A Decision Theoretic Framework for Measuring AI Reliance

Ziyang Guo
Northwestern Univerisity
Evanston, Illinois, USA
ziyang.guo@northwestern.edu

Yifan wu
Northwestern Univerisity
Evanston, Illinois, USA
yifan.wu@u.northwestern.edu

Jason Hartline
Northwestern Univerisity
Evanston, Illinois, USA
hartline@northwestern.edu

Jessica Hullman
Northwestern Univerisity
Evanston, Illinois, USA
jhullman@northwestern.edu

ABSTRACT

Humans frequently make decisions with the aid of artificially intelligent (AI) systems. A common pattern is for the AI to recommend an action to the human who retains control over the final decision. Researchers have identified ensuring that a human has appropriate reliance on an AI as a critical component of achieving complementary performance. We argue that the current definition of appropriate reliance used in such research lacks formal statistical grounding and can lead to contradictions. We propose a formal definition of reliance, based on statistical decision theory, which separates the concepts of reliance as the probability the decision-maker follows the AI's recommendation from challenges a human may face in differentiating the signals and forming accurate beliefs about the situation. Our definition gives rise to a framework that can be used to guide the design and interpretation of studies on human-AI complementarity and reliance. Using recent AI-advised decision making studies from literature, we demonstrate how our framework can be used to separate the loss due to mis-reliance from the loss due to not accurately differentiating the signals. We evaluate these losses by comparing to a baseline and a benchmark for complementary performance defined by the expected payoff achieved by a rational decision-maker facing the same decision task as the behavioral decision-makers.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Machine learning, reliance, decision making, rational decision-maker

ACM Reference Format:
Ziyang Guo, Yifan wu, Jason Hartline, and Jessica Hullman. 2024. A Decision Theoretic Framework for Measuring AI Reliance. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3630106.3658901

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658901

1 INTRODUCTION

AI-advised decision making, in which a human decision-maker has access to the recommendation of an artificial intelligence (AI system) and can choose whether or not to follow it, is often preferred as a means of retaining human control [3] in deploying predictive models. The motivation behind this approach is *complementary performance*; i.e., the human-AI team can outperform the AI or the human alone. However, many studies have shown that human-AI teams under-perform the AI alone in tasks where the AI's accuracy is higher than humans [3, 4, 6, 14, 18, 20, 21, 28]. One solution to this problem is to identify ways to ensure that the human, as the final decision-maker, has *appropriate reliance* on AI. Appropriate reliance is typically defined as submitting the AI recommendation when it is correct and not submitting it when it is not correct.

We argue that this definition of reliance lacks formal statistical grounding, leading to contradictions. For example, situations in which a human-AI team outperforms the human alone but underperforms the AI alone suggest that the human underrelies on the AI [3]. However, when researchers apply the above definition of appropriate reliance to their experimental results, they discover that the primary source of performance loss stems from the humans accepting the AI's inaccurate recommendations [6, 18, 21], considered over-reliance by the conventional definition.

Implicit in discussions of complementarity are assumptions of a human with some internal model of the data-generating process and an AI with its own model. Studying reliance implies that the human consults the AI recommendation, infers the probability that its decision is correct, then decides whether it is worth following its recommendation. Problems arise because defining appropriate reliance as submitting the AI's recommendation when it is correct and rejecting it when it is not confounds two challenges a human may face in an AI-advised decision-making: that of forming correct beliefs about the probability that the AI is correct, and that of making the optimal decision about whether to follow the AI conditional on one's beliefs. Without a definition that allows separation of different sources of performance loss, the analysis might misinterpret the reasons behind seemingly poor experiment results, leading researchers to prioritize less directly relevant follow-up actions for improving the team. For example, if the human has inaccurate beliefs about the probability that the AI is correct, this might stem from a lack of information about the prior probability

that the AI is correct (potentially addressable by providing the AI's accuracy on held-out data [33]), or from their failure to arrive at an accurate estimate of the AI's probability of being correct (potentially fixable via cognitive forcing functions [5, 12] or better explanations [3]). If the human correctly perceives the accuracy of the AI model, but uses the wrong decision rule to decide when to follow its recommendation, then the human may not understand the utility of different possible outcomes (e.g., a differential cost of using the AI's recommendation versus generating their own), or the researcher studying real-world human-AI teams may have assumed a utility function different from that used by the participant.

Another issue with the conventional definition of appropriate reliance is that it is a binary measure. Consequently, researchers cannot distinguish whether the human decision-maker mistakenly used (or did not use) the AI's recommendation in a situation where (A) the probability that relying on their own judgment would have been correct is similar to the probability that the AI was correct versus (B) very different. Intuitively, over-reliance is a bigger concern in B than in case A. We argue that the concept of reliance should be characterized within a continuous payoff space to allow for more fine-grained assessment.

We propose a formal definition of AI reliance. Following previous work on generating benchmarks for studies of information displays [31], our approach is grounded in statistical decision theory. Our definition separates the concepts of a reliance level (the probability that the human decision-maker goes with the AI recommendation) from the belief updating that a rational decision-maker is expected to do upon viewing an instance and associated AI recommendation. The framework we provide defines a benchmark for complementary performance representing the maximum attainable performance with the cooperation of AI and human and a baseline for complementary performance representing the maximum performance without any cooperation. We apply the framework to three well-regarded AI-advised decision making experiments from literature [3, 12, 21]. In all three cases, we show 1) that examining the results against the baseline and benchmark for complementary performance better reveals the limits of human behavioral performance and 2) specific sources of behavioral loss that help explain the experiment results but were not accounted for by the original interpretations of the results.

2 FORMULATING ASSUMPTIONS FOR STUDYING RELIANCE

In AI-advised decision-making scenarios [2, 29], the human makes a decision about a set of instances with the assistance of an AI recommendation. In formulating our definition of reliance below, we make several assumptions about this scenario:

(1) The human makes their own prediction about each instance prior to seeing the AI recommendation for that instance.

(2) The human consults the AI recommendation prior to making their decision.

There are two benefits to making these assumptions for AI-advised decision-making experiments. First, the assumptions ensure that participants neither anchor solely on the AI recommendations (completely neglecting to consider their own predictions) nor that they neglect to consult the AI recommendation at all [5, 12]. It is difficult to conceive of reliance in such cases. Second, and most importantly, by assuming we have access to the human's own prediction prior to their interaction with the AI recommendation, we can compare the results of experiments we run to a benchmark of complementary performance, which is attained by optimally combining the information contained in the human's predictions with that contained in the AI's recommendations, and a baseline of using either the AI or human only. We use human recommendation to refer to the human prediction prior to interaction with the AI recommendation.

3 DEFINITION OF RELIANCE

We define *appropriate reliance*, *over-reliance*, and *under-reliance* on AI recommendations in AI-advised decision making. Our framework conceives of three roles in the decision problem: a human recommender, an AI recommender, and a decision-maker. The two recommenders provide informational input to the decision-maker in the form of recommendations. The decision-maker chooses which recommender to follow on a decision task.

To formalize a decision task requires five key elements (Table 1): payoff-related states on which the decision is evaluated, a data genertaing model that generates the states and signals that inform about the state, the action, the information (i.e. signal) given to the decision-maker, and a scoring rule assessing the choice of action under the payoff-related state.

We define the reliance level of an decision-maker on the AI as the overall probability that she chooses the AI recommendation, conditional on the decision maker facing different recommendations from the human and the AI. The definition targets a conditional probability, because the reliance level cannot be defined when the human makes the same recommendation as the AI.

Definition 1 (Reliance). *The reliance level γ of any decision-maker on the AI is defined as the conditional probability $\gamma = \Pr[a = y^{AI} | y^{AI} \neq y^{H}]$ that the decision-maker chooses the AI recommendation, conditional on the AI recommendation y^{AI} being different from the human recommendation y^{H}.*

3.1 Rational Decision-Maker

We define the rational decision-maker in a binary-adoption decision task (Table 1) derived from the original one. This derived decision task limits the rational decision-maker to making a final decision by selecting between the human recommendation and the AI recommendation. We define the rational benchmark representing the expected performance of a rational Bayesian decision-maker who perfectly perceives the provided information in the signal and chooses the optimal action under the scoring rule for each decision task. The rational benchmark is the maximum payoff that can be expected from a behavioral decision-maker, i.e., the benchmark for complementary performance. Following the framework proposed by Wu et al. [31], we also define a baseline for expected performance using this rational Bayesian decision-maker. The rational baseline is the maximum payoff that can be expected from the behavioral decision-maker when they must choose between always going with either the AI or the human recommender, i.e., they do not consult the individual signals in making their decisions. The rational baseline represents the minimum threshold for achieving complementary performance, i.e., the baseline for complementary

	The original decision task	The derived binary-adoption decision task
Payoff-related state	θ = Ground truth $y \in Y$	$\hat{\theta} = (y, y^H, y^{AI})$ Ground truth $y \in Y$ Human recommendation $y^H \in Y$ AI recommendation $y^{AI} \in Y$
Data generating model	Feature values x from feature space X $(x, y) \sim \pi(X \times Y)$ Human recommendation y^H and AI recommendation y^{AI}: $(x, y^H) \sim \pi^H(X \times Y)$ $(x, y^{AI}) \sim \pi^{AI}(X \times Y)$ Explanation $e(y^{AI})$	
Action (choice)	$a \in Y$	$\hat{a} \in \{0 = \text{human}, 1 = \text{AI}\}$
Signal	$v = \{x, y^H, y^{AI}, e(y^{AI})\}$	
Scoring rule (payoff)	$S(a, \theta)$	$\hat{S}(\hat{a}, \hat{\theta}) = S(y^H, y)$ if $\hat{a}$ = human $\hat{S}(\hat{a}, \hat{\theta}) = S(y^{AI}, y)$ if $\hat{a}$ = AI

Table 1: Notation for original decision task and derived binary-adoption decision task in our framework.

performance. Using the rational benchmark and the rational baseline, we define the value of rational complementation, representing the expected improvement in payoff to a rational decision-maker that the joint human+AI setting provides over the better of either the AI or the human alone.

