which explained the decision making task to subjects and provided subjects with visual examples of the 9 bird species that they needed to identify during the experiment. After completing the tutorial, subjects were asked to complete a total of 26 bird species categorization tasks. In each task, the subject followed a three-step procedure: (1) First, the subject was asked to make an *initial* prediction independently. (2) Next, the subject was presented with bird species prediction given by the AI model M. (3) Finally, the subject was asked to make a *final* prediction, and they could freely decide whether their final prediction would be the same as their initial prediction or not.

The 26 tasks in the experiment were divided into two phases, where Phase 1 contained 16 tasks while Phase 2 included the remaining 10 tasks. For the $N = 16$ tasks in Phase 1, we randomly sampled them from the test dataset and ensured that half of them were tasks that human decision makers have high confidence in according to our pilot study results (i.e., s_i is above the median value), while for the other half of tasks, humans have low confidence in them. We then selected $n = 5$ tasks out of all 16 tasks in Phase 1 to deploy the adversarial attacks. The timing and type of the attacks were decided by the treatment that the subject was assigned. In particular, when the subject was assigned to treatments that deploy attacks on tasks where humans have high confidence, we used the model M_{conf} to predict humans' decision confidence in each task, and selected the 5 tasks with the highest likelihood of being high confidence to attack.

After completing the 16 tasks in Phase 1, subjects were asked to take a pause and fill out a survey to report their perceptions of the AI model's competence, reliability, and understandability, as well as their faith in the model, on a 7-point Likert scale. Then, in Phase 2, subjects were asked to complete the remaining 10 tasks with the assistance of the AI model M. Phase 2 was designed to measure how the timing and type of adversarial attacks in Phase 1 changed humans' trust in and reliance on the AI model. Thus, we did *not* deploy any attack to the AI model in Phase 2. Further, Phase 2 tasks were again sampled from the test dataset, but within the subset for which $0.33 \leq s_i \leq 0.67$ (i.e., humans' confidence in the task was neither too low or too high)[1]. Subjects in all

[1]We conjectured that in tasks that humans' confidence in their own decisions is neither too low or too high, the chance of observing variations in humans' reliance on the AI model due to their varying

(a) Trust (b) Reliance fraction (c) Switch fraction

Figure 1: Comparing the average values of subjects' trust in the AI model, and their reliance and switch fractions in Phase 2 across the four experimental treatments. Error bars represent the standard errors of the mean.

treatments saw the same set of 10 tasks in Phase 2, on which the AI model M's accuracy was 100%. Finally, after subjects completed all Phase 2 tasks, we asked them to complete a final survey to report their trust level in the AI model.

The base payment for the HIT was \$1.2. The HIT was open only to workers based in the US, and each worker could only take it once. Additionally, we included two attention check questions in the HIT, and only the data of those subjects who passed both attention checks was considered valid.

3.4 Experimental Results

After filtering the subjects who did not pass the attention check, we obtained valid data from 225 subjects for our experiment. To examine how the adversarial attack timing and type impact humans' trust in and reliance on the AI model in AI-assisted decision making, we used subjects' self-reported trust level in the final survey to quantify their trust in the AI model. To measure humans' reliance on the AI model, we adopted two metrics that are widely used in earlier research [Zhang *et al.*, 2020; Lai *et al.*, 2021]:

- **Reliance fraction**: In Phase 2, the fraction of tasks on which the subject's final decision was the same as the AI model's decision.
- **Switch fraction**: In Phase 2, the fraction of tasks on which the subject's final decision was the same as the AI model's decision, among all tasks that the subject's initial decision was different from that of the model's.

Clearly, for both metrics, the higher the value, the more the human decision maker relies on the AI model.

