Does Explainable Artificial Intelligence Improve Human Decision-Making?

Yasmeen Alufaisan,[1] **Laura R. Marusich,** [2] **Jonathan Z. Bakdash,**[3] **Yan Zhou,**[4] **Murat Kantarcioglu** [4]

[1] EXPEC Computer Center at Saudi Aramco
Dhahran 31311, Saudi Arabia
[2] U.S. Army Combat Capabilities Development Command Army Research Laboratory South
at the University of Texas at Arlington
[3] U.S. Army Combat Capabilities Development Command Army Research Laboratory South
at the University of Texas at Dallas
[4] University of Texas at Dallas
Richardson, TX 75080
yasmeen.alufaisan@aramco.com, {laura.m.cooper20.civ, jonathan.z.bakdash.civ}@mail.mil,
{yan.zhou2,muratk}@utdallas.edu

Abstract

Explainable AI provides insights to users into the *why* for model predictions, offering potential for users to better understand and trust a model, and to recognize and correct AI predictions that are incorrect. Prior research on human and explainable AI interactions has

typically focused on measures such as interpretability, trust, and usability of the explanation. There are mixed findings whether explainable AI can improve actual human decision-making and the ability to identify the problems with the underlying model. Using real datasets, we compare objective human decision accuracy without AI (control), with an AI prediction (no explanation), and AI prediction with explanation. We find providing any kind of AI prediction tends to improve user decision accuracy, but no conclusive evidence that explainable AI has a meaningful impact. Moreover, we observed the strongest predictor for human decision accuracy was AI accuracy and that users were somewhat able to detect when the AI was correct vs. incorrect, but this was not significantly affected by including an explanation. Our results indicate that, at least in some situations, the *why* information provided in explainable AI may not enhance user decision-making, and further research may be needed to understand how to integrate explainable AI into real systems.

Introduction

Explainable AI is touted as the key for users to "understand, appropriately trust, and effectively manage... [AI systems])" (Gunning 2017) with parallel goals of achieving fairness, accountability, and transparency (Sokol 2019). There are a multitude of reasons for explainable AI, but there is little empirical research for its impact on human decision-making (Miller 2019; Adadi and Berrada 2018). Prior behavioral research on explainable AI has primarily focused on human understanding/interpretability, trust, and usability for different types of explanations (Doshi-Velez and Kim 2017; Hoffman et al. 2018; Ribeiro, Singh, and Guestrin 2016, 2018; Lage et al. 2019).

To fully achieve fairness and accountability, explainable AI should lead to better human decisions. Earlier research demonstrated that explainable AI can be understood by people (Ribeiro, Singh, and Guestrin 2018). Ideally, the combination of humans and machines will perform better than either alone (Adadi and Berrada 2018), such as computer-assisted chess (Cummings 2014), but this combination may not necessarily improve the overall accuracy of AI systems. While (causal) explanation and prediction share commonalities, they are not interchangeable concepts (Adadi and Berrada 2018; Shmueli et al. 2010; Edwards and Veale 2018). Consequently, a "good" explanation, interpretable model predictions, may not be sufficient for improving actual human decisions (Adadi and Berrada 2018; Miller 2019) because of heuristics and biases in human decision-making (Kahneman 2011). Therefore, it is important to demonstrate whether, and what types of, explainable AI can improve the decision-making performance of humans using that AI, relative to performance using the predictions of "black box" AI with no explanations and for human making decisions with no AI prediction.

In this work, we empirically investigate whether explainable AI improves human decision-making using a two-choice classification experiment with real-world data. Using human subject experiments, we compared three different settings where a user needs to make decision 1) No AI prediction (Control), 2) AI predictions but no explanation, and 3) AI predictions with explanations. Our results indicate that, while providing the AI predictions tends to help users, the *why* information provided in explainable AI does not specifically enhance user decision-making.

Background and Related Work

Using Doshi-Velez and Kim's (2017) framework for interpretable machine learning, our current work focuses on: real humans, simplified tasks. Because our objective is on eval-

uating decision-making, we do not compare different types of explanations and instead used one of the best available explanations: anchor LIME (Ribeiro, Singh, and Guestrin 2018). We use real tasks here, although our tasks involve relatively simple decisions with two possible choices. Additionally, we use lay individuals rather than experts. Below, we discuss prior work that is related to our experimental approach.

Explainable AI/Machine Learning

While machine learning models largely remain opaque and their decisions are difficult to explain, there is an urgent need for machine learning systems that can "explain" its reasoning. For example, European Union regulation requires "right to explanation" for any algorithms that make decisions significantly impacting users with user-level predictors (Parliament and Council of the European Union 2016). In response to the lack of consensus on the definition and evaluation of interpretability in machine learning, Doshi-Velez and Kim (2017) propose a taxonomy for the evaluation of interpretability focusing on the synergy among human, application, and functionality. They contrast interpretability with reliability and fairness, and discuss scenarios in which interpretability is needed. To unmask the incomprehensible reasoning made by these machine learning/AI models, researchers developed explainable models that are built on top of the machine learning model to explain their decisions. The most common forms of explainable models that provide explanations for the decisions made by machine learning models are feature-based and rule-based models. The feature-based models resemble feature selection where the model outputs the top features that explain the machine learning prediction and their associated weights (Datta, Sen, and Zick 2016; Ribeiro, Singh, and Guestrin 2016). The rule-based models provide simple if-then-else rules to explain predictions (Ribeiro, Singh, and Guestrin 2018; Alufaisan et al. 2017). It has been shown that rule-based models provide higher human precision when compared to feature-based models (Ribeiro, Singh, and Guestrin 2018).