These three values construct a space of payoffs within which behavioral participants' performance can be quantified and compared. The rational benchmark also describes the appropriate reliance level, which maximizes the expected payoff. Throughout the paper, we use superscript r to denote notation for the rational decision-maker. For example, a^r is the action taken by the rational decision maker, and γ^r the rational decision-maker's reliance level.

• **Rational Baseline**, The rational baseline is the expected performance of the rational decision-maker without access to the signal on a randomly chosen decision task from the experiment. Without access to the signal, the rational decision-maker can only make decisions with prior beliefs based on her knowledge of the data-generating model and decision task. This is the better of the two scores achieved by the human alone and the AI alone.

$$\mathrm{R}_{\varnothing} = \max_{\hat{a}} \mathbf{E}_{\pi(\hat{\theta})}[\hat{S}(\hat{a}, \hat{\theta})] = \max_{\hat{a}} \mathbf{E}_{\pi(\hat{\theta})}[S(y^{\hat{a}}, y)].$$

• **Rational Benchmark**, The rational benchmark is the expected performance of the rational decision-maker with the signal on a randomly chosen decision task from the experiment. Let $a^r(v)$ be the action taken by the rational decision-maker given signal v. She chooses a^r to maximize her expected utility with $\pi(\hat{\theta}|v)$, the distribution of the payoff-related state conditioned on the signal v:

$$\mathrm{R} = \max_{a^r(\cdot)} \mathbf{E}_{\pi(v,\hat{\theta})}[\hat{S}(a^r(v), \hat{\theta})].$$

The rational benchmark upperbounds the expected performance of any behavioral decision-maker in the experiment.

• **Value of rational complementation**, The value of rational complementation is the increase in payoff over the rational baseline when the rational decision-maker sees the signal.

$$\Delta = \mathrm{R} - \mathrm{R}_{\varnothing}.$$

The value of rational complementation provides a scale for comparing expected performance in terms of the "lift" we see from having access to the information in the signals. In the context of AI-advised decision making, it also represents the maximum improvement of performance we can expect from a complementation of the human and the AI conditioned on the information structure of the signals. If we treat Δ as a comparative unit by normalizing all scores within the range where the baseline $\mathrm{R}_{\varnothing}$ is 0 and the benchmark R is 1, we get a sense of the proportion of possible score increase that different settings provide. For example, we could compare expected human performances B_{α} and B_{β} under two conditions α and β (e.g., α explanation and β explanation) by calculating $(\mathrm{B}_{\alpha} - \mathrm{B}_{\beta})/\Delta$.

Given the definitions above, we can define the appropriate reliance level as the reliance level of the rational decision-maker, conditional on the human recommendation being different from the AI recommendation, $y^H \neq y^{AI}$. Note that the appropriate reliance level maximizes the expected score of the decision.

Definition 2. *The **appropriate reliance level** γ^r is the rational decision-maker's reliance level on the AI,* $\gamma^r = \Pr[a^r = 1 | y^{AI} \neq y^H]$.

3.2 Behavioral Decision-Maker

The behavioral decision-maker who completes the decision task takes action a^b, and is evaluated by their expected performance on the task. We view the behavioral action as a random variable correlated with the signal, and hence also with the ground truth. Denote the joint distribution as $\pi(v, a^b, \theta)$.

• **Behavioral Performance**

$$\mathrm{B} = \mathbf{E}_{\pi(v,a^b,\theta)}[S(a^b, \theta)].$$

We define behavioral *under-reliance* and *over-reliance* by comparing behavioral reliance level γ^b to the appropriate reliance level γ^r.

Definition 3. *When $\gamma^b < \gamma^r$, the behavioral decision-maker **under-relies** on the AI.*

Definition 4. *When $\gamma^b > \gamma^r$, the behavioral decision-maker **over-relies** on the AI.*

In addition to the reliance level, we analyze the difference between the behavioral decision-maker's expected score and the rational decision-maker's expected score to measure decision quality.

To understand why we analyze the difference in score versus in the action space, consider the extreme case where the human recommender and the AI recommender are both uninformative about the ground truth. Adopting either the AI recommendation or the human recommendation would achieve an equally bad expected payoff, such that any reliance level between 0% and 100% would perform similarly. Simply evaluating the reliance level by comparing to the best reliance level ignores the close payoffs achieved by all reliance levels and leads to misleading conclusions.

We separate the behavioral decision-maker's loss in score into two sources: loss from mis-reliance, and what we term discrimination loss, referring to the loss from not accurately distinguishing when the AI recommender has better expected payoff than the human recommender or vice versa. To separate these sources of loss, we define another benchmark representing the expected score of a rational decision-maker who is constrained to a specific reliance level.

• **Mis-Reliant Rational benchmark** The expected score of a rational decision-maker with reliance level γ:

$$\begin{aligned} \mathrm{R}^{\mathrm{m}}(\gamma) = \max_{a^r(\cdot)} \quad & \mathbf{E}_{\pi(\hat{v},\hat{\theta})}[\hat{S}(a^r(\cdot),\hat{\theta})] \\ \text{s.t.} \quad & \Pr[a^r = 1 | y^{AI} \neq y^H] = \gamma \end{aligned}$$

Hence, the mis-reliant rational benchmark R^{m} represents the best score an decision-maker with a given reliance level γ could attain had they perfectly perceived the probability that the AI is correct relative to the probability that the human is correct on every decision task. By constraining a rational decision-maker to the same reliance level γ as each corresponding behavioral decision-maker, we can get a rational decision-maker who simulates the reliance level in the decision rule of the behavioral decision-maker but optimally perceives the signal and arrives at the Bayesian posterior beliefs on each instance. By comparing the expected score of these rational decision-makers and behavioral decision-makers, we can distinguish between the following sources of loss:

• **Reliance loss**, the loss from over- or under-relying on the AI, defined as $(\mathrm{R} - \mathrm{R}^{\mathrm{m}})/\Delta$. We measure reliance loss in payoff space rather than assessing the deviation from the optimal reliance level. The latter treats all errors identically, whereas using payoff space accounts for how big an error is in terms of lost payoff.

• **Discrimination loss**, the loss from not accurately differentiating the instances where the AI is better than the human from the ones where the human is better than the AI, defined as $(\mathrm{R}^{\mathrm{m}} - \mathrm{B})/\Delta$. Since R^{m} and B have the same reliance level and accept the same percentage of AI recommendations, the difference in the decisions of R^{m} and the decisions of B lies entirely in accepting the AI recommendations at different instances. R^{m} always accepts the top x% AI recommendations ranked by performance advantage over human recommendations, but B may not.

In other words, we decompose the difference between the best attainable performance in the study (R) and the observed behavior of study participants (B) into two parts. We show an example of the quantities, R, R^{m}, B, and $\mathrm{R}_{\varnothing}$, from our framework in Figure 1. Figure 1 illustrates how the behavioral performance B and mis-reliant rational benchmark R^{m} are bounded. B must be equal to or lower than the rational benchmark R. If B is higher than the rational baseline $\mathrm{R}_{\varnothing}$ (i.e., the better performance of either AI recommendations or human recommendations alone), we say B fulfills the requirement of complementary performance. R^{m} must fall between B and R.

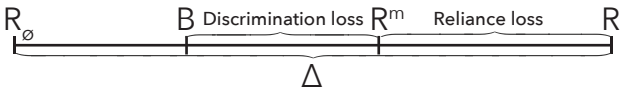

Figure 1: An example of the composition of the quantities defined in our framework. $\mathrm{R}_{\varnothing}$ and R can be calculated using knowledge of the experiment design, which in our framework includes the human recommendations and the AI recommendations in addition to the components of the decision problem (Table 1). R^{m} and B can be calculated given observed data on the human decision-maker's decisions in an AI-assisted scenario.