Figure 1 shows the comparisons on subjects' trust in the AI model as well as their reliance and switch fractions in Phase 2 across different treatments. Visually, we found that compared to adversarial attacks that are deployed on a random set of tasks, the ones that are deployed on tasks that humans have high confidence in appear to result in a larger decrease in the human decision maker's trust in and reliance on the AI model. In contrast, the type of the adversarial attacks does not appear to result in significant differences in humans' trust and reliance in our experiment. Our two-way ANOVA[2] test results further showed that the confidence-based attacks lead to a marginally larger decrease in humans' trust ($p = 0.07, \eta^2 = 0.015$) and a significantly larger decrease in humans' reliance fraction ($p = 0.031, \eta^2 = 0.021$) than attacks deployed on random tasks, while the difference in switch fraction is insignificant. We then looked into human subjects' self-reported perceptions of the AI model's competence, reliability, and understandability, as well as their faith in the AI model in the mid-point survey. Similar to our findings on trust and reliance, we also found a consistent trend that after experiencing attacks on tasks in which they had high confidence rather than on random tasks, subjects tended to consider the AI model as less reliable, less competent, less understandable, and they also had less faith in the AI model. Still, the type of attacks did not seem to change subjects' self-reported perceptions of the AI model much (see supplemental materials for details).

These results suggest that the *timing* of adversarial attacks in AI-assisted decision making can indeed affect humans' perceptions of and reactions to the AI model. This means that the attacker may be able to increase their utility (e.g., maximize the reduction of humans' trust in and reliance on the AI model) by strategically determining when to deploy the attacks. Meanwhile, we suspect that our lack of findings on the impacts of attack type on humans might be due to the representations learned by the neural models (i.e., ResNet) do not align well with human perceptions—what the model considers as distinct may be perceived as similar by humans. Future studies are needed to verify whether this finding still holds true when more human-compatible representations are used to select target labels for adversarial attacks.

4 Algorithmic Control of Attack Deployments

Results of our experimental study suggest that in AI-assisted decision making, the attacker may achieve a larger gain by deploying the adversarial attacks to the AI model on decision making tasks where humans are highly confident. However, the deployment of each attack may come with some cost (e.g., the cost of perturbing the task input, the risk of exposing the attacker). Thus, given a human decision maker who needs to complete a sequence of decision making tasks with the assistance of an AI model, a natural question to ask is how can an *attacker* dynamically and strategically determine *when* to deploy adversarial attacks to the AI model to maximize the reduction of the human decision maker's trust in and reliance

levels of trust in the model is the highest.

[2]Analysis of Variance (ANOVA) is a statistical test for identifying significant differences between group means.

on the AI model while taking the cost of attacks into consideration. To answer this question, we propose an algorithmic framework for controlling the deployment of adversarial attacks in AI-assisted decision making.

4.1 Modeling the Effects of Adversarial Attacks on Human Trust and Reliance

First, we aim to quantitatively characterize that in a sequence of AI-assisted decision making tasks, how the deployment of adversarial attacks impacts the human decision maker's trust in and reliance on the AI model. Following Li *et al.* [2023], we used an input-output hidden Markov model M_{trust} to do so. Specifically, consider a decision maker who needs to complete a sequence of T decision making tasks with the aid of an AI model. For each task t ($1 \leq t \leq T$), the decision maker is first provided with the task $\boldsymbol{x}_t \in \mathcal{X}$ and the AI model's decision recommendation $y_t^m \in \mathcal{Y}$ on it. Then, the decision maker needs to make a final decision $y_t^h \in \mathcal{Y}$ by deciding whether to accept or reject the AI model's recommendation (i.e., $d_t \in \{\text{accept}, \text{reject}\}$). Our model M_{trust} assumes that in each task t, there is a hidden "trust state" $z_t \in \mathcal{Z} = \{1, 2, \cdots, K\}$ (K is the total number of states) that reflects the decision maker's trust in the AI model. In addition, we define the "context" for each task t as $\boldsymbol{v}_t = (s_t, a_t)$, where $s_t \in \{0, 1\}$ is the decision maker's confidence on task t, while $a_t \in \{0, 1\}$ represents whether an adversarial attack is deployed on task t. With this, the details of the input-output hidden Markov model can be summarized as follows:

