When Algorithms Explain Themselves: AI Adoption and Accuracy of Experts' Decisions

Himabindu Lakkaraju* Chiara Farronato†

June 2022

PRELIMINARY AND INCOMPLETE
PLEASE DO NOT CIRCULATE

Abstract

We explore whether providing explanations about machine learning algorithms helps experts better use algorithmic recommendations. We design hypothetical surveys for doctors and recruiters, who have to make predictions about the likelihood of certain events to guide their treatment and hiring decisions. We progressively add explanations about how the prediction algorithm makes its decisions. We find that as experts receive more explanations about the behavior of the prediction algorithm, they increase their reliance on the algorithmic recommendation and approach the algorithm's accuracy level compared to experts who make predictions without algorithmic support. We also find that experts perform even better than the algorithm alone when given the most complete set of information about the algorithm's process, suggesting that there are valuable complementarities between the algorithm's and the expert's information set. However, we also find that explanations can backfire, by giving experts a false sense of trust in machine learning models. Our results have implications for the adoption and deployment of machine learning algorithms in contexts where experts, such as judges and doctors, make decisions.

1 Introduction

Despite the value that recommendation algorithms have played in technology platforms – from matching drivers and passengers on Uber to suggesting experiences and accommodations on Airbnb – their adoption has so far been slow ([4], [5]), particularly in contexts where skilled experts make decisions, such as doctors and judges([11]). While in many contexts a full substitution from humans to machines is fairly straightforward and efficient (in the case of taxi dispatch systems for example),

*Harvard Business School, hlakkaraju@hbs.edu
†Harvard Business School, CEPR, and NBER, cfarronato@hbs.edu

in other contexts, the role that the expert plays goes beyond making a simple prediction. In those contexts, experts can better account for ambiguity than machines, and experts can add a valuable "human touch" while performing the prediction task. In the case of doctors and patients, for example, having a human describe a diagnosis to a patient and address any related questions and fears may be preferred to receiving an email with an automated message. Machines can also make unexpected errors, which a human can more easily correct.

Despite the important role that experts play, multiple studies have shown that algorithms can and do better than the average expert [11, 13]. This raises the question of how we can nudge experts to better use algorithmic recommendations without losing the role that experts play in decoding ambiguity and non standard information. In this sense, we are interested in exploring the complementarities between the predictions of humans and algorithms and which set of nudges can better leverage these complementarities.

In this paper, we study the role of explanations in driving experts' decisions to follow algorithmic recommendations and the accuracy that the supported decision achieves. We give doctors hypothetical patients and ask them to make diagnostic predictions, with and without the support of algorithmic recommendations. When the algorithm provides recommendations, we also vary the degree of explanations provided with the recommendation. We progressively add explanations about the predicted probability of the assigned label; a list of top features leading to that particular recommendation; and the accuracy of a prediction solely based on the listed features. We conduct a similar experiment with recruiters who make hypothetical screening decisions to decide which candidates to interview for a job.

On the positive side, we find that adding explanations improve experts' decision making: their choices become more accurate compared to the group of doctors who are not given algorithmic recommendations, and compared to groups of doctors who are given fewer explanations. In the scenario with the most complete set of explanations, decision accuracy is even higher than the accuracy of the machine learning model alone, suggesting valuable complementarities from making a joint human-machine prediction.

On the negative side, providing explanations makes expert spend more time on each decision, perhaps due to the cognitive load associated with processing the additional information. Most importantly however, providing explanations can backfire. When experts are given a list of features that the algorithm considers when making a prediction, experts trust the algorithm more regardless of the importance of those features for the actual prediction. When the features are irrelevant, experts' over-reliance on the algorithm leads them to be even less accurate than the scenario where they are not provided any algorithmic advice. This over-reliance is somewhat corrected for doctors when they are given a measure of the accuracy of those features, but not for recruiters, who keep over-relying on the algorithmic recommendation.

Our results provide novel evidence on the role of explanations in the context of algorithmic recommendations. We show that experts can be nudged to adopt algorithmic recommendations, es-

pecially when those recommendations are accurate and are accompanied by convincing explanations of the algorithmic process.

