Investigations of Bias and Performance in Human-AI Teamwork in Hiring

Andi Peng,[1] Besmira Nushi,[2] Emre Kiciman,[2] Kori Inkpen,[2] Ece Kamar[2]
[1]Massachusetts Institute of Technology, [2]Microsoft Research

ABSTRACT

In AI-assisted decision-making, effective hybrid (human-AI) teamwork is not solely dependent on AI performance alone, but also on its impact on human decision-making. While prior work studies the effects of model accuracy on humans, we endeavour here to investigate the complex dynamics of how both a model's predictive performance and bias may transfer to humans in a recommendation-aided decision task.

We consider the domain of ML-assisted hiring, where humans—operating in a constrained selection setting—can choose whether they wish to utilize a trained model's inferences to help select candidates from written biographies. We conduct a large-scale user study leveraging a re-created dataset of real bios from prior work, where humans predict the ground truth occupation of given candidates with and without the help of three different NLP classifiers (random, bag-of-words, and deep neural network).

Our results demonstrate that while high-performance models significantly im- prove human performance in a hybrid setting, some models mitigate hybrid bias while others accentuate it. We examine these findings through the lens of decision conformity and observe that our model architecture choices have an impact on human-AI conformity and bias, motivating the explicit need to assess these complex dynamics prior to deployment.

CONTACT

Andi Peng
MIT CSAIL
andipeng@mit.edu
440-715-0384

INTRODUCTION

As **AI-powered decision tools** are increasingly deployed in real-world domains, a central challenge remains understanding how best to design models to **assist** humans. We investigate the question of how an AI-aided decision tool impacts both human **bias** and **accuracy** on a collaborative hiring task.

We make the following contributions:

1. To our knowledge, we present the first-ever experiment studying the propagation of both algorithmic performance and bias to human decision-making.
2. Our results reveal surprising findings, demonstrating that some models mitigate bias while others propagate and increase bias (even though original human and model biases span different regions). We interpret these results from a human-AI conformity lens and observe that high predictive performance from some model types do not necessarily increase human-model conformity, resulting in lower hybrid performance but less biased decisions.
3. We introduce our full crowdsourced data, comprised of 38,400 individual human judgements over 9,600 prediction tasks, as *Hybrid Hiring*: a first-ever large-scale dataset for studying human-AI collaborative decision- making trained, collected, and evaluated on real data.

Experimental Setup

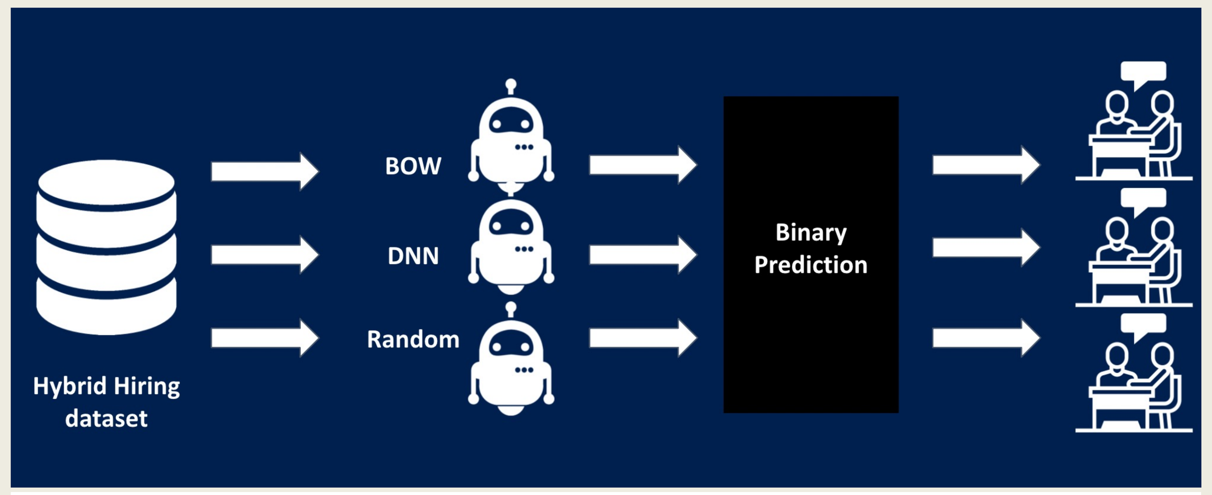


Figure 1: An example hybrid hiring workflow. A candidate dataset is used to train three NLP classifiers, which each outputs recommendations to human decision-makers.

We evaluate accuracy and bias of the resulting hybrid (human+AI) system.