- **Inputs**: $\boldsymbol{v}_t = (s_t, a_t)$, $t = 1, 2, \cdots, T$, where $s_t \in \{0, 1\}$ (0 means decision makers have low confidence in the task) and $a_t \in \{0, 1\}$ (0 means an adversarial attack is not deployed in the task).
- **Outputs**: $d_t \in \{0, 1\}, t = 1, 2, \cdots, T$ (0 means the decision maker rejects the AI recommendation).
- **Hidden States**: $z_t \in \mathcal{Z} = \{1, 2, \cdots, K\}, t = 1, 2, \cdots, T$; each value reflects a different trust state.
- **Transition Probability** (or Trust Dynamics Model): $\mathbb{P}_{TDM}(z_t | z_{t-1}, \boldsymbol{v}_t) = \mathbb{P}_{TDM}(z_t | z_{t-1}, s_t, a_t)$; e.g., $\mathbb{P}_{TDM}(z_t = k' | z_{t-1} = k, s_t = 0, a_t = 1)$ captures the likelihood of a decision maker transiting from latent trust state k in task $t-1$ to state k' in task t, given that the decision maker has low confidence in task t and an adversarial attack is deployed in task t.
- **Emission Probability** (or Decision Model): $\mathbb{P}_{DM}(d_t | z_t, \boldsymbol{v}_t) = \mathbb{P}_{DM}(d_t | z_t, s_t, a_t)$; e.g., $\mathbb{P}_{DM}(d_t = 1 | z_t = k, s_t = 0, a_t = 1)$ indicates the probability of a decision maker relying on the AI recommendation in task t, when their current latent trust state is k, they have low confidence in task t and an adversarial attack is deployed on task t [3].

With a set of training data collected from Q human decision makers on their reliance behavior in AI-assisted decision making under adversarial attacks, i.e., $\mathcal{D}_{\text{train}} = \{\{\boldsymbol{x}_t^q, a_t^q, d_t^q\}_{t=1}^T\}_{q=1}^Q$, we can first estimate decision maker q's confidence on task t using the predictive model M_{conf} (i.e., $s_t^q = M_{\text{conf}}(\boldsymbol{x}_t^q)$), and then learn the input-output hidden Markov model M_{trust} via the expectation-maximization (EM) algorithm [McLachlan and Krishnan, 2007].

4.2 Attacker Utility Maximization

Given a learned model M_{trust} that characterizes the impacts of adversarial attacks on decision makers' trust and reliance, we next explore how the attacker should deploy the adversarial attacks (i.e., determine the values of a_t) dynamically and strategically. As discussed earlier, when deploying adversarial attacks in AI-assisted decision making, the attacker needs to trade-off between the "gains" brought up by the reduction of decision makers' trust in the AI model, which can be reflected in their decreased level of reliance on the AI model, and the "cost" of the attacks. Without loss of generality, we assume that the attacker can observe the decision maker's reliance decisions (i.e., d_t) in the AI-assisted decision making tasks as they deploy the attacks, and the attacker gets a utility of w_a (or w_r) whenever they observe the decision maker accepting (or rejecting) the AI model's decision recommendation, while the cost for deploying an attack is c. We further assume that the attacker has a quasi-linear utility function:

$$U = w_a N_{accept} + w_r N_{reject} - c N_{attack}$$

where N_{accept} (or N_{reject}) represents the number of times the human decision maker accepts (or rejects) the AI model's recommendations, and N_{attack} represents the number of adversarial attacks deployed. The attacker's goal is to control the deployment of attacks on the AI model in a sequence of T tasks in an online fashion—after completing the first t_c tasks and obtaining the human decision maker's interaction history so far (i.e., $\{s_t, a_t, d_t\}_{t=1}^{t_c}$), the attacker observes the content of the $(t_c + 1)$-th task $\boldsymbol{x}_{t_c+1}$ (thus human confidence on this task can be estimated, $s_{t_c+1} = M_{\text{conf}}(\boldsymbol{x}_{t_c+1})$) and needs to determine whether or not to deploy an adversarial attack on it in order to maximize their utility.