Lou et al. (2012) investigate the generalized additive models (GAMs) that combine single-feature models through a linear function. GAMs are more accurate than simple linear models, and can be easily interpreted by users. Their empirical study suggests that a shallow bagged-tree with gradient boosting is the best method on low to medium dimensional datasets. Anchor LIME is an example of the current state-of-the-art explainable rule-based model (Ribeiro, Singh, and Guestrin 2018). It is a model-agnostic system that can explain predictions generated by any machine learning model with high precision. The model provides rules, referred to as anchors, to explain the prediction for each instance. A rule is an anchor if it sufficiently explains the prediction locally such that any changes to the rest of the features, features not included in the anchor, do not effect the prediction. Anchors can be found in two different approaches: bottom-up approach and beam search. Wang et al. (2017) present a machine learning algorithm that produces Bayesian rule sets (BRS) comprised of short rules in the disjunctive normal form. They develop two probabilistic models with prior parameters that allow the user to specify a desired size and shape and balance between accuracy and interpretability. They apply two priors—beta-binomials and Poisson distribution—to constrain the rule generation process and provide theoretical bounds for reducing computation by iteratively pruning the search space. In our experiments, we use anchor LIME to provide explanations for all our experimental evaluation due to the high human precision of anchor LIME as reported in Ribeiro, Singh, and Guestrin (2018).

Human Decision-Making and Human Experiments with Explainable AI

A common reason for providing explanation is to improve human predictions or decisions (Keil 2006). People are not necessarily rational (i.e., maximizing an expected utility function). Instead, decisions are often driven by heuristics and biases (Kahneman 2011). Also, providing more information, even if relevant, does not necessarily lead people to making better decisions (Gigerenzer and Brighton 2009). Bounded rationality in human decision-making using satisfying with constraints (Gigerenzer and Brighton 2009) is an alternative theory to heuristics and biases (Kahneman 2011). Regardless of the theoretical account for human decision-making, people, which can include experts (Dawes, Faust, and Meehl 1989), generally do not make fully optimal decisions.

At a minimum, explainable AI should not be detrimental to human decision-making. The literature on decision aids (a computational recommendation or prediction, typically without an explicit explanation) has mixed findings for human performance. Sometimes these aids are beneficial for human decision-making, whereas at other times they have negative effects on decisions (Kleinmuntz and Schkade 1993; Skitka, Mosier, and Burdick 1999). These mixed findings may be attributable to absence of explanations; this can be investigated through human experiments testing AI predictions with explanations compared with AI predictions alone.

Most prior human experiments with explainable AI have concentrated on interpretability, trust, and subjective measures of usability, such as preferences and satisfaction, with work on decision-making performance remaining somewhat limited (Miller 2019; Adadi and Berrada 2018). Earlier results suggest explainable AI can increase interpretability (e.g. Ribeiro, Singh, and Guestrin 2018), trust (e.g. Lakkaraju and Bastani 2020; Ribeiro, Singh, and Guestrin 2016; Selvaraju et al. 2017), and usability (e.g. Ribeiro, Singh, and Guestrin 2018) to varying degrees, but this does not necessarily translate to better performance on real-world decisions about the underlying data, such as whether to actually use the AI's prediction, whether the AI has made an error, and the role of explanations. In fact, recent work has shown that subjective measures commonly assessed (e.g., preference and trust) do not predict actual human performance (Buçinca et al. 2020; Zhang, Liao, and Bellamy 2020); similarly, performance on common proxy tasks such as predicting the AI's decision also may not be indicative of actual decision-making performance (Buçinca et al. 2020).

These findings highlight the need for more study of the impact of AI explanation on objective human performance, not just proxy or subjective measures.

In the limited studies that do examine the effect of explanation on human decision-making performance, there are mixed findings about whether the explanation provides an additional benefit over AI prediction alone. For example, some researchers found that human performance was better when an AI prediction was accompanied by explanation than performance with the prediction alone (Buçinca et al. 2020; Lai and Tan 2019) However, other studies did not show any additional benefit of explanation over AI prediction alone (Green and Chen 2019), with some even showing evidence of worse performance with explanation (Poursabzi-Sangdeh et al. 2018; Zhang, Liao, and Bellamy 2020).

The two papers finding a benefit for explanations consisted of a task in which users made decisions about the fat content in pictures of food (Buçinca et al. 2020) and judgments about whether text from hotel reviews were genuine or deceptive (Lai and Tan 2019). They also both used a simple binary choice as the decision-making task. In contrast, the work finding no improvement in decision accuracy with explainable AI used datasets that comprised variables and outcomes, including probabilistic assessments for risks with recidivism and loan outcomes (Green and Chen 2019), decisions about real estate valuations (Poursabzi-Sangdeh et al. 2018), and predictions about income (Zhang, Liao, and Bellamy 2020). In addition, instead of simple binary choices, these studies used prediction of values along a continuum (Green and Chen 2019; Poursabzi-Sangdeh et al. 2018), and binary choice with the option to switch after seeing the model prediction (Zhang, Liao, and Bellamy 2020).

