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

Andi Peng, [1*] **Besmira Nushi,** [2] **Emre Kıcıman,** [2] **Kori Inkpen,** [2] **Ece Kamar** [2]

[1] Massachusetts Institute of Technology
[2] Microsoft Research
andipeng@mit.edu, {benushi, emrek, kori, eckamar}@microsoft.com

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 improve 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.

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 (Kleinberg et al. 2018). Ergo, a growing body of literature has arisen to study these *screening* or *recommendation* systems (Kleinberg et al. 2019), where a ML model acts as a data filtering mechanism to provide inferences as recommendations for a human decision-maker (Gillies et al. 2016). These collaborative settings call for a different evaluation process prior. If the model were to operate alone, the typical evaluation pipeline would involve measuring and reporting various predictive performance metrics (i.e. *how accurate is the model in solving the task?*), as well as checks for potential biases that may favor or disfavor groups based on sensitive attributes such as gender, age, or ethnicity (i.e. *does the model exhibit lower predictive performance for a given group?*)

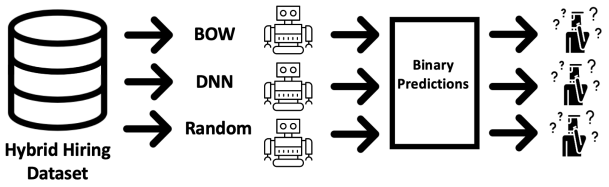


Figure 1: An example hybrid hiring workflow. A candidate dataset is used to train three NLP classifiers, which each outputs recommendations to human decision-makers. We evaluate accuracy and bias of the resulting system.

(Mehrabi et al. 2021; Barocas, Hardt, and Narayanan 2017). Both axes (*performance* and *bias*) are important for real-world deployment and exhibit different social implications in practice (Barocas, Hardt, and Narayanan 2017).

If the AI is instead intended to *assist* the human rather than act as sole arbiter, then assessing resulting performance involves understanding the interaction between human and machine. When a human makes a decision with the help of an AI recommendation, they can either bring in their own perspectives in choosing how to utilize the model or may choose to solve the task alone. Thus, *hybrid* (human-AI) performance depends on how the model alters the human decision, requiring an evaluation of a different nature that looks at how humans choose to conform to specific models.

Previous work has taken this approach in investigating how model accuracy transfers to hybrid accuracy (Lai and Tan 2019; Bansal et al. 2019a; Green and Chen 2019; Feng and Boyd-Graber 2019), illustrating that although hybrid systems designed for collaboration can improve accuracy beyond that of the human or AI alone, high model accuracy does not always transfer into high hybrid accuracy (Yin, Wortman Vaughan, and Wallach 2019). However, despite this increasing focus on human-AI collaboration, the way predictive bias inherent in ML models transfer to human decisions is not well understood at all. Specifically, it is not clear how biases from different model architectures would influence human bias or whether a more biased model would ultimately propagate to a human decision-maker at a higher rate than a less biased one like in the case of accuracy. The

[*] Work done while an AI Resident at Microsoft Research.

two in combination (predictive performance and bias) result in complex dynamics that may alter how both percolate down to a human decision-maker.

In this work, we investigate this by conducting a large-scale study to assess how a realistic hybrid system performs on both overall accuracy and bias (difference in predicting male vs. female candidates). We choose the domain of hiring due to a rich literature of human and algorithmic biases documented, with the question at play being: "Do I think this candidate is a good fit for this job?" Our human study leverages a large-scale text dataset (De-Arteaga et al. 2019) consisting of real candidate bios and employs three different NLP classifiers as assistance in predicting occupation from bio. We test how these models perform in isolation vs. when utilized as recommendations by humans in a hybrid system. To minimize side effects from other system properties (e.g., UX experience, confidence, etc.) we keep the interface presentation *unchanged* in all conditions and display only the final model recommendation as an aid. Figure 1 illustrates our hybrid experimental setup.

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 of our deployed models mitigate hybrid 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.

The above contributions provide important insights previously under-studied in both human-AI collaboration and algorithmic fairness literatures, and raise critical concerns and trade-offs that need to be investigated prior to deploying similar models in practice, particularly since our work revealed significant differences in model conformity, even *without an interface change*. Inspired by these results, we propose future directions in studying the impact of different ML models in hybrid decision-making scenarios.