We solve this utility maximization problem using the proposed behavior model M_{trust}, and the attacker's utility function. Specifically, when the decision maker's current trust state is $z = k$ and their predicted confidence in the next task is s, the utility for the attacker to take action a to deploy ($a = 1$) or not deploy ($a = 0$) an attack in the next task is given by $\mathcal{R}(k, s, a) = \mathbb{E}_{k' \sim \mathbb{P}_{TDM}(\cdot | k, s, a)}[\mathbb{P}_{DM}(1 | k', s, a) w_a + \mathbb{P}_{DM}(0 | k', s, a) w_r - \mathbb{1}(a = 1) c]$. However, after completing t_c tasks, instead of knowing the decision maker's precise trust state, the attacker can only estimate a *distribution* of the trust states using M_{trust}, and we denote this distribution as the "state belief" $\boldsymbol{b}(t_c) = (b(1), \ldots, b(K))$. Given a state belief $\boldsymbol{b}$, the attacker's expected utility when humans' confidence is s and they take action a in the next task is then defined as $\rho(\boldsymbol{b}, s, a) = \mathbb{E}_{k \sim \boldsymbol{b}}[\mathcal{R}(k, s, a)]$.

Further, we note that after completing t_c tasks, the attacker needs to decide whether to deploy an attack on the next task to maximize their utility in the rest $T - t_c$ tasks, instead of just the immediate next task. So, we define $EU_{max}(\boldsymbol{b}, s, a, l)$ as the *maximum* expected utility that the attacker can obtain when there are l more tasks to complete, while the belief of the decision maker's current trust state distribution is $\boldsymbol{b}$, the

[3] While human decision makers will not directly know whether an adversarial attack is deployed on a task, they may indirectly perceive it through y_t^m, thus it may affect trust and reliance.

predicted human confidence on the next task is s, and the deployment of adversarial attack on the next task is decided by a. Given this definition, it is clear that after completing t_c tasks, the attacker's optimal decision for the deployment of an adversarial attack on the next task should be calculated as $a_{t_c+1} = \text{argmax}_{a\in\{0,1\}} EU_{max}(\boldsymbol{b}(t_c), s_{t_c+1}, a, T - t_c)$. This means that the attacker should launch an attack on the next task when $a_{t_c+1} = 1$, and vice versa.

The exact values of the maximum expected utility EU_{max} can be computed recursively. For the case where there is only one task remaining, i.e. $l = 1$, we have $EU_{max}(\boldsymbol{b}, s, a, l) = \rho(\boldsymbol{b}, s, a)$. Otherwise, we have:

$$EU_{max}(\boldsymbol{b}, s, a, l) = \rho(\boldsymbol{b}, s, a) + \sum_{d\in\{0,1\}} (\mathbb{E}_{k\sim\boldsymbol{b}}[\mathbb{E}_{k'\sim\mathbb{P}_{TDM}(\cdot|k,s,a)}[\mathbb{P}_{DM}(d|k', s, a)V(\boldsymbol{b}'_{s,a,d}, l-1)]])$$

where $V(\boldsymbol{b}, l) = \max_{a\in\{0,1\}} EU_{max}(\boldsymbol{b}, s, a, l)$. Since the attacker can not observe the task content beyond the immediate next task, we assume that humans' confidence for future tasks follows a Bernoulli distribution with probability 0.5, i.e. $s \sim \mathcal{B}(0.5)$. Therefore, $V(\boldsymbol{b}, l) = \max_{a\in\{0,1\}} \mathbb{E}_{s\sim\mathcal{B}(0.5)}[EU_{max}(\boldsymbol{b}, s, a, l)]$. In addition, $\boldsymbol{b}'_{s,a,d}$ is the attacker's updated state belief, when their initial belief is $\boldsymbol{b}$, their attack decision on the next task (on which humans' confidence is s) is a, and the decision maker's reliance decision on that task is d. $\boldsymbol{b}'_{s,a,d}$ can be computed as follows:

$$b'_{s,a,d}(k') \propto \mathbb{E}_{k\sim\boldsymbol{b}}[\mathbb{P}_{TDM}(k'|k, s, a)\mathbb{P}_{DM}(d|k', s, a)]$$

Finally, after the attacker implements their attack decision a_{t_c+1} on the $t_c + 1$-th task and observes the decision maker's reliance decision d_{t_c+1}, the attacker will update their state belief as $\boldsymbol{b}(t_c + 1) = \boldsymbol{b}'_{s_{t_c+1}, a_{t_c+1}, d_{t_c+1}}$.

Note that the time complexity to compute the exact values of $EU_{max}(\boldsymbol{b}, s, a, l)$ is exponential, which is intractable in practice especially for a long task sequence. Therefore, we set a limit τ on the maximum number of tasks that the attacker will look ahead in order to make their attack deployment decision in the next task. That is, the optimal decision for the deployment of an attack on the next task $t_c + 1$ can be approximated as: $a_{t_c+1} = \text{argmax}_{a\in\{0,1\}} EU_{max}(\boldsymbol{b}(t_c), s_{t_c+1}, a, N')$ where $N' = \min(\tau, T - t_c)$.

4.3 Evaluations

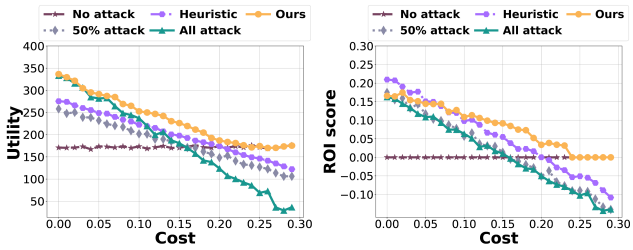
To examine whether attackers can improve their utility by following the proposed algorithmic framework to deploy adversarial attacks, we conduct a set of evaluations.

Baselines. We consider the following attack deployment strategies as our baselines: 1) *No attack*: no attack will be deployed on any task. 2) *50% random attack*: an attack will be deployed on each task with a probability of $p = 0.5$. 3) *Heuristic*: for each task, an attack will be deployed if M_{conf} predicts humans' confidence on it to be high. 4) *All attack*: attacks will be deployed on all tasks.

Evaluation procedure. Our evaluations are conducted on synthetic datasets. In particular, to generate the evaluation data, we assume that human decision makers' reliance behavior on the AI model in AI-assisted decision making under adversarial attacks follows a behavior model M_{behavior}. As detailed later, M_{behavior} may reflect real human behavior as we observed in our human subject experiment, or reflect some assumed human behavior. Given M_{behavior}, we first generate a training dataset $\mathcal{D}_{\text{train}}$, reflecting 500 decision makers each completing a set of 20 randomly sampled bird categorization tasks, while adversarial attacks are deployed on 50% of these tasks and decision makers' reliance on the AI model in each task is decided by M_{behavior}. Using this training dataset $\mathcal{D}_{\text{train}}$, we can learn the hidden Markov model M_{trust}.