Besides dataset and task differences, there are two other distinctions among these papers. Only a single paper assessed decision-making under time pressure (Zhang, Liao, and Bellamy 2020) and only two papers informed users if their decisions were correct or incorrect (Green and Chen 2019; Zhang, Liao, and Bellamy 2020). Our study design uses datasets of multiple variables and outcomes and provides correct/incorrect feedback, but also uses a very straightforward binary choice task. This combination could potentially resolve the disparity in results from the studies above.

Methods

In this section, we first describe the two datasets used in our experiments. We then provide the details of our experimental design and hypotheses, participant recruitment, and general demographics of our sample.

Dataset

To conduct our experiments, we choose two different datasets that have been heavily used in prior research that tries to understand algorithmic fairness and accountability issues. For example, the COMPAS dataset has been used to detect potential biases in criminal justice system (Angwin et al. 2016). The Census income dataset, which has been used to test many machine learning techniques, involves predictions of individuals' income status. This has been associated with potential biases in making decisions such as access to credit and job opportunities.

We choose these datasets primarily because they both involve real-world contexts that are understandable and engaging for human participants. Further, the two datasets differ widely in number of features and in the overall accuracy classifiers can achieve in their predictions. This allows us to explore the effects of these differences on human performance; in addition, it ensures that our findings are not limited only to a specific dataset. We briefly discuss each dataset in more detail below.

COMPAS stands for Correctional Offender Management Profiling for Alternative Sanctions (Angwin et al. 2016). It is a scoring system used to assign risk scores to criminal defendants to determine their likelihood of becoming a recidivist. The data has 6,479 instances and 7 features. These features are: gender, age, race, priors count, and charge degree risk score, and whether the defendants re-offended in two years or not. We let the binary re-offending feature be our class.

Census income (CI) data contains information used to predict individuals' income (Dua and Graff 2017). It has 32,561 instances and 14 features. These features are: age, workclass, education, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, and country. The class value is low income (less or equal to 50K) or high income (greater than 50K). We preprocessed the dataset to allow equal class distribution [1].

Experimental Design

Prior results demonstrating people interpret, trust, and prefer explainable AI, suggesting it will improve the accuracy of human decisions. Hence, our primary hypotheses are that explainable AI would aid human decision-making. The hypotheses (H.) are as follows:

H. 1 Explainable AI enhances decision-making process compared to only an AI prediction (without explanation) and a control condition with no AI.

[H. 1.a] A participant performs above chance in prediction tasks.

H. 2 A participant's decision accuracy is positively associated with AI accuracy.

H. 3 Average participant's decision accuracy does not outperform AI accuracy.

H. 4 Participants outperform AI accuracy more often with explainable AI over AI prediction alone.

H. 5 Participants follow explainable AI recommendation more often than AI only recommendation.

H. 6 Explainable AI increases participants' decision confidence.

H. 7 A participant's decision confidence is positively correlated with the accuracy of his/her decision.

[1] The CI dataset is from 1994. We adjusted for inflation by using a present value of 88k. From 1994 to January 2020 (when the experiment was run) inflation in the U.S. was 76.45%: https://www.wolframalpha.com/input/?i=inflation+from+1994+to+jan+2020

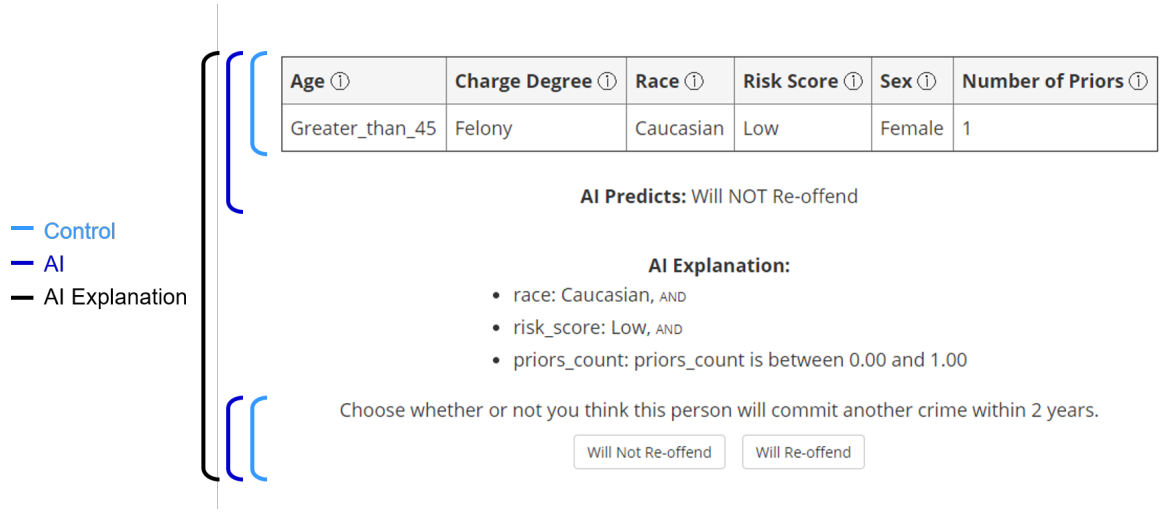

Figure 1: Example from the study demonstrating the information appearing in the three AI conditions for a trial from the COMPAS dataset condition.