The result that both accuracy and explainability of the algorithm's prediction matter for experts' trust in the algorithm emphasizes an important trade-off when deploying algorithmic recommendations. On one end, experts are persuaded by highly accurate algorithms, which calls for sophisticated neural nets and black-box type of techniques. On the other hand, experts are also persuaded by explanations of the factors contributing to the prediction, which calls for simpler models, such as decision trees or regression techniques. We are exploring this trade-off directly in follow-up surveys.

Our work is related to the recent and growing literature on the adoption of machine learning algorithms to aid in prediction tasks ([3]), which is frequently described as a general purpose technology applicable across a broad range of industries and problems ([10]). The literature has typically focused on the substitutability between humans and algorithms ([8]) on a task-by-task basis ([2], [1]), although more recently complementarities across tasks within an organization have been recognized as possible drivers of adoption ([4], [5]). Our research is most closely connected to studies of how AI algorithms improve, rather than substitute, human decision making ([7], [6], and [9]).

The paper is organized as follow. Section 2 describes our study design. Section 3 presents our results, and Section 4 concludes.

2 Study Design

We want to test whether adding information about how a recommendation algorithm reaches a particular recommendation helps experts improve their decisions. For this we have designed a set of treatments that progressively add explanations about the algorithm's prediction process.

We have two parallel sets of experiments, one for doctors and one for recruiters. Doctors were asked whether a particular patient is likely to develop breast cancer in the following 2 years (see Table 1 for an example). Recruiters were asked whether a particular candidate is likely to perform well in an entry-level software engineering job. We call recruiters and doctors *experts*.

Experts were placed in one of 5 conditions (Table 2):

1. *Expert only*: Experts were asked to answer their relevant question without the support of any algorithmic recommendation;

2. *Expert + Prediction Label*: Experts were given the label ("is the patient at risk of breast cancer?" in the medical context; "will this candidate perform well as an entry-level software engineer?" in the recruiting context) from a prediction model. The classification model used for the medical predictions is a deep neural network with 5 hidden layers and 10 nodes per layer. Its underlying accuracy is 89.09%. The model used for the hiring predictions is a deep neural network with 5 hidden layers and 15 nodes per layer, with an underlying accuracy of 88.77%.

Table 1: Survey Template

Below is some information about a patient: Gender: Female Age: 32 Pre-existing conditions: Diabetes, High blood pressure Family health history: Anxiety, Depression ESR (Erythrocyte Sedimentation Rate): Normal (<= 20 mm/hr) Exercise: 3 times a week Smoker: At least 5 times a week Alcohol: Once a month Current symptoms: Cold, Cough, Fever Commonly experienced symptoms: Cold, Menstrual Cramps Avg. WBC (last 3 years): Normal Avg. RBC (last 3 years): Normal Avg. HGB (last 3 years): Low Avg. Glucose (last 3 years): High Number of breast exams (last 3 years): 2 Status of breast exams (last 3 years): Normal The ML model that we built, which has an overall accuracy of 89.09% (See link for detailed FPR, TPR, FNR, TNR values), predicts that this patient is not likely to be diagnosed with breast cancer within the next 2 years with a probability of 0.86. Using the information provided above, can you predict if this patient is likely to be diagnosed with breast cancer within the next 2 years?

The table provides a template for the questions given to doctors in Treatment 3 (see Figure 2 for a list of treatments.

3. *Expert + Label + Predicted probability of success*: In addition to the predicted label, experts were also given the predicted probability of success from the classification model. *Success* here means being at risk of cancer in the medical context, and being a good candidate for a software engineering job in the recruiting context.

4. *Expert + Label + Predicted Probability + 4 Features*: In addition to the predicted label and predicted probability, experts were also told 4 features that the algorithm used in its prediction. Since some features are more important than others in driving the prediction, we further break

Table 2: Treatment Arms

Treatment	1	2	3	4	5
	Expert	Expert Label	Expert Label Probability	Expert Label Probability 4 Features	Expert Label Probability 4 Features Feature Accuracy
Sub-Condition					
A				4/4 Top Features	4/4 Top Features
B				3/4 Top Features	3/4 Top Features
C				2/4 Top Features	2/4 Top Features
D				1/4 Top Features	1/4 Top Features
E				0/4 Top Features	0/4 Top Features

The table describes the five main treatment conditions, 1 *through* 5*, as well as the five sub-conditions, A through E, for treatments* 4 *and* 5*.*

down this treatment into 5 sub-conditions, described below.