4 APPLYING THE FRAMEWORK TO AI RELIANCE STUDIES

We discuss how to apply the framework to AI reliance studies using an example.

Experiment design and data collection. The first step in applying the framework is to formulate the experiment design as a decision problem by defining the ground truth state, data-generating model, action space, signal, and scoring rule. Imagine we run an experiment studying AI-advised recidivism decisions with 200 humans, where each completes 20 trials. In each trial they view a profile of the defendent, and must predict whether the defendent will be re-arrested. The participants are assisted with an AI model that is deterministic and calibrated on the ground truth. We equally divide the 200 participants into two groups, randomly assigning 100 to one explanation condition and the other 100 to a different explanation condition. All participants first do the 20 instances by themselves before they see any AI recommendations, then make final decisions on the same 20 instances with the AI assistance. For every correct decision on the second batch of trials, the participant receives \$0.5 as incentivization. The decision tasks are formalized in Table 2 in Appendix B. When the experiment is complete, we have collected 4000 decision observations in total. Each observation includes information about the profile of the defendent, the outcome of whether the defendent is re-arrested, the human recommendation on the first batch of trials, the AI recommendation, the explanation of the AI recommendation, and the final decision on the second batch of trials.

Rational baseline $R_{\varnothing}$. Recall that the rational baseline represents the expected performance of the rational decision-maker without access to the signal on the derived binary-adoption decision task from the experiment. Hence, the best action is the better of always following the AI and always following the human recommendation. We estimate the rational baseline by identifying the best-response to the empirical distribution of states in the 4000 observations experiment. This calculation is illustrated in Algorithm 1 in Appendix A.

(Approximating) Rational benchmark R. To calculate the rational benchmark we identify the best response to each signal. When the signal space has finite size, we can calculate the rational benchmark by simulating the best response to each signal on the empirical

distribution of the experiment observations. However, for a large number of decision tasks in the literature (including, e.g., the demonstrations in Section 5), the signal space has near infinite size (e.g., it involves text documents) such that each experimental observation might involve a different unique signal. Thus, the identified best response action may overfit to the data relative to the true expected score of the rational decision-maker on a randomly chosen decision task from the experiment. We approximate the rational benchmark by designing an upperbound and a lowerbound.

• Upperbound: Overfitting to the empirical distribution. We calculate the rational benchmark on the empirical joint distribution $\tilde{\pi}(\hat{\theta}, v)$ over the payoff-relevant state $\hat{\theta}$ and the signal v, treating the empirical distribution as the true data generating model. Algorithm 2 in Appendix A calculates this empirical distribution.

To see why this is an upperbound and why we call it overfitting, consider the case where the signal space is continuous. Each entry in the experiment data has a distinct signal. Without repetition, it is impossible to approximate the true distribution of the payoff-relevant state $\hat{\theta}$ conditioning on each signal v. Treating the empirical distribution as the true data generating model, there is no randomness in the payoff-relevant state given the rational decision-maker's knowledge.

• Lowerbound: Learning the best response on the optimally discretized empirical distribution to avoid overfitting. Assuming continuity on the joint distribution $\tilde{\pi}(\hat{\theta}, v)$ over the payoff-relevant state $\hat{\theta}$ and the signal v, we approximate the rational benchmark by coarsening the signal space into finite discrete signals $\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k$, and calculating the best response on the empirical distribution over the discretized space $\{\tilde{v}_i\}_i$. An example using the k-means algorithm to discretize the signals is shown in Algorithm 3 in Appendix A.

To see why this is an lowerbound on the rational benchmark, first note that the rational decision-maker with the true data generating model can always perform the same discretization as the algorithm on the signal space, and such discretization to the signal can only decrease the expected performance. It remains to make sure the discretization is not too fine, such that the estimate on the empirical distribution is close to the rational decision-maker's expected payoff on the discretized signal (i.e. the estimate does not overfit to the data points from the experiment). We ensure this by performing cross-validation on the estimated average payoff. We randomly split the experiment data into a training set and a test set. Intuitively, increasing the number of clusters k leads to an expected payoff closer to the rational benchmark, but a higher gap between the estimated payoff on the clustering set and the test set (a.k.a. the generalization error). We select k to balance the increase in expected payoff and the generalization error.

The calculation of the rational benchmark hence takes an empirical distribution as input. For a finite signal space, the rational benchmark is calculated on the empirical distribution. For an infinite signal space, the upperbound is calculated on the empirical distribution, while the lowerbound is calculated on the discretized empirical distribution. Regardless of which bound we are calculating, given an empirical distribution (e.g, the 4000 observations), we simulate the rational decision-maker's decision. For each observation, the rational decision-maker receives a signal (raw signal or discretized signal) and calculates the posterior distribution of states given the signal by Bayes rule, denoted as $\pi(\hat{\theta}|v) = \frac{\pi(\hat{\theta},v)}{\pi(v)}$. We pick the action with higher expected payoff under the posterior distribution on the current observation. We repeat this process for all observations and then take the expectation on all the rational benchmarks we get. We can take the conditional expectation across different conditions, e.g., different explanations. This calculation is illustrated in Algorithm 4 in Appendix A.

Behavioral performance B. The expected performance of a behavioral decision-maker's final decision is estimated on the joint behavior of the behavioral decision-makers in the experiment, denoted as $\pi(v, \theta, a^b)$. We can use the observations to directly represent the joint behavior of the behavioral decision-makers or estimate using a model trained on the observations to predict the behavioral decisions[1]. This calculation is illustrated in Algorithm 5 in Appendix A.

(Approximating) Mis-reliant rational benchmark R^m. The mis-reliant rational benchmark is the expected score of a rational decision-maker with the same behavioral reliance level as the human participant. To calculate this, we simulate the rational decision-maker completing the same set of trials as the behavioral decision-makers do but additionally constrain the reliance level to be the same as the reliance level produced by the behavioral decision-makers. In our example experiment, each behavioral decision-maker completes 20 trials with reliance levels, $\gamma^b = \Pr[a^b = y^{AI} | y^{AI} \neq y^H]$. As the rational decision-maker traverses the 4000 observations, like behavioral participants she should engage in 20 consecutive trials for each set. Suppose that the signals that the rational decision-maker receives in the 20 consecutive trials are $v_1, \ldots, v_{20}$. For each signal v_i, the rational decision-maker knows the posterior payoffs, i.e., $\mathbf{E}_{\pi(\hat{\theta}|v_i)}[S(y^{AI}, y)]$ and $\mathbf{E}_{\pi(\hat{\theta}|v_i)}[S(y^H, y)]$. Then, the rational decision-maker ranks the signals in decreasing order of $\mathbf{E}_{\pi(\hat{\theta}|v_i)}[S(y^{AI}, y)] - \mathbf{E}_{\pi(\hat{\theta}|v_i)}[S(y^H, y)]$ and accepts the AI recommendation from the first signal in the sorted list, up to a γ^b fraction of 20 signals. We take the expectation over all observations (or conditionally on the manipulated variable of interest depending on the study design). This calculation is illustrated in Algorithm 6 in Appendix A. Note that estimation of the mis-reliant rational benchmark faces the same risk of overfitting as the rational benchmark. When the signal space is infinite, we approximate the mis-reliant rational benchmark the same way that we do the rational benchmark by calculating the upper- and lower-bound.

Quantifying uncertainty. All the quantities calculated by the above algorithms are point estimates of the expectations. To get a robust estimate, we bootstrap to compute the expectation. For each iteration in bootstrapping, we sample from the 4000 observations, and run the four algorithms on the ratio of the sample. The estimations of the expected payoff generated through iterations quantify the uncertainty. This calculation is illustrated in Algorithm 7 in Appendix A.

[1]When we estimate the joint behavior by a model, how good the estimates of behavioral performance are will depend on how well the model predicts the behavioral data.

5 DEMONSTRATION

We apply our framework to three AI-advised decision making experiments [3, 12, 21]. [2] We reanalyze the reliance levels of behavioral decision-makers within the payoff space by comparing to the rational baseline and the rational benchmark. We also identify the discrimination loss.[3]

5.1 On Human Prediction with Explanations and Predictions of Machine Learning Models [21]

Lai and Tan [21] compare different approaches to integrate an AI in the task of detecting deception in hotel reviews.

5.1.1 Experiment design. Following [24], participants are asked to look up a hotel review and then make a decision on whether the review is genuine or deceptive. Lai and Tan [21] proposed seven conditions with different levels of AI assistance along a hypothesized spectrum from full human agency to full automation: no information from the AI, only example-based explanation, only highlight-feature explanation, only heatmap explanation, only predicted label, predicted label with random heatmap explanation, predicted label with example-based explanation, predicted label with heatmap explanation, and predicted label with accuracy. Since the reliance problem we study is defined only for the scenario where the AI recommendation is provided to the human decision maker, we analyze only the five conditions including AI information. The decision task is summarised in Table 4 in Appendix B.