To investigate these hypotheses, we used a 2 (Dataset: Census and COMPAS) x 3 (AI condition: Control, AI, and AI with Explanation) between-participants experimental design. The three AI conditions were:

- Control: Participants were provided with no prediction or information from the AI.
- AI: Participants were provided with only an AI prediction.
- AI with Explanation: Participants received an AI prediction, as well as an explanation of the prediction using anchor LIME (Ribeiro, Singh, and Guestrin 2018).

To achieve more than 80% statistical power to detect a medium effect size for this design, we planned for a sample size of $N = 300$ (50 per condition).

In all conditions, each trial consists of a description of an individual and a two-alternative forced choice for the classification of that individual. Each choice was correct on 50% of the trials, thus chance performance for human decision-making accuracy was 50%. Additionally, an AI prediction and/or explanation may appear, depending on the AI condition (see Figure 1). After a decision is made, participants are asked to enter their confidence in that choice, on a Likert scale of 1 (No Confidence) to 5 (Full Confidence). Feedback is then displayed, indicating whether or not the previous choice was correct.

We compared the prediction accuracy of Logistic Regression, Multi-layer Perceptron Neural Network with two layers of 50 units each, Random Forest, Support Vector Machine (SVM) with rbf kernel and selected the best classifier for each dataset. We chose a Multi-layer Perceptron Neural Network for Census income data where it resulted in an overall accuracy of 82% and SVM with rbf kernel for COMPAS data with an overall accuracy of 68%. Census income accuracy closely matches the accuracy reported in the literature (Dua and Graff 2017; Alufaisan, Kantarcioglu, and Zhou 2016) and COMPAS accuracy matches the results published by ProPublica (Angwin et al. 2016). We split the data to 60% for training and 40% for testing to allow enough instances for the explanations generated using anchor LIME (Ribeiro, Singh, and Guestrin 2018).

In our behavioral experiment, 50 instances were randomly sampled without replacement for each participant. Thus, AI accuracy was experimentally manipulated for participants (Census: mean AI accuracy = 83.85%, sd = 3.67%; COMPAS: mean AI accuracy = 69.18%, sd = 4.65%). Because of the sample size and large number of repeated trials per participant, there was no meaningful difference in mean AI accuracy for participants in the AI condition vs. those in the AI explanation condition ($p = 0.90$).

Participant Recruitment and Procedure

We developed the experiment using jsPsych (De Leeuw 2015), and hosted it on the Volunteer Science platform (Radford et al. 2016) [2]. Participants were recruited using Amazon Mechanical Turk (AMT) and were compensated $4.00 each. We collected data from 50 participants in each of the six experimental conditions, for a total of 300 participants (57.67% male). Most participants were 18 to 44 years old (80.67%). This research was approved as exempt (19-176) by the Army Research Laboratory's Institutional Review Board.

Participants read and agreed to a consent form, then received instructions on the task, specific to the experimental condition they were assigned to. They completed 10 practice trials, followed by 50 test trials and a brief questionnaire assessing general demographic information and comments on strategies used during the task. The median time to complete the practice and test trials was 18 minutes.

Results and Discussion

In this section we analyze and describe the effects of dataset, AI condition, and AI accuracy on the participants' decision-

[2] https://volunteerscience.com/

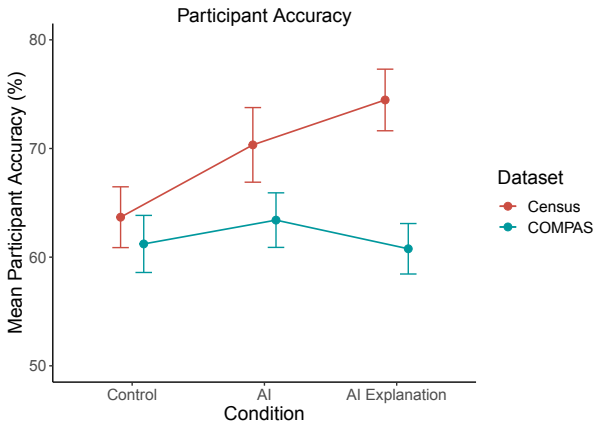

Figure 2: Mean participant accuracy in each AI and dataset condition. Error bars represent 95% confidence intervals.