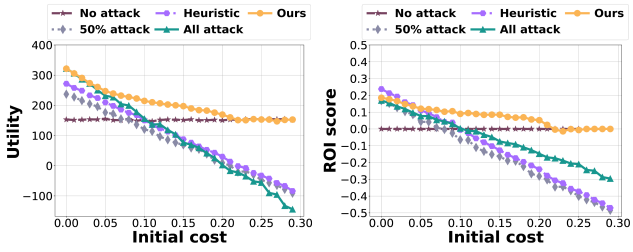
We next simulate five groups of test datasets, each corresponding to a unique attack deployment strategy (i.e., 4 baselines + the proposed). For each group, we simulate 50 decision makers each completing a set of 20 bird categorization tasks, while the deployment of attacks on each task is controlled by the strategy used for that group, and decision makers' reliance on the AI model is again decided by the behavior model M_{behavior}. In our evaluations, we set $w_a = 0, w_r = 0.5, \tau = 2$. We also consider two different scenarios for the cost of attacks: "fixed" and "changeable". In the fixed cost scenario, the cost c for each attack is a constant, and we vary it from 0 to 0.3 to examine how the performance of different attack deployment strategies changes under attack cost of different magnitude. However, in the changeable cost scenario, we assume the value of c increases exponentially with the number of attacks, as represented by $c_t = (1 + \alpha)^{N_{attack}(t)} c_0$—$N_{attack}(t)$ is the number of attacks deployed before task t, α is a hyperparameter that controls the growth of cost and is set to 0.05 in our evaluation, and c_0 is the cost for the first attack (we also experiment with different c_0 values from 0 to 0.3 in our evaluations). We repeat the simulations 5 times and compare the performance of different attack deployment strategies on: (1) the attacker's average utility obtained on the test dataset (2) the average cost-effectiveness ratio of each attack (ROI), which is calculated as the difference between the attacker's utility obtained by following the given strategy X and by following the "No attack" strategy, divided by the number of attacks deployed in strategy X. For both metrics, larger values indicate better performance.

Results for Behavior Model I: M_{behavior} reflects real-world human behavior. As the first evaluation, we utilize the data that we collected from Phase 1 of our human subject experiment in Section 3 to learn a model M_{trust} that can reflect the trust and reliance behavior of the real-world human subjects in AI-assisted decision making under adversarial attacks. We then use M_{trust} as our behavior model M_{behavior} to generate the synthetic evaluation datasets. Figure 2 shows the comparisons across different attack deployment strategies when the attack cost is fixed. As shown in the figure, when the cost is low, placing an attack brings up significantly more gains than cost for the attacker, and our method helps attacker to obtain similar utility as the baseline "All attack" strategy. In contrast, when the attack cost is very high, our method results in similar utility as the baseline "No attack" strategy. Importantly, when the cost is moderate, our method consistently results in a higher attacker utility compared to any other baseline strategies. It is also evident from Figure 2b that when following the proposed strategy to deploy attacks, the cost-effectiveness ra-



(a) Overall utility (b) ROI score

Figure 2: Comparison of attack deployment strategies with fixed attack cost: Decision maker's behavior model is learned from the data of our human-subject experiment.

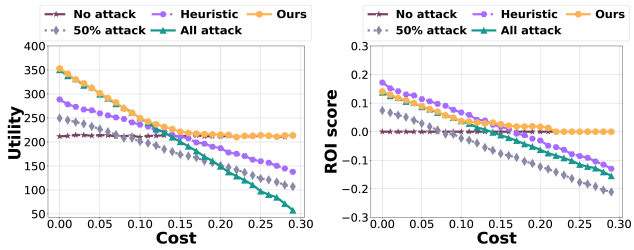


(a) Overall utility (b) ROI score

Figure 3: Comparison of attack deployment strategies with increasing attack cost: Decision maker's behavior model is learned from the data of our human-subject experiment.

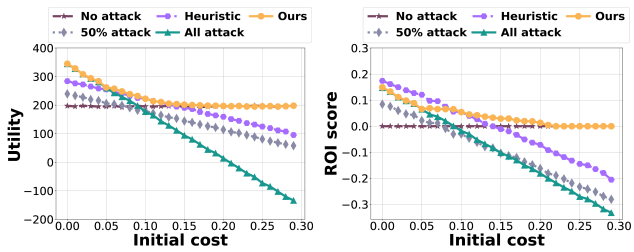
tio of each attack is almost always the highest, especially for the moderate level of cost, which indicates that the attacks are the most "efficient." Similar evaluation results for the changeable cost scenario are presented in Figure 3. Again, we find our approach consistently outperforms other baseline strategies in helping the attacker achieve higher utility and deploy more efficient attacks.