5. *Expert + Label + Predicted Probability + 4 Features + Feature Accuracy*: In addition to the elements in previous treatment arms, experts were told the accuracy of a prediction based on the 4 features listed. To calculate this accuracy metric, we keep those 4 features constant and manipulate the other predictors. If the prediction can change dramatically depending on the value that the other predictors take, then the accuracy of the 4 features provided to the expert will be low. As in Treatment 4, we randomly choose the features and break down this treatment into 5 sub-conditions.

For treatments 4 and 5 above, we create additional sub-conditions. We randomly select four features to disclose to the experts from the entire list of features used in the classification model. The list of disclosed features can include 0, 1, 2, 3, or 4 of the top 4 features predicting the outcome. The top 4 features are obtained by enumerating over all possible combinations of features to determine which set of 4 features contribute the most to each individual prediction. To this end, we take each set of 4 features and an instance, hold the values of these features unchanged, and perturb the remaining feature values slightly (by adding a small amount of Gaussian noise) to see how much the predictive probability of the corresponding instance changes [12]. Smaller changes in the predictive probability indicate that the set of 4 features chosen contribute the most to the prediction of that instance. After determining the top 4 features in this manner, we then consider the following sub-conditions:

A *4/4 top features*: Four of the features listed are among the top 4 predictors;

B *3/4 top features*: Three of the features listed are among the top 4 predictors;

C *2/4 top features*: Two of the features listed are among the top 4 predictors;

D *1/4 top features*: One of the features listed is among the top 4 predictors;

E *0/4 top features*: None of the features listed is among the top 4 predictors.

For the medical context, we have 28 doctors in treatments 1, 2, 3, 4.A, and 5.A, and 16 doctors in 4.B through 4.E and 5.B through 5.E. In the recruiting context, we have 32 recruiters in treatments 1, 2, 3, 4.A, and 5.A, and 16 recruiters in 4.B through 4.E and 5.B through 5.E.

In every treatment, each expert is given 10 randomly drawn people (patients for doctors and job candidates for recruiters) for whom they have to make decisions. Thus in total, we have 204 doctors spread across the various treatments and 2,040 medical questions. For recruiters, we have 224 experts and 2,240 hiring questions.

We are interested in two relevant outcomes. The first and most obvious outcome is decision accuracy, defined as the number of correct decisions out of all possible decisions. The second is decision time, i.e., how long the expert takes to process the information given and make their decision.

3 Results

In this section we describe our results. Figure 1 plots the results for doctors. Figure 1a presents the results for the five main treatments 1, 2, 3, 4.A, and 5.A. On the left-hand side we plot accuracy. Progressively adding explanations about the model recommendation helps doctors trust the recommendation more, which increases decision accuracy from 78% at baseline to 95%. This is a sizable increase. It is also worth noting that in the last two treatments, the doctor's judgement combined with the algorithm's prediction and explanations perform better than the expert (78% accuracy) or the algorithm (89% accuracy) on their own, suggesting valuable complementarities in the information used by the expert and the algorithm.

Providing information from a prediction algorithm lengthens the time doctors take to make their diagnostic decisions. On the right-hand side of Figure 1a we plot the average time, in minutes, that doctors take to answer each question. Every time doctors are provided with algorithmic recommendations, they take at least 30% longer to reach a decision. The scenarios where doctors take the longest time are those where doctors are given only the prediction (treatment 2) or the prediction with the complete set of explanations (treatment 5.A). In those two treatments, the time to reach a decision almost doubles, from 2 minutes and 10 seconds to around 4 minutes. This result can be explained by the fact that receiving a simple prediction leads doctors to reflect over whether they should trust the algorithm or not. Adding explanations increases trust, which speeds up decision

making, but also progressively increases the cognitive load needed to process additional information, which slows down decision making.