5.1.2 Analysis. The conclusions drawn by Lai and Tan [21] include: AI-advised decisions were better when the AI system interfered more with the human decision-maker's process, and trust in the AI recommendation increased with more AI-based information. Trust was evaluated by the rate at which the AI recommendations were accepted. Their results ranking the AI-based conditions by both performance and trust is (from worst to best) were: no predicted label < only predicted label < predicted label with random heatmap explanation < predicted label with example-based explanation < predicted label with heatmap explanation < predicted label with accuracy. Using our approach, we examine the ranking of behavioral performance within the scale created by the rational baseline and rational benchmark. Instead of evaluating reliance as rate of acceptance of AI recommendations, we evaluate the reliance level of the behavioral decision-makers in payoff space.

Extending the author's original conclusions, we find that **the rational baseline dominates almost all other quantities in our framework except the rational benchmark**, including the behavioral performance and the mis-reliant rational benchmark across all explanation conditions, as shown in Figure 2 (**the rational baseline** and **the rational benchmark**). Additionally, **the rational benchmark only improves marginally over the rational baseline**, i.e., the rational decision-maker does not gain much from access to human recommendations, as shown in Figure 2A (**the rational benchmark** and **the rational baseline**). Consequently, it is hard to expect behavioral decision-makers to achieve complementary performance. These findings suggest that the experimental design was poorly suited for studying complementary performance, because the AI consistently outperforms the human.

Using our approach, we extend the authors' results by observing that **different explanation conditions result in different levels of discrimination loss and reliance loss.** For example, the condition with heatmap explanations and the condition directly providing model accuracy show similar reliance loss (Figure 2C) but the discrimination loss in the latter is smaller than the former. This suggests why showing accuracy can help the behavioral decision-makers achieve higher performance than heatmap explanations: the accuracy information helps the behavioral decision-makers better differentiate instances where the AI predictor outperforms the human predictor from those where the human predictor outperforms the AI predictor, presumably because it provides information on the joint distribution of the AI recommendation and the ground truth that is absent from the heatmap explanations.

5.2 Does the Whole Exceed its Parts? [3]

Bansal et al. [3] use an online crowdsourced experiment to investigate the effects of explanations on the degree of complementary performance achieved by AI-advised humans. In contrast to prior studies like [21], Bansal et al. [3] controlled the AI's accuracy to be comparable to the humans', to avoid the AI being obviously better than human performance on the task.

5.2.1 Experiment design. The experiment compares human-AI team decisions across four approaches to explaining AI recommendations: no explanation, explanation for the most confident AI recommendation, explanations for the top-2 most confident AI recommendations, and adaptively showing explanations for the top-1 or top-2 most confident AI recommendations, randomly assigned between subjects. The participants are tasked with using the AI recommendation and its explanation for two tasks: sentiment classification and LSAT (multiple-choice questions where one of four choices is the correct answer). Because the manipulation of interest (explanation types) and conclusions drawn about the complementary performance of the human-AI teams across different explanation types are the same between the two tasks, we analyze only the results of the LSAT task. The decision task is summarised in Table 3 in Appendix B.

5.2.2 Analysis. Bansal et al. [3] drew several conclusions from their results: AI-advised decision making achieved complementary performance (i.e., a higher payoff than expected of the human or AI alone), and presenting explanations to the human-AI team led to no observable performance improvements using null hypothesis significance testing (NHST) with $\alpha = 0.05$. The authors speculated that the reason they did not observe improvement from explanations is because people over-relied on the AI when explanations are provided. This is supported by evidence that providing explanations increased decision performance when the AI was correct

[2] We use the upper bound (overfit) method to approximate the rational benchmarks and the mis-reliant rational benchmark, i.e., estimating the empirical distribution using the observations of signals and payoff-relevant state and treating the empirical distribution as the true data generating model. We confirmed our conclusions from this approach using the approximation of the rational benchmark with discretized signals in Appendix C.

[3] See our supplementary materials for complete analysis with full code and original data: https://osf.io/2cbxf/?view_only=fd9c2e8e1dd24aa787af05dadafe4bcf

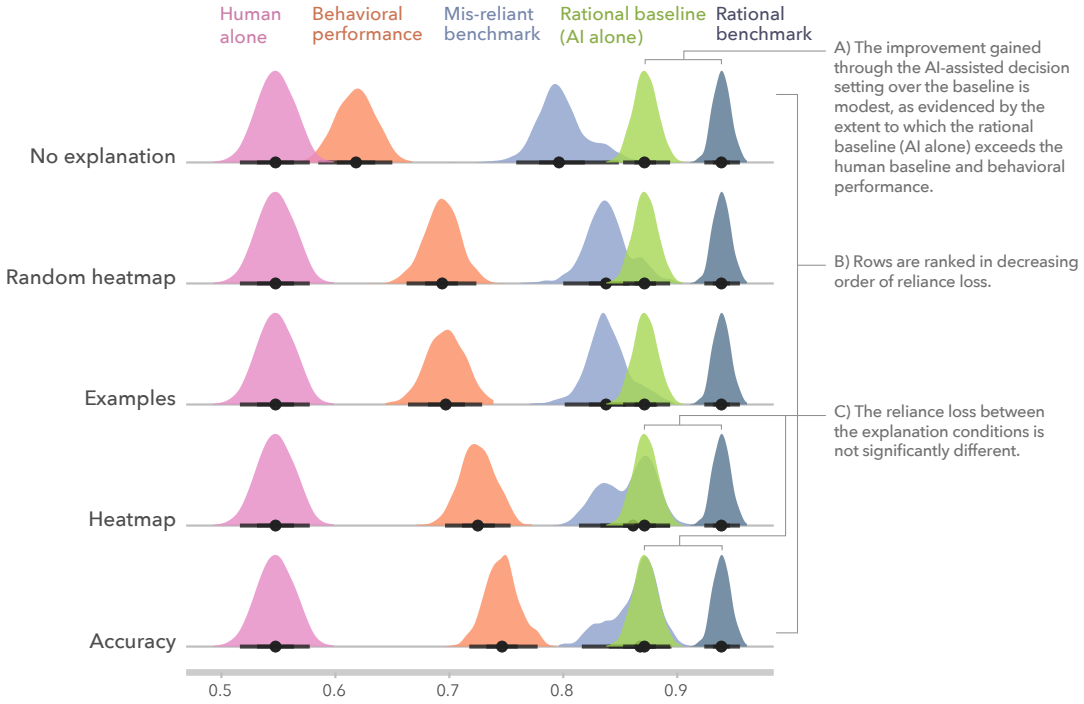

Figure 2: Expected payoffs of benchmarks, baselines, and observed performance in Lai and Tan [21].

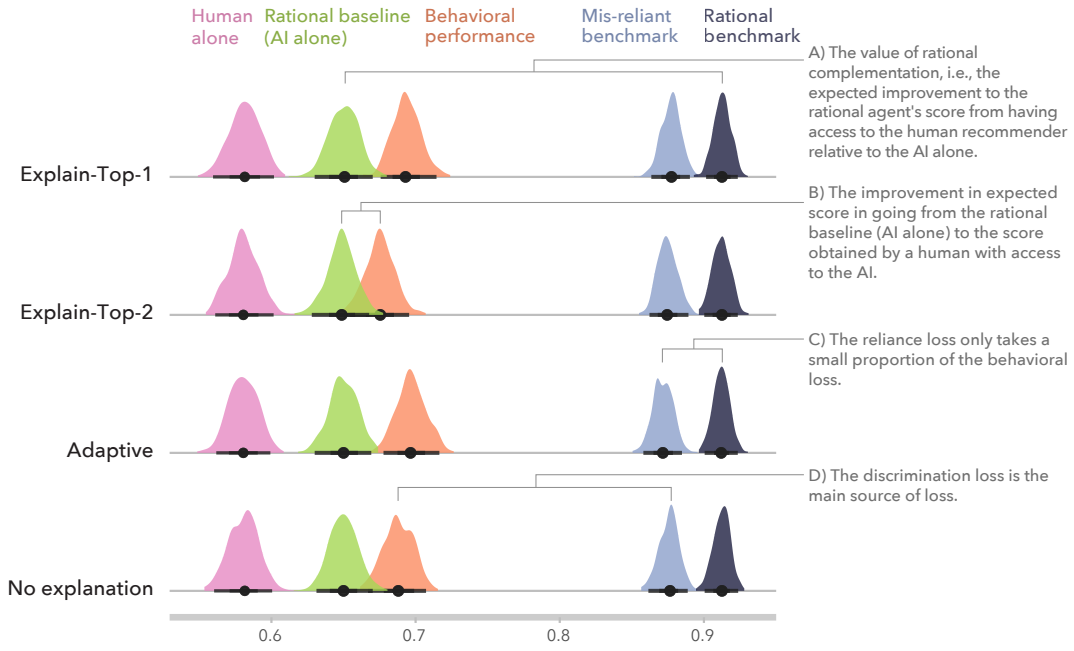

Figure 3: Expected payoffs of benchmarks, baselines, and observed performance in Bansal et al. [3].

and decrease it when the AI was incorrect. We use our framework to evaluate this conclusion. Specifically, we compare the observed behavioral payoffs to the rational baseline and rational benchmark, and evaluate the reliance level of participants in payoff space by comparing the behavioral payoffs to the mis-reliant rational benchmark. Our results are shown in Figure 3.