RESULTS

We conduct a crowdsourced study across three conditions (model-only, human-only, and hybrid (human+AI) and evaluate:
Predictive performance (true positive rate (TPR))
1. Bias(differential TPR in classifying female vs. male candidates (ΔTPR, or TPR_f - TPR_m))

	Human	Rand	H+R	DNN	H+DNN	BOW	H+BOW
attorney	0.60	0.51^β	0.57	0.79^α	0.66^α	0.78^α	0.70^α
paralegal	0.60	0.49^β	0.56	0.87^α	0.68^α	0.78^α	0.70^α
physician	0.52	0.49^β	0.52	0.85^α	0.61^α	0.85^α	0.66^α
surgeon	0.61	0.51^β	0.61	0.89^α	0.68^α	0.82^α	0.74^α
professor	0.59	0.51^β	0.59	0.85^α	0.70^α	0.87^α	0.75^α
teacher	0.53	0.50^β	0.54	0.86^α	0.61^α	0.87^α	0.74^α

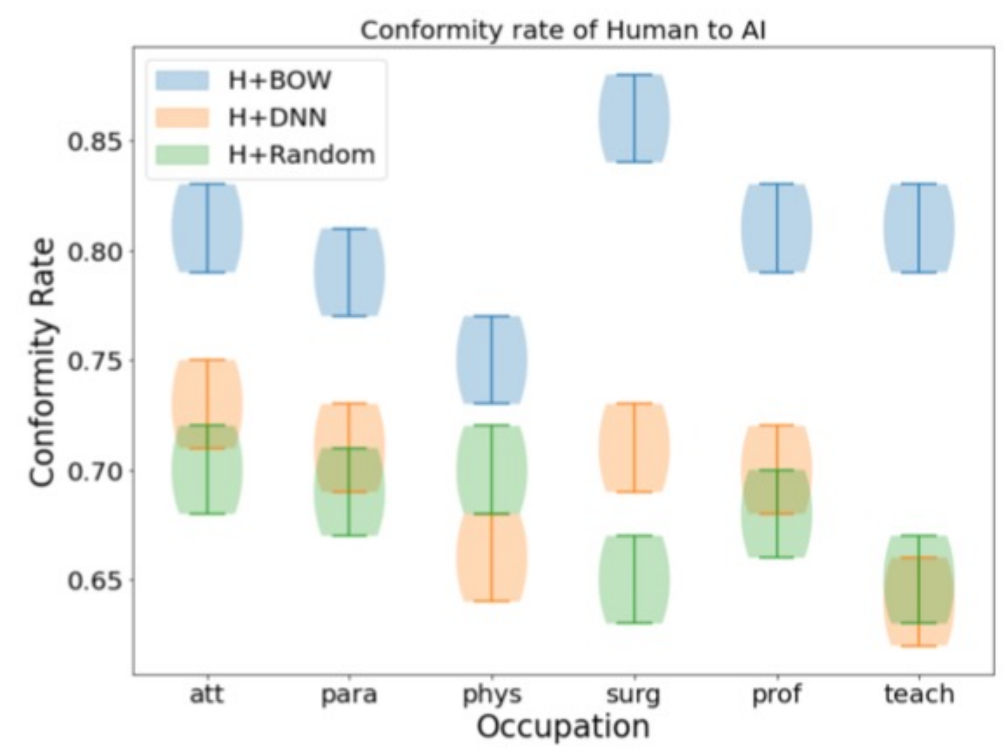
$^\alpha$ Greater than the Human condition, significant at $p < 0.01$. Also in yellow.
$^\beta$ Less than the Human condition, significant at $p < 0.01$. Also in green.

Table 1: *TPR* (predictive performance) on the same candidate slates across conditions. Pairwise comparisons are made between the human (base condition) and each corresponding model. Higher TPR models (DNN and BOW) consistently translate into higher TPR hybrid systems (H+DNN and H+BOW) whereas a lower TPR model (Random) does not impede performance (H+R).

	Human	Rand	H+R	DNN	H+DNN	BOW	H+BOW
attorney	-0.02	-0.04	-0.02	-0.04	-0.03	-0.06	-0.03
paralegal	0.09*	0.03	0.07	0.11*	0.03	0.23*	0.15*
physician	-0.02	0.02	-0.00	0.09*	-0.00	0.05	0.06
surgeon	-0.06	-0.04	-0.13*	-0.07*	-0.03	-0.16*	-0.16*
professor	0.02	0.04	0.00	-0.04	-0.03	-0.06	-0.03
teacher	0.10*	-0.03	0.03	0.03	0.02	0.04	0.07

* $TPR_f \neq TPR_m$, significant at $p < 0.01$. Also in pink.

Table 2: *Bias* (ΔTPR) across conditions for tested occupations. Within each slate, we conduct a pairwise comparison between TPR_f and TPR_m to see whether a significant difference is present. If so, that condition exhibits a significant ΔTPR.