Next we move to discussing the results in Figure 1b, which focus on treatments 4 and 5 and the related sub-conditions.[1] We start with the effect of providing 4 features on decision accuracy (left-hand side). Whenever the list of features coincides with the 4 most relevant features for the prediction, doctors seem capable of discerning whether to rely on the algorithmic predictions or their own prediction such that in the aggregate they perform better than the algorithm (93% accuracy). However, doctors seem unable to distinguish between more and less relevant features. As the list of features includes fewer and fewer of the 4 most relevant features for the prediction (i.e., moving from sub-conditions A to E), decision accuracy rapidly degrades, from 93% to 63%, significantly lower than the accuracy of doctors' independent choices. This is worrisome, because it suggests that the mere list of features leads doctors to trust the algorithm more, regardless of the relevance of those feature. Decision time remains constant across all $4.A$ through $4.E$ treatments, suggesting that doctors are not faster or slower as a function of the relevance of the features disclosed.

It is comforting to see that doctors seem to adjust their over-reliance on the algorithmic prediction when they are given a measure of feature relevance (treatments $5.A$ through $5.E$). A pairwise comparison of the green and yellow bars on the left hand side of Figure 1b shows that adding feature accuracy allows doctors to make better decisions for any set of features, whether or not they are relevant. The improvement is particularly sizable when none of the features provided are among the 4 most relevant features (sub-condition E), which suggests that doctors' beliefs about feature relevance are biased upward. The improvement is also large enough to surpass the baseline accuracy of doctors when they make independent decisions.

Figure 2 reproduces analogous results for recruiters. The majority of the results are consistent across doctors and recruiters, so here we only highlight the differences.

The first difference is on the time it takes recruiters to reach a decision as they are given progressively more explanations about the algorithm process (right-hand-side plot of Figure 2a). For doctors, the time first increases between treatment 1 and 2, then decreases as we add the first explanation, and finally monotonically increases as we add increasingly more explanations. For recruiters instead, the time is monotonically increasing from treatment 1 to treatment $5.E$. The increase in time is also smaller than the increase for doctors, at least in percent terms.

The second difference is in the role that feature accuracy plays in guiding recruiters (left-hand-side plot of Figure 2b). Just like doctors, recruiters over-rely on algorithms when they are given a list of features. This leads to poor decision accuracy when those features are irrelevant in the model. Unlike doctors however, recruiters are unable to calibrate their reliance on the algorithm when given information about feature accuracy (the green and yellow bars are pair-wise indistinguishable from one another).

[1] The results for Treatment 5 and its sub-conditions will be completed.

4 Conclusion

In contexts where experts make decisions, adoption of algorithmic predictions has been slow. In this paper, we show that explaining machine learning algorithms can be an effective way to convince experts to rely on algorithmic recommendations and improve decision accuracy, but these nudges can backfire.