Extending the authors' original conclusions, we find that **despite the behavioral decision-makers achieving complementary performance, there is still considerable room for improvement**, shown as the distance between **the behavioral performance** and **the rational benchmark** (Figure 3A and B). The **behavioral payoff** surpasses the **rational baseline**, as shown in all rows representing different explanation conditions in Figure 3. This comparison leads to the authors' conclusion that complementary performance is observed in every condition. However, comparing to the **rational benchmark**, the **behavioral decision-makers** only improve a small proportion over the **rational baseline** (Figure 3). Our analysis more clearly demonstrates the remaining need to identify ways to bridge the remaining substantial gap.

Applying NHST as in the original study, we corroborate the authors' conclusion that there are **no significant improvements for explanation conditions over the no explanation condition**. Using our approach we confirm there are not significant reductions in either discrimination loss or reliance loss. For example, in Figure 3 (**behavioral performance** and **mis-reliant rational benchmark**), the behavioral decision-makers in the no explanation and the adaptive explanation condition achieve similar performance; the same is true of the Explain-Top-1 and Explain-Top-2 conditions.

Further extending the original conclusions, we find that **despite the over-reliance shown by the original paper, poor reliance itself is not the main source of loss.** While the behavioral decision-makers' reliance levels across all conditions *are* higher than the optimal reliance level in expectation represented by the rational benchmark, our analysis suggests that miscalibrated reliance of the behavioral decision-makers does not lead to substantial loss in payoff. As shown in Figure 3C, **the mis-reliant rational benchmarks** across all conditions are very close to **the rational benchmark**, such that reliance loss is very minor compared to the total behavioral losses.

Instead our approach shows that **the behavioral decision-makers have substantially lower performance compared to the rational benchmark due to large discrimination loss** (i.e., accepting the AI recommendations for the wrong instances), as shown in Figure 3D. Combined with the evidence that the behavioral decision-makers have low reliance loss, this could suggest that the explanations be designed specifically to help users distinguish the intance where the AI is expected to succeed from those where the AI is expected to fail, instead of aiming to calibrate the human's overall trust in the AI's accuracy or adjusting the human's decision rule. For example, explanations could give information on the joint distribution of AI recommendation and the ground truth, i.e., $\pi(y^{AI}, y)$ rather than focusing on describing only the decision rule of AI, e.g., as in LIME [25] or SHAP [23].

5.3 The Impact of Algorithmic Risk Assessments on Human Predictions and its Analysis via Crowdsourcing Studies [12]

Fogliato et al. [12] conduct an online crowdsourcing experiment where participants face the task of assessing a defendant's risk of re-arrest after viewing the defendant's profile. The experiment investigates the research questions of whether anchoring effects impact participants' recommendations and whether the evaluation of participants' decisions depends on the types of recommendations (probablity or binary decision), both of which can be modeled as decision tasks in our framework.

5.3.1 Experiment Design. The experiment compares AI-assisted human recommendations under two different conditions: anchoring and non-anchoring. Participants assigned to the anchoring condition see the question presented together with the AI's recommendation, while under the non-anchoring condition, participants are asked to predict the risk before seeing AI recommendation and then to revise their assessment after having AI recommendation. In each question, participants are shown the profile of a defendant, including demographics, current charge, and criminal history. Participants are asked to report: 1) the probability of the defendant being re-arrested from $[0, 100\%]$, and 2) a binary choice of whether the defendant will be re-arrested within a given duration or not. The decision tasks for probabilty and binary decision are summarised in Table 5 in Appendix B.

5.3.2 Analysis. Fogliato et al. [12] report that 1) the probability of re-arrest reported by the participants did not uniformly map to their binary decision, such that behavioral predictive performance and reliance level must be considered separately, and 2) no clear differences between participants' accuracy, false positive rate, false negative rate, positive predicted values, or AUC were found between the anchoring and no anchoring condition. Our analysis of their results is shown in Figure 4 for the binary decision task and the probabilistic decision task.

Corroborating with the authors' conclusion, by putting both tasks on the same payoff scale, we find that **people are better at the probability task than the decision task.** First, we observe that the behavioral decision-makers doing the probability task can achieve higher performance than those doing the binary decision task overall. For example, **the behavioral performance** for the probability task is much higher than **the behavioral performance** for the binary decision task (Figure 4). Second, **the behavioral performance** is higher than **the performance of the human only baseline** in the probabilistic task while they perform similarly in the decision task, as shown in Figure 4B. These results corroborate the conclusion by Fogliato et al. [12] that there is no deterministic decision rule that describes how the participants' probability estimates map to their binary decisions.

We also find that **the rational baselines and the rational benchmarks differ for each task between the anchoring and the no anchoring conditions, suggesting a need to reconsider Fogliato et al. [12]'s conclusion about the similarity between anchoring and no anchoring**. As shown in Figure 4A, **the rational baseline** in the anchoring condition is slightly higher than in

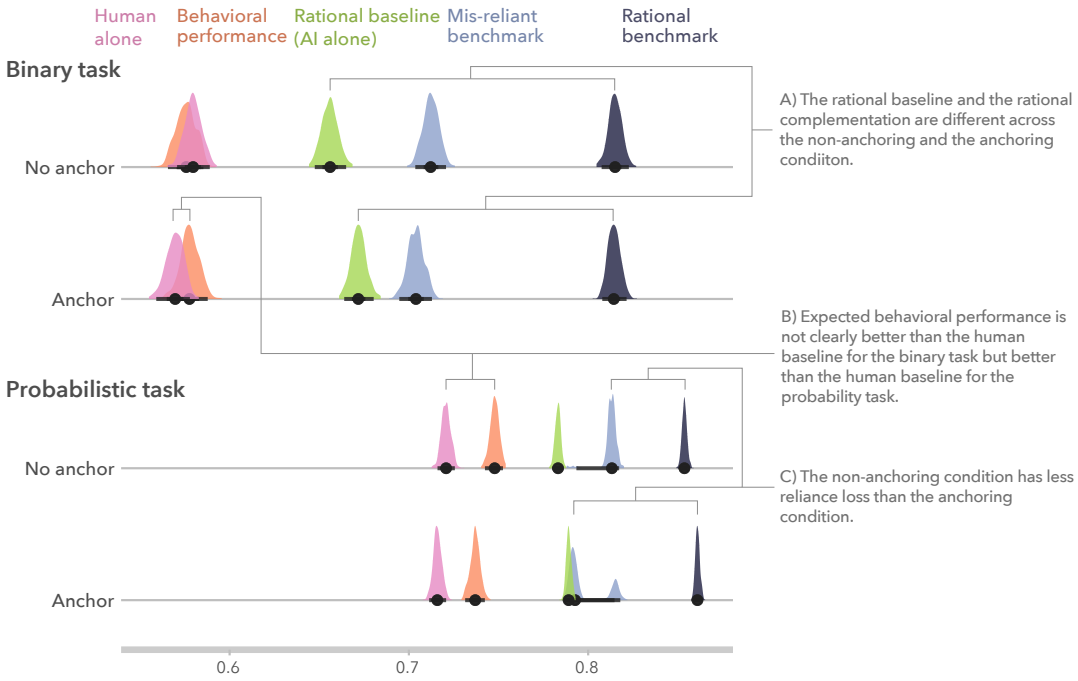

Figure 4: Expected payoffs of benchmarks, baselines, and observed performance in Fogliato et al. [12]

the non-achoring condition. This implies just comparing the absolute performance of the behavioral decision-makers can mislead. Despite the behavioral performance being similar across the conditions in terms of absolute values, the behavioral decision-makers have better relative performance in the non-anchoring condition than the anchoring condition when compared to the rational baseline and the rational benchmark.

Similarly, contradicting the authors' conclusion, we find that **the behavioral decision-makers' reliance is closer to the appropriate reliance under the non-anchoring condition than the anchoring condition in both tasks.** As shown in Figure 4C, the reliance loss ($\frac{R-R^m}{R-R_\varnothing}$) is lower for the no anchoring condition, while the discrimination loss ($\frac{R^m-B}{R-R_\varnothing}$) is slightly higher. This suggests that letting the behavioral decision-makers make a decision by themselves first (a.k.a., the non-anchoring effects) can improve their reliance, but not necessarily help them distinguish between the signals where the AI recommendation is expected to outperform the human recommendation and the signals where the human recommendation is expected to outperform the AI recommendation.