Results of Behavior Model II: M_{behavior} reflects behavior of two types of decision makers. Our Behavior Model I assumes that all decision makers adjust their trust and determine their reliance on AI based on the same input-output hidden Markov model. In reality, however, there may exist different "types" of decision makers who respond to AI recommendations in different ways [Meissner and Keding, 2021]. To examine whether the proposed method can adequately capture the behavior of a population of decision makers of different types and deploy attack strategies effectively, we construct a second behavior model for generating the evaluation dataset. Here, we assume the probability for a decision maker j to rely on the AI model on a task is $acc_j^{s,a}$, where s is their confidence in the task and a represents if an attack is deployed on the task. We then created two distinct types of decision makers based on this assumption. The first type of decision maker is rather skeptical of AI and has low level of reliance on AI under all circumstances. As a result, we set $acc_i^{1,0} = 0.4$, $acc_i^{1,1} = 0.3$, $acc_i^{0,0} = 0.2$, and $acc_i^{0,1} = 0.1$ for this type of decision makers. The second type, however, is quite willing to rely on AI in general, except for if they observe the AI to be obviously "wrong". Thus, we have



(a) Overall utility (b) ROI score

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



(a) Overall utility (b) ROI score

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

$acc_i^{1,0} = 0.9, acc_i^{1,1} = 0.2, acc_i^{0,0} = 0.8, acc_i^{0,1} = 0.6$ for them. When generating our training and test datasets, each decision maker is drawn from these two types uniformly randomly. Again, the performance of our proposed method is evaluated against other baselines under both the fixed and changeable cost scenarios, and the results, as depicted in Figure 4 and Figure 5, demonstrate that our method outperforms all baselines, further validating its robustness (see the supplemental materials for more details).

5 Conclusions

In this paper, we explore how attackers may strategically deploy adversarial attacks in AI-assisted decision making to reduce human decision makers' trust in and reliance on the AI model. Using a randomized human-subject experiment, we first show that the timing of adversarial attacks largely influence how much humans decrease their trust and reliance, with the decrease being particularly salient when attacks occur on decision making tasks where humans are highly confident themselves. We then propose an algorithmic framework to enable attackers to dynamically and strategically determine the deployment of adversarial attacks in AI-assisted decision making, taking both the gains coming from humans' reduction of trust and reliance and the cost of attacks into account. Multiple evaluations show that compared to using other fixed or heuristic attack deployment strategies, our approach helps attackers obtain higher utility from their attacks.

Ethical Statement

This study was approved by the Institutional Review Board of the authors' institution. Through this work, we hope to draw the community's attention to the fact that attackers may leverage their understandings on humans' trust dynamics in AI models to strategically deploy adversarial attacks in AI-assisted decision making and negatively influence human decision makers' perceptions and behavior. In addition, we hope insights revealed in this study can inform more effective defense methods to protect the effectiveness of human-AI collaborations in AI-assisted decision making. For example, as attackers may believe the decision making tasks where humans are highly confident as the most efficient targets to deploy attacks on, defenders should either use additional efforts to ensure the robustness of AI model's outputs on these tasks, or to deploy honeypots on these tasks to detect attackers.

Acknowledgments

We thank the support of the National Science Foundation under grant IIS-1850335 and IIS-2229876 on this work. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

References

[Akhtar and Mian, 2018] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

[Bansal *et al.*, 2021a] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11405–11414, 2021.

[Bansal *et al.*, 2021b] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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*, pages 1–16, 2021.

[Buçinca *et al.*, 2021] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 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, 2021.