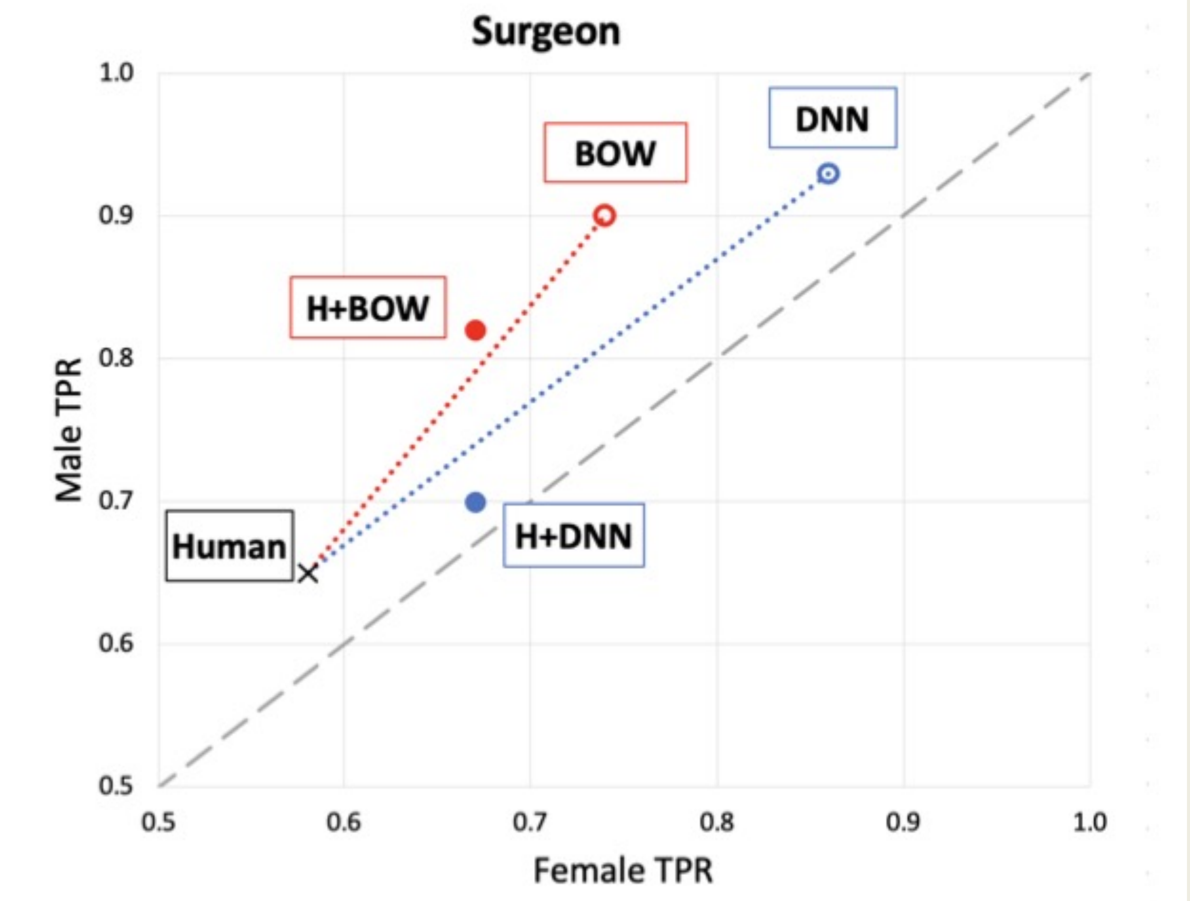


Figure 4: A visual of bias within the surgeon task, plotted again female (x-axis) and male (y-axis) TPRs. The center (grey) line represents an unbiased model. The bottom left represents a less accurate model, and the top right more accurate. Interpolation (dotted) lines are drawn to represent the expected trendline if no consistent difference across hybrid conditions existed. We see that DNN helps mitigate human bias (the resulting hybrid ΔTPR is close to the unbiased line) whereas BOW appears to *induce bias* (resulting in a hybrid ΔTPR farther from the line).

Discussion

Impact on Model Deployment
Our work calls into light critical concerns and trade-offs that need to be investigated prior to deploying similar models in practice in the world, particularly since results revealed significant differences in model conformity, even *without an interface change*.

Dataset Release
We introduce our full data as `Hybrid Hiring`, a large-scale dataset for studying human-AI decision-making that is collected and evaluated on real-world data. Comprised of 38,400 human judgements over 9,600 prediction tasks across seven conditions, our dataset represents a first of its kind released to study human decision-making in the loop with trained inferences.