6 DISCUSSION

We contribute a formal definition of reliance and corresponding framework for interpreting losses in behavioral decision-making performance within the baseline and benchmark for complementary performance. The first source of loss concerns the difference in the rate at which the behavioral decision-maker relies on the AI relative to the appropriate level of reliance defined by the decision problem, calculated in payoff space. The second source of loss concerns the difference in score between a behavioral decision-maker and the best score a rational decision-maker who relies on the AI at the same rate as the behavioral decision-maker but who perfectly perceives the posterior probabilities could achieve. By contributing clear comparison points in the form of performance *benchmarks* to the design and interpretation of studies of human reliance on AI, our work enables researchers to identify the upper-bound of complementary performance and how far the human-AI team is from this optimal attainable performance.

Closest to the motivation of our work, Fok and Weld [13] motivate the need for a notion of "strategy-graded reliance," where appropriate reliance is determined from the relative expected performance of the human and the AI, over "outcome-graded reliance" based on the human's acceptance of AI advice conditioned on its post-hoc correctness. Several other prior works propose studying reliance using conditional probability (e.g., [26, 27, 30, 32]) to separate cases where the human recommendation is better than the AI recommendation from cases where the AI recommendation is better than the human recommendation. We unambigously define strategy-guided reliance and show how to calculate optimal reliance and disentangle sources of behavioral loss.

Our framework enables evaluating reliance in payoff space, in contrast to prior work which has evaluated reliance in action space only [3, 26, 32]. Studying reliance only in the action space still neglects sensitivity in the payoff, such as the magnitude of improvement that the human recommendation provides over the AI recommendation or vice versa. Defining a measurement of reliance in payoff space also enables the calculation of a benchmark to compare with, which we show in our demonstrations to be highly valuable for learning from a reliance evaluation.

Decoupling sources of behavioral loss in human AI-advised decisions is important for designing and interpreting AI-advised decision-making experiments, which helps to build better understanding and test hypothesis about the source of behavioral loss. In recent years, numerous papers [1, 3, 3–12, 15–17, 19, 22, 30, 33–35] have employed user studies to investigate how various factors contribute to enhancing the complementary performance of human-AI teams. Without a well-grounded notion of reliance, such studies have limited ability to draw conclusions from a decision-making task on how good the reliance is and whether action should be taken to improve it. For example, in our demonstration of Bansal et al. [3], we find that the reliance level differing from optimal is not the main source of behavioral loss. This intepretation would suggest follow-up actions like calibrating human's trust on the AI in general (e.g., by making sure they have internalized information about its accuracy), but this may not adequately address challenges they face in discriminating which signals warrant accepting the AI's prediction. We also admit that while distinguishing reliance from discrimination loss in human-AI team performance may be useful to drive further improvements when there is a large discrepancy between these, in practice actions taken to improve one form of loss will likely affect the other.

Importantly, our framework hypothesizes two distinct roles in the decision-making process to separate human recommendations without AI assistance from the the process by which the human makes the final decision with access to human recommendations and AI recommendations. This setup allows researchers to better interpret experiments and design the decision process they study; however, the generalizability of our framework to alternative study set-ups still holds. Our framework can be applied to situations where the human is both making a recommendation and making the final decision, i.e., where the human recommender and decision-maker are the same person. However, without constraints, they might ignore the AI and just submit the human recommendation or anchor on the AI without thinking to make the decision by themselves. Both of these two cases cause inaccurate measurement of reliance, since AI recommendations and human recommendations are not consulted in human's decision rule. Efforts should be made to align with the assumptions of our framework to facilitate the interpretation of experimental results.

6.1 Limitations

We formalize the AI-advised decision-making problem into a binary choice of whether to adopt a human recommendation or an AI recommendation. However, this may not be suitable for every real world case. For example, when the recommendation space is continuous (e.g., regression), the human decision-maker is likely to make a decision that is different from the human recommendation or the AI recommendation. Future work could extend our definition to continuous recommendation spaces.

We only identify two losses affecting human decision-makers, though more fine-grained losses may exist in AI-advised decision-making and be worth analyzing. For example, discrimination loss can be caused by two possible reasons: misidentifying the probability that the AI is correct or misidentifying the probability that the human is correct. Improving the former implies better conveying the AI's accuracy, while improving the latter implies giving information on the human's average performance on the task. More fine-grained behavioral losses can increase learning from experimental results and imply more targeted improvement of designs. Future work can seek to identify and separate such additional behavioral losses and explore possible design choices to address them.

ACKNOWLEDGMENTS

We would like to thank the authors of Lai and Tan [21], Bansal et al. [3], and Fogliato et al. [12], who provided their data for demonstration in this paper.

REFERENCES

[1] Maryam Ashoori and Justin D. Weisz. 2019. In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes. arXiv:1912.02675 [cs.CY]

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. 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* (, Yokohama, Japan,) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[4] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 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 (2021), 1–21.

[6] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*. 160–169. https://doi.org/10.1109/ICHI.2015.26

[7] Valerie Chen, Nari Johnson, Nicholay Topin, Gregory Plumb, and Ameet Talwalkar. 2022. Use-case-grounded simulations for explanation evaluation. *Advances in Neural Information Processing Systems* 35 (2022), 1764–1775.

[8] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–32.

[9] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[10] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[11] Shi Feng and Jordan Boyd-Graber. 2019. What can AI do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 229–239. https://doi.org/10.1145/3301275.3302265

[12] Riccardo Fogliato, Alexandra Chouldechova, and Zachary Lipton. 2021. The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–24.

[13] Raymond Fok and Daniel S Weld. 2023. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. *arXiv preprint arXiv:2305.07722* (2023).

[14] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[15] Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831* (2020).

[16] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[17] Benjamin D Horne, Dorit Nevo, John O'Donovan, Jin-Hee Cho, and Sibel Adalı. 2019. Rating reliability and bias in news articles: Does AI assistance help everyone?. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 247–256.

[18] Maia Jacobs, Melanie F Pradier, Thomas H McCoy Jr, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry* 11, 1 (2021), 108.

[19] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. 2018. To trust or not to trust a classifier. *Advances in neural information processing systems* 31 (2018).

[20] Igor Kononenko. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine* 23, 1 (2001), 89–109.

[21] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[22] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.

[23] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[24] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557* (2011).

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[26] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[27] Jakob Schoeffer, Johannes Jakubik, Michael Voessing, Niklas Kuehl, and Gerhard Satzger. 2023. On the Interdependence of Reliance Behavior and Accuracy in AI-Assisted Decision-Making. *arXiv preprint arXiv:2304.08804* (2023).

[28] Michelle Vaccaro and Jim Waldo. 2019. The effects of mixing machine learning and human judgment. *Commun. ACM* 62, 11 (2019), 104–110.

[29] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[30] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th international conference on intelligent user interfaces*. 318–328.

[31] Yifan Wu, Ziyang Guo, Michails Mamakos, Jason Hartline, and Jessica Hullman. 2023. The Rational Agent Benchmark for Data Visualization. *arXiv preprint arXiv:2304.03432* (2023).

[32] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 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.

[33] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[34] Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2016. Trust and reliance based on system accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. 223–227.

[35] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

A THE ALGORITHMS FOR CALCULATIONS IN THE FRAMEWORK

This appendix includes all the algorithms in the form of pseudocode for all the calculations we introduce in Section 4.