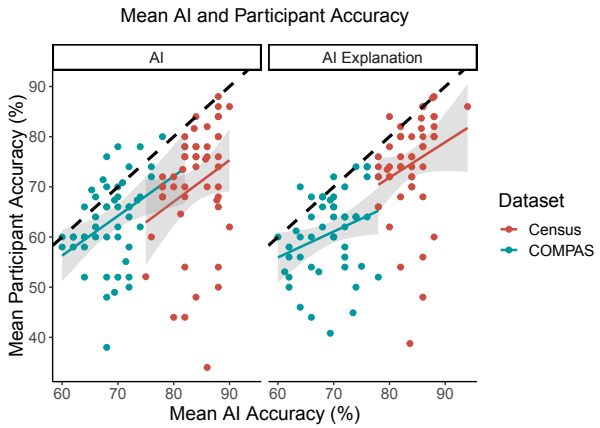

Figure 3: Mean AI accuracy (per participant) and mean participant accuracy by AI and AI Explanation and the two datasets. The shaded areas represent 95% confidence intervals.

making accuracy, ability to outperform the AI, adherence to AI recommendations, confidence ratings, and reaction time.

Participant Decision-Making Accuracy

We compared participants' mean accuracy in the experiment across conditions using a 2 (Dataset) x 3 (AI) factorial Analysis of Variance (ANOVA) (see Figure 2). We found significant main effects, with a small effect size for AI condition ($F(2, 294) = 8.19, p < 0.001, \eta^2 = 0.04$) and a nearly large effect for dataset condition ($F(1, 294) = 46.51, p < 0.001, \eta^2 = 0.12$). In addition, there was a significant interaction with a small effect size ($F(2, 294) = 8.38, p < 0.001, \eta^2 = 0.05$), indicating that the effect of AI condition depended on the dataset. Specifically, the large effect for increased accuracy with AI was driven by the Census dataset.

Contrary to H. 1, explainable AI did not substantially improve decision-making accuracy over AI alone. We followed up on significant ANOVA effects by performing pairwise comparisons using Tukey's Honestly Significant Difference. These post-hoc tests indicated that participants who viewed the Census dataset showed improved accuracy over control when given an AI prediction ($p < 0.01$) and higher accuracy with AI explanation versus control ($p < 0.001$), but there was no statistically significant difference in participant accuracy for AI compared to AI explanation ($p = 0.28$). Whereas the COMPAS dataset had no significant differences in participant accuracy across pairwise comparisons for the three AI conditions ($ps > 0.75$). Also, the mean participant accuracy for the COMPAS control condition (mean = 63.7%, sd = 9.24%) was comparable to participant accuracy for prior decision-making research using the same dataset (mean = 62.8%, sd = 4.8%) (Dressel and Farid 2018).

There was strong evidence supporting H. 1.a, the vast majority of participants had mean accuracy exceeding guessing (50% accuracy). The overall participant accuracy across all conditions was 65.65% (sd = 10.92%), with 90% (or 270 out of 300) participants performing above chance on the classification task. This indicates that the task was challenging but feasible for almost all participants.

AI Accuracy and Participant Decision-Making Accuracy

We also evaluated the effect of the randomly varied AI accuracy for each participant on their decision-making accuracy. We used linear regression to analyze this relationship, specifying participant accuracy as the dependent variable and the following as independent variables: mean AI accuracy (per participant), AI condition, and dataset condition, see Figure 3. Regressions are represented by the solid lines with the shaded areas representing 95% confidence intervals. The control condition is not included in the analysis or figure, because the accuracy of the AI is not relevant if no AI prediction is presented to the participant. The overall regression model was significant with a large effect size, $F(4, 195) = 21.23, p < 0.001, R^2_{adjusted} = 0.29$. Consistent with H. 2, there was a large main effect for AI accuracy ($\beta = 0.70, p < 0.001, R^2 = 0.28$). Also, there was a small AI accuracy and dataset interaction ($\beta = -0.07, p < 0.01, R^2 = 0.03$), reflecting the same interaction depicted in Figure 2. There were no significant regression differences for dataset or AI versus AI Explanation, $ps > 0.60$; there was no significant effect of dataset because it largely drove AI accuracy. Note it is not just that participants perform better with the higher mean AI accuracy of the Census dataset, both datasets had large positive relationships with participant accuracy and corresponding mean AI accuracy shown in Figure 3.

Outperforming the AI Accuracy An interesting question is whether the combination of AI and human decision-making can outperform either alone. The previous analyses showed that the addition of AI prediction information improved human performance over controls with humans alone. We also evaluated how often the human decision-making accuracy outperformed the accuracy of the corresponding mean AI prediction accuracy, which was experimentally manipulated. Although most participants per-

Dataset/Condition	AI	AI Explanation
Census	0	1
COMPAS	10	3

Table 1: Number of participants with decision accuracy exceeding their mean AI accuracy.