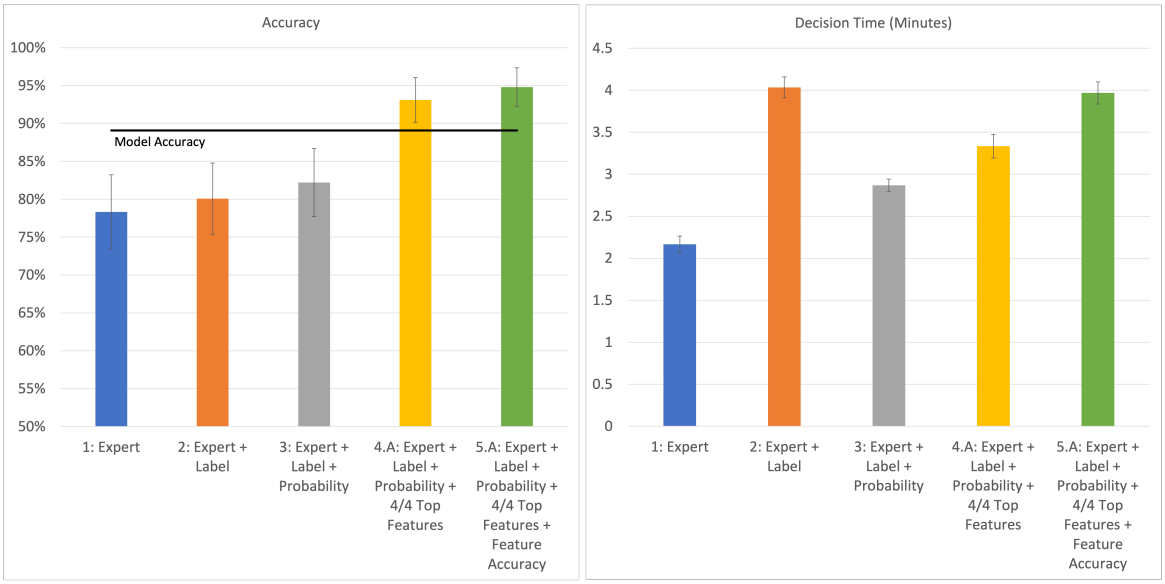
We give hypothetical decisions to doctors and recruiters about the likelihood that patients will develop breast cancer and the likelihood that job candidates will perform well in an entry-level software engineering job. We vary the availability of an algorithmic recommendation and a series of explanations about how the model reaches a particular recommendation. Experts' trust in algorithmic recommendations increases as more information about the black-box is provided to them. With the most complete set of explanations, experts are able to discern when to rely on the algorithm and when to overrule it, such that the combined accuracy is higher than the accuracy of the expert or the machine alone. Adding explanations makes experts slower in reaching a decision, compared to the scenario where they are not given any input from an algorithm, providing a valuable trade-off when decision time is as essential as accuracy.

However, providing explanations can lead experts to rely on algorithmic predictions more than necessary, which dramatically reduces the accuracy of the supported decision. Indeed, when given explanations, experts rely on algorithms regardless of the validity of those explanations. As a result, when explanations are not relevant, experts end up making decisions that are less accurate than without any algorithmic support. In the medical context, adding information about the relevance of the explanation can adjust over-reliance, although in the hiring context explanation relevance does not seem to help recruiters discern between better and worse explanations.

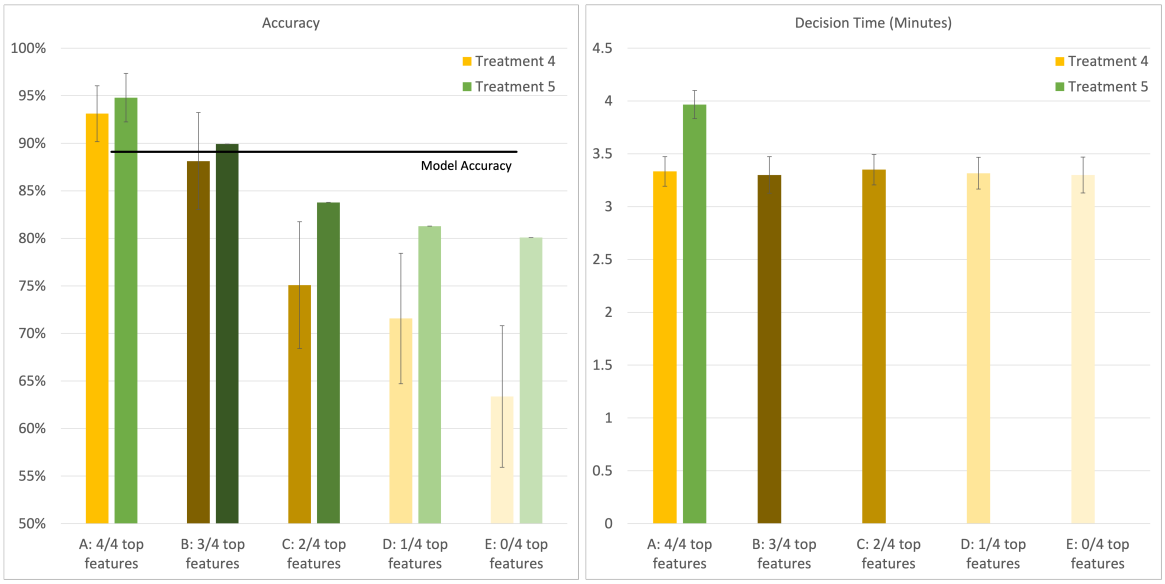
Our results have implications for the deployment of machine learning algorithms into settings where experts are the main decision makers. Explanations play an important role in convincing experts to trust recommendations, but if the baseline algorithm is inaccurate or many features are equally relevant in driving prediction, our results suggest that we cannot fully rely on experts to effectively discern what is and what is not relevant information.