Input: the experimental data D with each row representing one experimental trial, and the scoring rule for the derived binary-adoption decision task $\hat{S}$
Output: the rational baseline $R_{\varnothing}$
$payoff \leftarrow 0$;
for $action \leftarrow \{0, 1\}$ *(action = 0 follow human, 1 follow AI)* **do**
  for $row \in D$ **do**
    $\hat{\theta} \leftarrow$ the state realized in row;
    $payoff \leftarrow payoff + \hat{S}(action, \hat{\theta})$;
  end
  $payoff_{action} \leftarrow payoff/$the number of rows in D;
end
$R_{\varnothing} = \max\{payoff_0, payoff_1\}$;

Algorithm 1: Rational baseline

Input: the experimental data D with each row representing one experimental trial, the space of derived binary-adoption states $\hat{\Theta}$, and the space of signals V
Output: the empirical distribution $\tilde{\pi}(\hat{\theta}, v)$
$\tilde{\pi}(\hat{\theta}, v) \leftarrow 0\,\mathbb{1}_{|\hat{\Theta}|}\mathbb{1}_{|V|}^{\top}$;
/* Initializing a matrix with all 0. */
for $row_i \in D$ **do**
  $\hat{\theta}_i \leftarrow$ the state realized in row_i;
  $v_i \leftarrow$ the signal realized in row_i;
  $\tilde{\pi}(\hat{\theta}_i, v_i) \leftarrow \tilde{\pi}(\hat{\theta}_i, v_i) + 1$;
end
$\tilde{\pi}(\hat{\theta}, v) \leftarrow \tilde{\pi}(\hat{\theta}, v)/|\tilde{\pi}(\hat{\theta}, v)|$;
/* Normalizing to get the joint distribution. */

Algorithm 2: Calculating the empirical distribution

Input: the experimental data D with each row representing one experimental trial, the total number of clusters K, the space of derived binary-adoption states $\hat{\Theta}$, and the space of signals V
Output: the empirical distribution $\tilde{\pi}(\hat{\theta}, \tilde{v})$ on the optimally discretized space
$\tilde{\pi}(\hat{\theta}, \tilde{v}) \leftarrow 0\,\mathbb{1}_{|\hat{\Theta}|}\mathbb{1}_{K}^{\top}$;
/* Initializing a matrix with all 0. */
$\{v_i\} \leftarrow$ all signals realized in D;
$kmeans \leftarrow initialize\_kmeans(\{v_i\}, K)$;
/* Training the K-means model. */
for $row_i \in D$ **do**
  $\hat{\theta}_i \leftarrow$ the state realized in row_i;
  $v_i \leftarrow$ the signal realized in row_i;
  $\tilde{v}_i \leftarrow kmeans(v_i)$;
  $\tilde{\pi}(\hat{\theta}_i, \tilde{v}_i) \leftarrow \tilde{\pi}(\hat{\theta}_i, \tilde{v}_i) + 1$;
end
$\tilde{\pi}(\hat{\theta}, \tilde{v}_i) \leftarrow \tilde{\pi}(\hat{\theta}, \tilde{v}_i)/|\tilde{\pi}(\hat{\theta}, \tilde{v}_i)|$;
/* Normalizing to get the joint distribution. */

Algorithm 3: Discretizing signals using the cluster generated by K-means

Input: the experimental data D with each row representing one experimental trial, the joint distribution between states and signals $\pi(\hat{\theta}, v)$, and the scoring rule for the derived binary-adoption decision task $\hat{S}$
Output: the rational benchmark R
$payoff \leftarrow 0$;
for $row \in D$ **do**
  $v \leftarrow$ the signal realized in row;
  $\pi(\hat{\theta}|v) = \pi(\hat{\theta}, v)/\pi(v)$;
  /* the posterior distribution of the binary-adoption state */
  $action \leftarrow \text{argmax}_{\hat{a} \sim \{\text{human,AI}\}} E_{\hat{\theta} \sim \pi(\hat{\theta}|v)}(\hat{S}(\hat{a}, \theta))$;
  /* the action made on the posterior distribution */
  $\hat{\theta} \leftarrow$ the state realized in row;
  $payoff \leftarrow payoff + \hat{S}(action, \hat{\theta})$;
end
$R \leftarrow payoff/$the number of row in D;

Algorithm 4: Rational benchmark

Input: the experimental data D with each row representing one experimental trial, the joint behavior $\pi(v, \theta, a^b)$, and the scoring rule S
Output: the behavioral performance B
$payoff \leftarrow 0$;
for $row \in D$ **do**
  $v \leftarrow$ the signal realized in row;
  $\theta \leftarrow$ the state realized in row;
  $action \leftarrow$ action drawn from $\pi(a^b|\theta, v) = \pi(v, \theta, a^b)/\pi(\theta, v)$;
  $payoff \leftarrow payoff + S(action, \theta))$;
end
$\mathrm{B} \leftarrow payoff/$the number of row in D;

Algorithm 5: Behavioral performance

Input: the experimental data D with each row representing one experimental trial, the joint distribution between states and signals $\pi(\hat{\theta}, v)$, the scoring rule for the original decision task S, and the scoring rule for the derived binary-adoption decision task $\hat{S}$
Output: the mis-reliant rational benchmark R^{m}
$P \leftarrow \{P_1, \ldots, P_M\}$;
/* The sets of trials finished by each participant; M participants in total. */
for $i \in \{1, \ldots, M\}$ **do**
  $P_i \leftarrow filter(D, participant\_id == i)$;
end
$payoff \leftarrow 0$;
for $P_i \in P$ **do**
  Sort P_i in decreasing order of $\mathbf{E}_{\pi(\hat{\theta}|v)}[S(y^{AI}, y)] - \mathbf{E}_{\pi(\hat{\theta}|v)}[S(y^{H}, y)]$;
  $\{v_j\} \leftarrow \{$the signal realized in $row_j\}_{row_j \in P_i}$;
  $\{\theta_j\} \leftarrow \{$the state realized in $row_j\}_{row_j \in P_i}$;
  $\{a^b_j\} \leftarrow \{$action drawn from $\pi(a^b|\theta_j, v_j)\}_{row_j \in P_i}$;
  $\gamma^b \leftarrow \sum_{row_j \in P_i} \mathbb{1}[a^b_j = y^{AI}_j \& y^{AI}_j \neq y^H_j]$;
  $N \leftarrow$ the number of rows in P_i;
  $\{a^r_j\} \leftarrow \{AI\}_{j \in \{1,\ldots,\gamma^b\}} \cup \{human\}_{j \in \{\gamma^b+1,\ldots,N\}}$;
  $\{\hat{\theta}_j\} \leftarrow$ $\{$the binary-adoption state realized in $row_j\}_{row_j \in P_i}$;
  $payoff \leftarrow payoff + \sum_{j \in [N]} S(a^r_j, \hat{\theta}_j))$;
end
$\mathrm{R}^{\mathrm{m}} \leftarrow payoff/$the number of rows in D;

Algorithm 6: Mis-reliant rational benchmark

Input: the experimental data D with each row representing one experimental trial, total number of iterations T, the sample size k, prior distribution of the binary-adoption state $\pi(\hat{\theta})$, the joint distribution between states and signals $\pi(\hat{\theta}, v)$, the joint behavior $\pi(v, \theta, a^b)$, the scoring rule S, and the scoring rule for derived binary-adoption decision task $\hat{S}$
Output: the distribution of the rational baseline $\{\mathrm{R}_{\varnothing i}\}_{i \in [T]}$, the rational benchmark $\{\mathrm{R}_i\}_{i \in [T]}$, the behavioral performance $\{\mathrm{B}_i\}_{i \in [T]}$, and the mis-reliant rational benchmark $\{\mathrm{R}^{\mathrm{m}}_i\}_{i \in [T]}$
for $i \in [T]$ **do**
  $\tilde{D} \leftarrow sample(D, k)$;
  $\mathrm{R}_{\varnothing i} \leftarrow$ Rational baseline$(\tilde{D}, \pi(\hat{\theta}), \hat{S})$;
  $\mathrm{R}_i \leftarrow$ Rational benchmark$(\tilde{D}, \pi(\hat{\theta}, v), \hat{S})$;
  $\mathrm{B}_i \leftarrow$ Behavioral performance$(\tilde{D}, \pi(v, \hat{\theta}, a^b), S)$;
  $\mathrm{R}^{\mathrm{m}}_i \leftarrow$ Mis-reliant rational baseline$(\tilde{D}, \pi(\hat{\theta}, v), S, \hat{S})$;
end

Algorithm 7: Quantifying uncertainty

B FORMALIZED DECISION TASKS

	The original decision task	The derived binary-adoption decision task
Payoff-related state	θ = Ground truth $y \in \{0,1\}$ Be re-arrested or not	$\hat{\theta} = (y, y^H, y^{AI})$ Ground truth $y \in \{0,1\}$ Human recommendation $y^H \in \{0,1\}$ AI recommendation $y^{AI} \in \{0,1\}$
Data generating model	A profile x of a defendent who is randomly drawn from the defendent population Ground truth y drawn from a distribution conditioned on x. The human recommendation y^H is produced by the decision rule of the human predictor, represented by the joint behavioral $\pi(y^H, x, y)$ AI recommendation y^{AI} for the profile x The explanation $e(y^{AI})$	
Action (choice)	$a \in \{0,1\}$ Be re-arrested or not	$\hat{a} \in \{0 = \text{human}, 1 = \text{AI}\}$
Signal	$v = \{x, y^H, y^{AI}, e(y^{AI})\}$	
Scoring rule (payoff)	$S(a,\theta) = 0.5 \times \mathbb{1}[a = \theta]$	$\hat{S}(\hat{a},\hat{\theta}) = S(y^H, y)$ if $\hat{a}$ = human $\hat{S}(\hat{a},\hat{\theta}) = S(y^{AI}, y)$ if $\hat{a}$ = AI

Table 2: Example of original and derived binary-adoption decision task in hypothetical recidivism experiment

Payoff-related state	θ = Correct answer $y \in \{A,B,C,D\}$	$\hat{\theta} = (y, y^H, y^{AI})$ Ground truth $y \in \{A,B,C,D\}$ Human recommendation $y^H \in \{A,B,C,D\}$ AI recommendation $y^{AI} \in \{A,B,C,D\}$
Data generating model	Question x drawn from the scope of LSAT questions Correct answer y for x AI recommendation y^{AI} for x Human recommendation y^H: $y^H \sim \pi(x, y^H)/\pi(x)$ Explanation $e(y^{AI})$	
Action (choice)	$a \in \{A,B,C,D\}$	$a \in \{0 = \text{human}, 1 = \text{AI}\}$
Signal	$v = \{x, y^H, y^{AI}, e(y^{AI})\}$	
Scoring rule (payoff)	$S(a,\theta) = \mathbb{1}[a = \theta]$	$\hat{S}(\hat{a},\hat{\theta}) = S(y^H, y)$ if $\hat{a}$ = human $\hat{S}(\hat{a},\hat{\theta}) = S(y^{AI}, y)$ if $\hat{a}$ = AI

Table 3: Bansal et al. [3] decicion task under our framework.