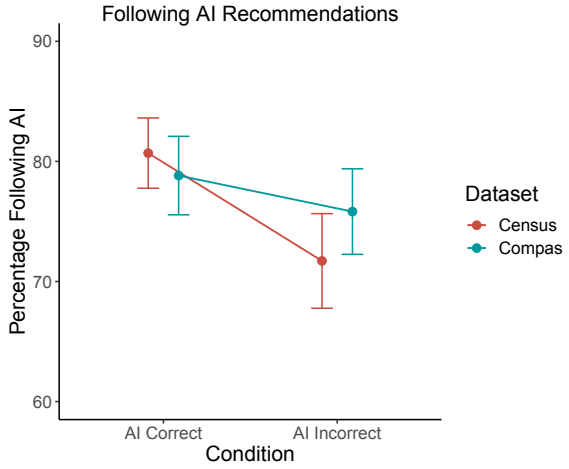

Figure 4: Mean proportion of participant choices matching AI prediction as a function of whether the AI correct/incorrect and the dataset condition. Error bars represent 95% confidence intervals. To simplify this figure, results were collapsed for the AI and AI Explanation conditions which did not have a significant main effect, $p = 0.62$.

formed well above chance, only a relatively small number of participants had decision accuracy exceeding their mean AI prediction (7% or 14 out of 200). This result largely supports H. 3 and also shown above in Figure 3 where each dot represents an individual; and dots above the black dashed line show the participants that outperformed their mean AI prediction. The black dashed line shows equivalent performance for mean AI accuracy and mean participant accuracy.

Adherence to AI Model Predictions Participants followed the AI predictions more often when the AI was correct versus when the AI was incorrect, indicating some recognition of when the AI makes bad predictions (see Figure 4, $F(1, 196) = 36.15, p < 0.001, \eta_p^2 = 0.16$). This was consistent with participant sensitivity to AI recommendations, evidence for H. 2. Also, participants were better able to recognize correct vs. incorrect AI predictions when they were in the Census condition, demonstrated in the significant interaction between AI correctness and dataset, $F(1, 196) = 9.01, p < 0.01, \eta_p^2 = 0.04$. None of the remaining ANOVA results were significant, $ps > 0.16$. Thus, there was no evidence for higher adherence to recommendations with explainable AI, which rejected H. 5.

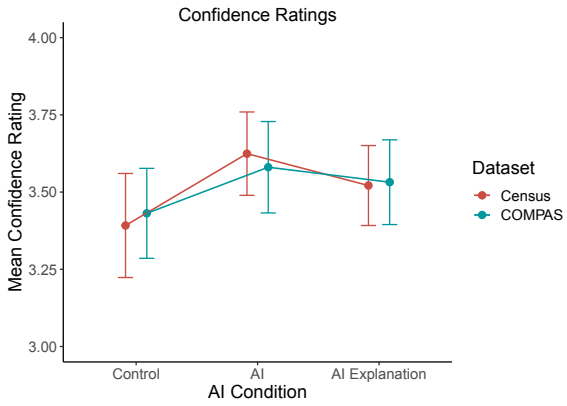

Figure 5: Mean participant confidence ratings in each AI condition. Error bars represent 95% confidence intervals.

Confidence Ratings

We found that AI (without and with explanation) resulted in slightly increased mean confidence. There was a small effect of AI condition on mean confidence (see Figure 5, $F(2, 294) = 3.58, p = 0.03, \eta^2 = 0.02$). Post hoc tests indicated participants had significantly lower mean confidence in the control condition than AI, $p < 0.03$, but there were no statistical differences for other pairwise comparisons, $ps > 0.25$. This contradicted H. 6, and there was no evidence of a confidence increase with explanations. In addition, there was no evidence for a main effect of dataset condition or interaction, $ps > 0.84$.

Confirming H. 7, we found a positive relationship for accuracy and confidence rating within individuals indicating that participants' confidence ratings were fairly well-calibrated with their actual decision accuracy. We calculated each participant's mean accuracy at each confidence rating they used, and then conducted a repeated measures correlation (Bakdash and Marusich 2017) ($r_{rm} = 0.48, p < 0.001$).

Additional Results

We also assessed reaction time and summarize self-reported decision-making strategies. These results are exploratory, there were no specific hypotheses. There was no significant main effect of AI condition on participants' reaction time ($F(2, 294) = 2.13, p = 0.12, \eta^2 = 0.01$). There was only a main effect of dataset condition ($F(1, 294) = 28.52, p < 0.001, \eta^2 = 0.09$), where participants took an average of 1600 ms longer in the Census condition than the COMPAS condition (see Figure 6). This effect was most likely due to the Census dataset having more variables for each instance than the COMPAS dataset, and thus requiring more reading time on each trial. The addition of an explanation did not meaningfully increase reaction time over an AI prediction only.

Subjective measures, such as self-reported strategies and measures of usability, often diverge from objective measures of human performance (Andre and Wickens 1995; Nisbett and Wilson 1977) such as actual decisions (Buçinca et al.

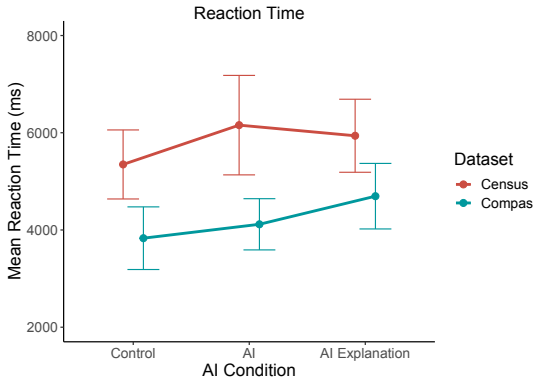

Figure 6: Mean reaction time in each AI and dataset condition. Error bars represent 95% confidence intervals.