[Chiang and Yin, 2022] Chun-Wei Chiang and Ming Yin. Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In *27th International Conference on Intelligent User Interfaces*, pages 148–161, 2022.

[Chong *et al.*, 2022] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice. *Computers in Human Behavior*, 127:107018, 2022.

[Dietvorst *et al.*, 2015] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

[Evtimov *et al.*, 2017] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2(3):4, 2017.

[Eykholt *et al.*, 2018] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

[Ghaffari Laleh *et al.*, 2022] Narmin Ghaffari Laleh, Daniel Truhn, Gregory Patrick Veldhuizen, Tianyu Han, Marko van Treeck, Roman D Buelow, Rupert Langer, Bastian Dislich, Peter Boor, Volkmar Schulz, et al. Adversarial attacks and adversarial robustness in computational pathology. *Nature Communications*, 13(1):1–10, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Jia *et al.*, 2020] Yunhan Jia Jia, Yantao Lu, Junjie Shen, Qi Alfred Chen, Hao Chen, Zhenyu Zhong, and Tao Wei Wei. Fooling detection alone is not enough: Adversarial attack against multiple object tracking. In *International Conference on Learning Representations (ICLR'20)*, 2020.

[Kim *et al.*, 2022] Sunnie SY Kim, Nicole Meister, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Hive: evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision*, pages 280–298. Springer, 2022.

[Lai and Tan, 2019] Vivian Lai and Chenhao Tan. 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*, pages 29–38, 2019.

[Lai *et al.*, 2021] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

[Lee *et al.*, 2021] Jieun Lee, Genya Abe, Kenji Sato, and Makoto Itoh. Developing human-machine trust: Impacts of prior instruction and automation failure on driver trust in partially automated vehicles. *Transportation research part F: traffic psychology and behaviour*, 81:384–395, 2021.

[Li *et al.*, 2023] Zhuoyan Li, Zhuoran Lu, and Ming Yin. Modeling human trust and reliance in ai-assisted decision making: A markovian approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[Lu and Yin, 2021] Zhuoran Lu and Ming Yin. 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*, pages 1–16, 2021.

[Ma *et al.*, 2023] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[McLachlan and Krishnan, 2007] Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2007.

[Meissner and Keding, 2021] Philip Meissner and Christoph Keding. The human factor in ai-based decision-making. *MIT Sloan Management Review*, 63(1):1–5, 2021.

[Nourani *et al.*, 2020] Mahsan Nourani, Joanie King, and Eric Ragan. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 112–121, 2020.

[Povolny and Trivedi, 2020] Steve Povolny and Shivangee Trivedi. Model hacking adas to pave safer roads for autonomous vehicles. *McAfee Advanced Threat Research*, 2020.

[Qiu *et al.*, 2020] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 19–37. Springer, 2020.

[Rechkemmer and Yin, 2022] Amy Rechkemmer and Ming Yin. 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*, pages 1–14, 2022.

[Samangouei *et al.*, 2018] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, 2018.

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[Tejeda *et al.*, 2022] Heliodoro Tejeda, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. Ai-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior*, 5(4):491–508, 2022.

[Wah *et al.*, 2011a] C Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011. Technical Report CNS-TR-2011-001, 2011.

[Wah *et al.*, 2011b] Catherine Wah, Steve Branson, Pietro Perona, and Serge Belongie. Multiclass recognition and part localization with humans in the loop. In *2011 International Conference on Computer Vision*, pages 2524–2531. IEEE, 2011.

[Wang *et al.*, 2022] Xinru Wang, Zhuoran Lu, and Ming Yin. Will you accept the ai recommendation? predicting human behavior in ai-assisted decision making. In *Proceedings of the ACM Web Conference 2022*, pages 1697–1708, 2022.

[Xiao *et al.*, 2018] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018.

[Yuan *et al.*, 2019] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.

[Zhang *et al.*, 2020] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 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*, pages 295–305, 2020.