	The original decision task	The derived binary-adoption decision task
Payoff-related state	θ = Ground truth $y \in \{0,1\}$ Deceptive or genuine	$\hat{\theta} = (y, y^H, y^{AI})$ Ground truth $y \in \{0,1\}$ Human recommendation $y^H \in \{0,1\}$ AI recommendation $y^{AI} \in \{0,1\}$
Data generating model	Ground truth $y \sim \text{Bernoulli}(0.5)$, indicating whether the review is written by a person who has been going to the hotel or not. Review text x generated by the person $x \sim \pi(x, y)/\pi(y)$ Human recommendation y^H $y^H \sim \pi(x, y^H)/\pi(x)$ AI recommendation y^{AI} for x Explanation $e(y^{AI})$	
Action (choice)	$a \in \{0,1\}$ Deceptive or genuine	$\hat{a} \in \{0 = \text{human}, 1 = \text{AI}\}$
Signal	$v = \{x, y^H(x), y^{AI}(x), e(y^{AI})\}$	
Scoring rule (payoff)	$S(a,\theta) = \mathbb{1}[a = \theta]$	$\hat{S}(\hat{a},\hat{\theta}) = S(y^H, y)$ if $\hat{a}$ = human $\hat{S}(\hat{a},\hat{\theta}) = S(y^{AI}, y)$ if $\hat{a}$ = AI

Table 4: Lai and Tan [21] decision task under our framework.

	The binary decision task	The probabilistic decision task	The binary-adoption decision task
Payoff-related state	θ = Ground truth $y \in \{0,1\}$ Re-arrest or not		$\hat{\theta} = (y, y^H, y^{AI})$ Ground truth $y \in \{0,1\}$ Human recommendation $y^H \in \{0,1\}$ AI recommendation $y^{AI} \in \{0,1\}$
Data generating model	A defendent p is randomly drawn from the defendent population. The profile x for p Ground truth y for p Human recommendation y^H $y^H \sim \pi(x, y^H)/\pi(x)$ AI recommendation y^{AI} for x AI's confidence score $e(y^{AI})$		
Action (choice)	$a \in \{0,1\}$ Re-arrest or not	$a \in [0,1]$ Probability of re-arrest	$\hat{a} \in \{0 = \text{human}, 1 = \text{AI}\}$
Signal	$v = \{x, y^H(x), y^{AI}(x), e(y^{AI})\}$		
Scoring rule (payoff)	$S(a,\theta) = 1 - (a-\theta)^2$		$\hat{S}(\hat{a},\hat{\theta}) = S(y^H, y)$ if $\hat{a}$ = human $\hat{S}(\hat{a},\hat{\theta}) = S(y^{AI}, y)$ if $\hat{a}$ = AI

Table 5: Fogliato et al. [12] decision task under our framework.

C THE RESULTS OF DEMONSTRATIONS USING DISCRETIZED SIGNAL APPROXIMATION

This appendix includes our additional results for demonstrations in Section 5, where we use the discretized signals to approximate the rational benchmark and the mis-reliant rational benchmark. We subsequently re-check the conclusions we get in Section 5 with the results shown in this appendix. All the conclusions analyzed under the results of approximation using discretized signals corroborate with the conclusions we get in Section 5.

C.1 Does the Whole Exceed its Parts? [3]

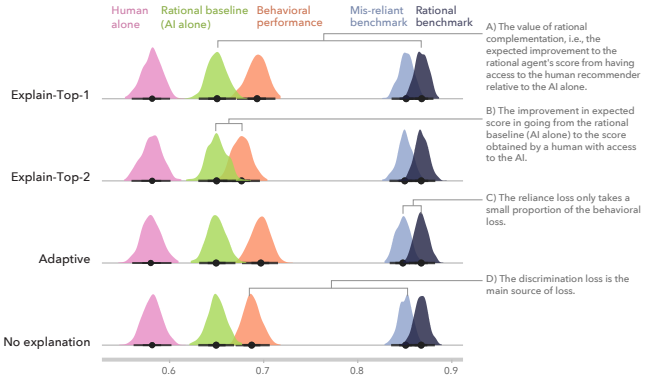

Figure 5: Estimated payoffs of the experiment data in Bansal et al. [3].

First, the results also show considerable room for improvement to achieve to **the rational benchmark**, as shown in Figure 5A and B. Second, no significant improvement by displaying explanations is evidenced in the results. As shown by Figure 5, the **behavioral performance** and the**mis-reliant rational benchmark** perform similarly across the explanation conditions and the no explantion condition. Third, the reliance loss is modest to the behavioral loss, while the discrimination loss is the main source of loss, as shown in Figure 5C and D.

C.2 On Human Predictions with Explanations and Predictions of Machine Learning Models [21]

First, similarly to what we get in Section 5, **the rational baseline** dominates all other quantities defined by our framework except **the rational benchmark**, leading to the conclusion about the failure of complementary performance in the decision task. Second, **the rational benchmark** only shows marginal improvement over **the rational baseline**, as shown in Figure 5A. Third, the explanations can improve the behavioral performance and the reliance, as shown in Figure 5C. Finally, we observed the same pattern of reliance loss and discrimination loss in the results, e.g., Figure 5D.

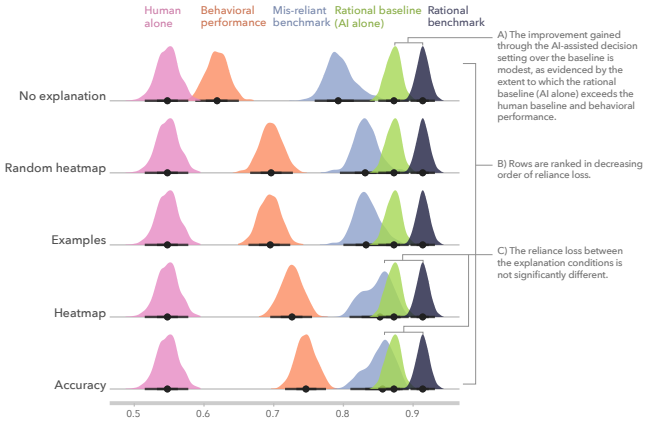

Figure 6: Estimated payoffs of the experiment data in Lai and Tan [21].

C.3 The Impact of Algorithmic Risk Assessments on Human Predictions and its Analysis via Crowdsourcing Studies [12]

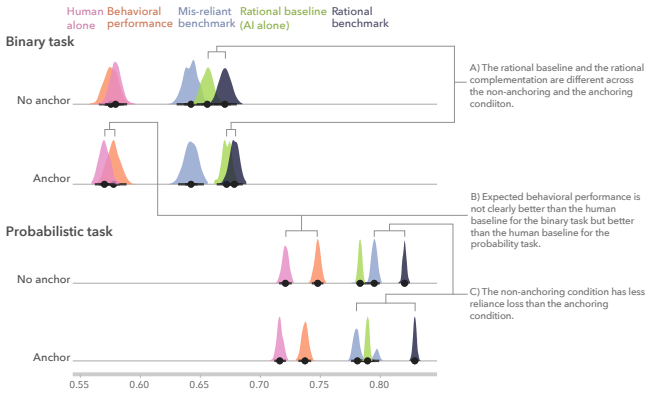

Figure 7: Estimated payoffs of the experiment data in Fogliato et al. [12].

First, we also find the quantities under our framework act differently between the probabilistic decision task the the binary decision task. For example, **the behavioral performance** exceeds **the performance of human predictions** in the probabilistic decision task while acts the same in the binary decision task (Figure 5B). Second, **the rational baseline** and **the rational benchmark** have different values on the anchoring effect condition and the non-anchoring effect condition, as shown in Figure 5A. Finally, the anchoring effect condition can improve the reliance loss over the non-anchoring effect condition, as shown in Figure 5C.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009