2020). Participants self-reported varying strategies to make their decisions, yet there was a clear benefit for AI prediction (without and with explanation). In the AI and AI explanation conditions: $n = 80$ indicated using the data without mentioning AI, $n = 39$ reported using a combination of the data and the AI, and only $n = 16$ said they primarily used, trusted, or followed the AI. Despite limited self-reported use of the AI in the two relevant conditions, decision accuracy was higher with AI (Figure 2), strongly associated with AI accuracy (Figure 3), and there was some sensitivity to whether the AI was followed when it was correct versus incorrect (Figure 4). Nearly 80% of user comments could be coded, blank and nonsense responses could not be coded.

Discussion

Our results show providing an AI prediction enhances human decision accuracy, but in opposition to the hypotheses adding an explanation positively impact decisions and increase the ability to outperform the AI. This finding is in line with some previous studies that also found no added benefit for explanation over AI prediction alone (Green and Chen 2019; Poursabzi-Sangdeh et al. 2018; Zhang, Liao, and Bellamy 2020). This suggests that it was not the simplicity of the decision-making task that accounts for the opposite findings in Buçinca et al. (2020) and Lai and Tan (2019), since the current study also uses a relatively straightforward binary decision. Rather, it may be the case that tasks with highly intuitive datasets are required for explanation to improve performance. Future work may address this question directly.

One possible explanation for findings of no added benefit of explanation is that providing more information, even if task-relevant, does not necessarily improve human decision-making accuracy (Gigerenzer and Brighton 2009; Goldstein and Gigerenzer 2002; Nadav-Greenberg and Joslyn 2009). This phenomenon is attributed to cognitive limitations and people using near-optimal strategies, and corresponds with Poursabzi-Sangdeh et al.'s (2018) findings that explanation caused information overload, reducing people's ability to detect AI mistakes. However, their study and most other papers did not use the speeded response paradigm we used here, suggesting this was not solely attributable to participants responding as quickly as possible.

The lack of a significant, practically-relevant effect for explainable AI was not due to lack of statistical power or ceiling performance - nearly all participants consistently performed above chance, but well below perfect accuracy. These findings also illustrate the need to compare decision-making with explainable AI to other conditions including no AI and AI prediction without explanation. If we did not have an AI only (decision aid) condition a reasonable but flawed inference would have been that explainable AI enhances decisions.

Limitations The present findings have limited generalizability to decision-making with other datasets and explainable AI techniques. For example, the effectiveness of explanations, or lack thereof, for human decision-making may depend on a variety of factors: the specific explanation technique, the properties of the dataset, and the task itself (such as probabilistic predictions vs two choice decisions). Nevertheless, this paper demonstrates that one cannot assume explainable AI will necessary improve human decisions and the need to evaluate objective measures, of human performance, in explainable AI.

Conclusions and Future Work

Much of the existing research on explainable AI focuses on the usability, trust, and interpretability of the explanation. In this paper, we fill in the research blank by investigating whether explainable AI can improve human decision-making. We design a behavioral experiment in which each participant recruited using Amazon Mechanical Turk is asked to complete 50 test trials in one of six experimental conditions. Our experiment is conducted on two real datasets to compare human decision with an AI prediction and an AI with explanation. Our experimental results demonstrate that AI predictions alone can generally improve human decision accuracy, while the advantage of explainable AI is not conclusive. We also show that users tend to follow AI predictions more often when the AI predictions are accurate. In addition, AI with or without explanation can increase the confidence of human users which, on average, was well-calibrated to user decision accuracy.

In the future, we plan to investigate whether explainable AI can help improve fairness, safety, and ethics by increasing the transparency of AI models. Human decision-making is a key outcome measure, but is certainly not the only goal for explainable AI. We also plan to explore the difference of distributions in the error space between human decision and AI predictions, especially at decision boundaries. Also, whether human-machine collaboration is feasible through interactions in closed feedback loops. We will also expand our datasets to include other data format such as images.

Acknowledgments

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied,

of the U.S. Army Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation. M.K. and Y.Z. were supported by ARL under grant W911NF-17-1-0356. We thank Jason Radford for help with implementing the experiment on the Volunteer Science platform and Katelyn Morris for independently coding the comments. We also thank Dan Cassenti for input on the paper, and Jessica Schultheis and Alan Breacher for editing the paper.

Ethical Impact

For many important critical decisions, depending on the AI model prediction may not be enough. Furthermore, many recent regulations such as GDPR (Parliament and Council of the European Union 2016) allow potential audit of AI prediction by a human. Therefore, it is critical to understand whether the explanations provided by explainable AI methods improve the overall prediction accuracy, and help human decision makers to detect errors. Our results indicate that although the existence of an AI model may improve human decision-making, the explanations provided may not automatically improve the accuracy. We believe that our results could help ignite the needed research to explore how to better integrate explanations, AI models and human operators to have better outcomes compared AI models or humans alone.