Finally, we show that both accuracy and explainability are important in guiding experts' reliance on algorithms, which creates an important trade-off when deploying algorithms. On one end, experts are persuaded by highly accurate algorithms, which calls for sophisticated neural nets and black-box type of techniques. On the other hand, experts are also persuaded by explanations of the factors contributing to the prediction, which calls for simpler models, such as decision trees or regression techniques. We are exploring this trade-off directly in follow-up surveys.

Figure 1: Doctors



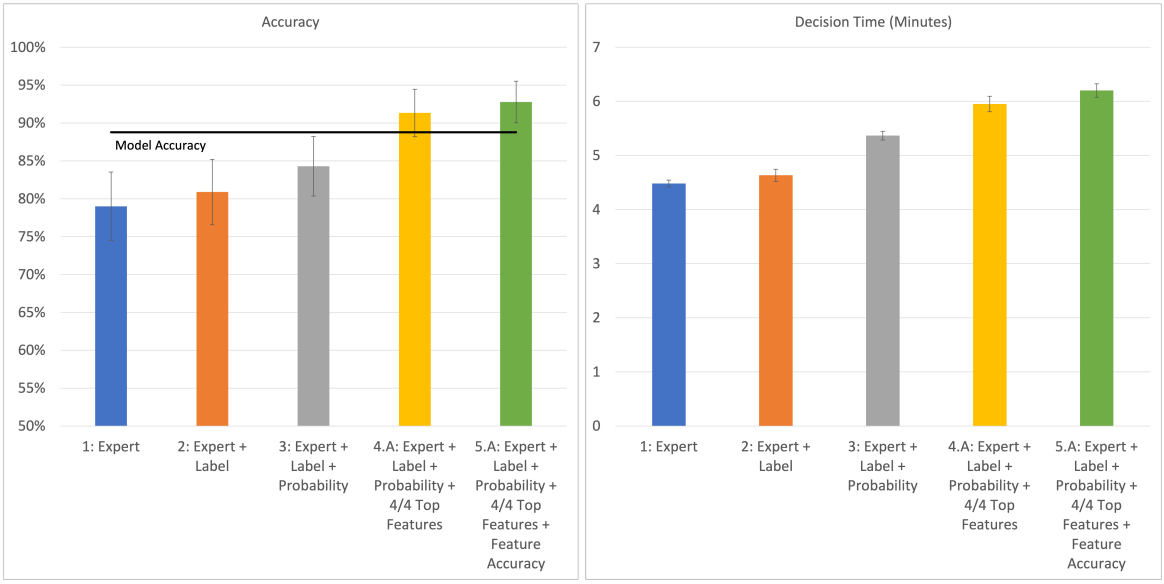
(a) Effect of Adding Explanations



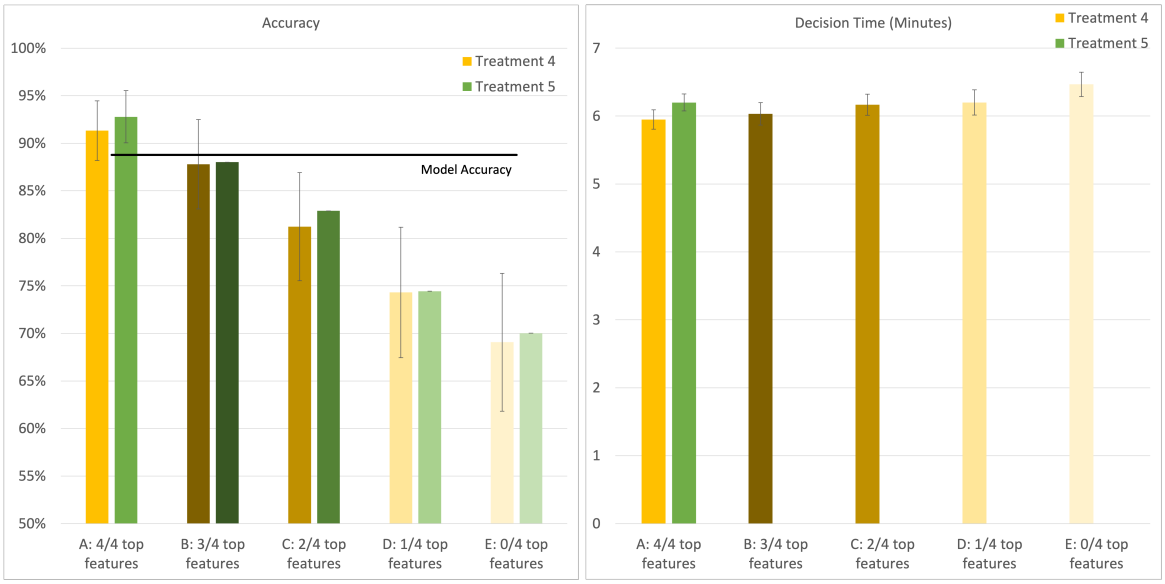
(b) Effect of Better Explanations

The figures focus on doctors. Panel a plots the results across treatments 1, 2, 3, 4.*A*, *and* 5.*A. Panel b plots results across sub-conditions A through E for treatments* 4 *and* 5. *On the left column we plot accuracy, while on the right column we plot the decision time in minutes. 95% confidence intervals are denoted by vertical lines. (Note: results for treatments* 5.*B through* 5.*E to be completed).*

Figure 2: Recruiters



(a) Effect of Adding Explanations



(b) Effect of Better Explanations

The figures focus on recruiters. Panel a plots the results across treatments 1, 2, 3, 4.*A*, *and* 5.*A*. *Panel b plots results across sub-conditions A through E for treatments* 4 *and* 5. *The plot is otherwise identical to Figure 1.*

References

[1] Daron Acemoglu and Pascual Restrepo. Artificial intelligence, automation, and work. In *The economics of artificial intelligence: An agenda*, pages 197–236. University of Chicago Press, 2018.