Another consideration is the data itself, models created using systematically biased data will simply mirror and reinforce patterns inherent to the data. For example, in the COMPAS dataset a high risk score for re-offending has far lower accuracy for black defendants than white ones (Angwin et al. 2016). Furthermore, multiple AI systems for predicting criminal activity have relied on "dirty" data with racial bias due to flawed polices and procedures (Richardson, Schultz, and Crawford 2019). One solution may be to combine approaches for fair machine learning (Corbett-Davies and Goel 2018) with explainable AI, while considering dataset properties such as its provenance.

References

Adadi, A.; and Berrada, M. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6: 52138–52160.

Alufaisan, Y.; Kantarcioglu, M.; and Zhou, Y. 2016. Detecting Discrimination in a Black-Box Classifier. In *In Proceedings of the 2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*. IEEE.

Alufaisan, Y.; Zhou, Y.; Kantarcioglu, M.; and Thuraisingham, B. 2017. From Myths to Norms: Demystifying Data Mining Models with Instance-Based Transparency. In *In Proceedings of the 2017 IEEE 3rd International Conference on Collaboration and Internet Computing (CIC)*. IEEE.

Andre, A. D.; and Wickens, C. D. 1995. When users want what's not best for them. *Ergonomics in design* 3(4): 10–14.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bakdash, J. Z.; and Marusich, L. R. 2017. Repeated measures correlation. *Frontiers in psychology* 8: 1–13. doi: 10.3389/fpsyg.2017.00456.

Buçinca, Z.; Lin, P.; Gajos, K. Z.; and Glassman, E. L. 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*, IUI '20, 454–464. doi:10.1145/3377325.3377498.

Corbett-Davies, S.; and Goel, S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* .

Cummings, M. M. 2014. Man versus machine or man+ machine? *IEEE Intelligent Systems* 29(5): 62–69.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *In Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP)*, 598–617.

Dawes, R. M.; Faust, D.; and Meehl, P. E. 1989. Clinical versus actuarial judgment. *Science* 243(4899): 1668–1674.

De Leeuw, J. R. 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods* 47(1): 1–12.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* .

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4(1): eaao5580.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. URL http://archive.ics.uci.edu/ml. Accessed on 05.2020.

Edwards, L.; and Veale, M. 2018. Enslaving the algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"? *IEEE Security & Privacy* 16(3): 46–54.

Gigerenzer, G.; and Brighton, H. 2009. Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science* 1(1): 107–143.

Goldstein, D. G.; and Gigerenzer, G. 2002. Models of ecological rationality: the recognition heuristic. *Psychological review* 109(1): 75.

Green, B.; and Chen, Y. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3(CSCW). doi:10.1145/3359152. URL https://doi.org/10.1145/3359152.

Gunning, D. 2017. Explainable artificial intelligence (xai). URL https://www.darpa.mil/attachments/XAIProgramUpdate.pdf. Accessed on 05.2020.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608* .

Kahneman, D. 2011. *Thinking, fast and slow*. Macmillan.

Keil, F. C. 2006. Explanation and understanding. *Annu. Rev. Psychol.* 57: 227–254.

Kleinmuntz, D. N.; and Schkade, D. A. 1993. Information displays and decision processes. *Psychological Science* 4(4): 221–227.

Lage, I.; Chen, E.; He, J.; Narayanan, M.; Kim, B.; Gershman, S.; and Doshi-Velez, F. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006* .

Lai, V.; and Tan, C. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38.

Lakkaraju, H.; and Bastani, O. 2020. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 79–85. ACM. doi:10.1145/3375627.3375833.

Lou, Y.; Caruana, R.; and Gehrke, J. 2012. Intelligible Models for Classification and Regression. In *KDD'12, Beijing, China*. ACM.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267: 1–38.

Nadav-Greenberg, L.; and Joslyn, S. L. 2009. Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making* 3(3): 209–227.

Nisbett, R. E.; and Wilson, T. D. 1977. Telling more than we can know: verbal reports on mental processes. *Psychological review* 84(3): 231.

Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of such Data, and Repealing Directive 95/46/ EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Vaughan, J. W.; and Wallach, H. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* .

Radford, J.; Pilny, A.; Reichelmann, A.; Keegan, B.; Welles, B. F.; Hoye, J.; Ognyanova, K.; Meleis, W.; and Lazer, D. 2016. Volunteer science: An online laboratory for experiments in social psychology. *Social Psychology Quarterly* 79(4): 376–396.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 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.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Richardson, R.; Schultz, J. M.; and Crawford, K. 2019. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online* 94: 15.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Shmueli, G.; et al. 2010. To explain or to predict? *Statistical science* 25(3): 289–310.

Skitka, L. J.; Mosier, K. L.; and Burdick, M. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51(5): 991–1006.

Sokol, K. 2019. Fairness, Accountability and Transparency in Artificial Intelligence: A Case Study of Logical Predictive Models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 541–542.

Wang, T.; Rudin, C.; Doshi-Velez, F.; Liu, Y.; Klampfl, E.; and MacNeille, P. 2017. A Bayesian Framework for Learning Rule Sets for Interpretable Classification. *Journal of Machine Learning Research* 18(70): 1–37. URL http://jmlr.org/papers/v18/16-003.html.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. E. 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*, FAT* '20, 295–305. doi:10.1145/3351095.3372852.