[2] Daron Acemoglu and Pascual Restrepo. Robots and jobs: Evidence from us labor markets. *Journal of Political Economy*, 128(6):2188–2244, 2020.

[3] Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press, 2018.

[4] Ajay K Agrawal, Joshua S Gans, and Avi Goldfarb. Ai adoption and system-wide change. *NBER Working Paper No. 28811*, 2021.

[5] Timothy Bresnahan. *Artificial Intelligence Technologies and Aggregate Growth Prospects*, page 132–170. Cambridge University Press, 2021.

[6] Sean Cao, Wei Jiang, Junbo L Wang, and Baozhong Yang. From man vs. machine to man+ machine: The art and ai of stock analyses. Technical report, National Bureau of Economic Research, 2021.

[7] Sukwoong Choi, Namil Kim, Junsik Kim, and Hyo Kang. How does ai improve human decision-making? evidence from the ai-powered go program. *Evidence from the AI-Powered Go Program (July 26, 2021)*, 2021.

[8] Jörg Claussen, Christian Peukert, and Ananya Sen. The editor and the algorithm: Economic returns to data and externalities in online news. 2019.

[9] Bo Cowgill, Fabrizio Dell'Acqua, and Sandra Matz. The managerial effects of algorithmic fairness activism. In *AEA Papers and Proceedings*, volume 110, pages 85–90, 2020.

[10] Avi Goldfarb, Bledi Taska, and Florenta Teodoridis. Could machine learning be a general purpose technology? a comparison of emerging technologies using data from online job postings. Technical report, National Bureau of Economic Research, 2022.

[11] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

[12] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018.

[13] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.