Related Work

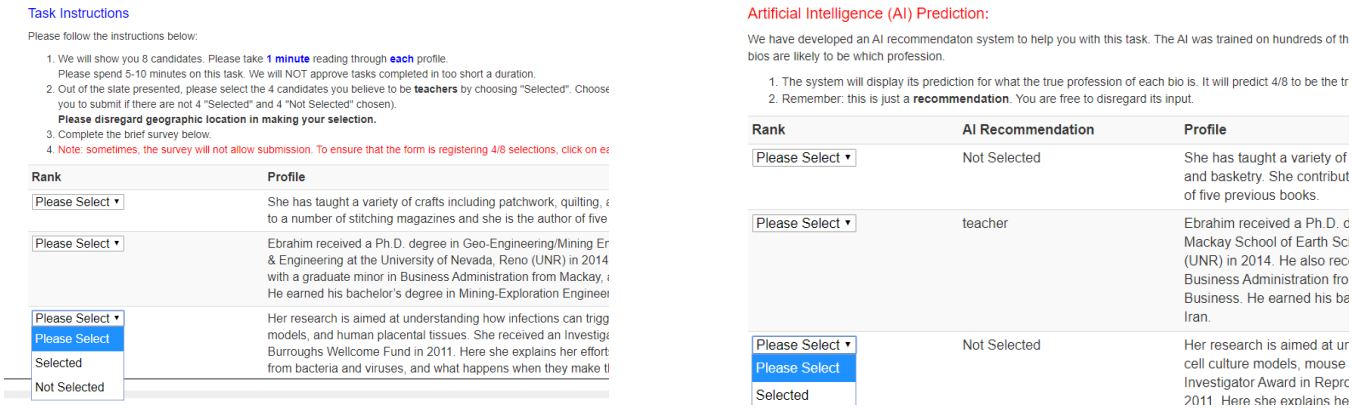
Algorithmic Bias It is now more important than ever to quantify and understand model biases that reinforce the disadvantaged status of different groups (Nosek and Banaji 2002; Sweeney 2013). While ML achieves higher-still accuracy, a key question becomes: *accurate, but for whom* (Barocas and Selbst 2016)? Hiring, long a discriminatory practice (Isaac, Lee, and Carnes 2009), has received specific renewed interest due to a rise in automated decision systems deployed with alarmingly detrimental effects towards female candidates (The Guardian 2018; Raghavan et al. 2020).

Spurred by such concerns, the ML community has responded with a rapidly growing body of literature on algorithmic fairness efforts. A brief overview ranges from approaches that seek to mitigate bias using techniques that are "unaware" of protected attributes like race and gender (Dwork et al. 2012) to more sophisticated techniques that seek to impose fairness as a "constraint" (Hardt, Price, and Srebro 2016). In practice, any method that relies on protected attributes for model training stands at odds with anti-discrimination law, which forbids the usage of these features in model prediction, even if the purpose is to mitigate bias (Dwork and Ilvento 2018; Gonen and Goldberg 2019).

Human Bias Complex decision tasks, limited cognitive resources, uncertain information, and a human tendency to aspire to reduce overall decision load together lead to a *bounded rationality* model of human decision-making, where cognitive biases come into play (Simon 1955; Cunningham 2013; Kahneman 2003). These biases are best described as heuristics, or mental shortcuts, that humans take when evaluating large amounts of uncertain information in a messy world (Thaler and Sunstein 2008). One particular form of bias that has been found to be especially detrimental is that of gender bias, particularly when evaluating candidates in professional settings. There is evidence that gender inequalities in the workplace stem, at least in part, from biased attitudes directed against women from those who hold sexist or innate preferences for a particular gender in different professions (Koch, D'Mello, and Sackett 2015). For instance, a study found that the higher a participant scored on a hostile sexism personality test, the more likely they were to recommend a male candidate rather than female for a managerial position (Masser and Abrams 2004).

Human-AI Collaboration The concept that decision processes adapt over time to adjust to changing preferences has led to *preference construction*, or decision-makers formalizing which option they prefer (Lichtenstein and Slovic 2006; Thaler and Sunstein 2008). It is of no surprise that systems designed to produce recommendations in key stages of decision-making have been found to have immense impact on final outcomes (Mandl et al. 2011). In these cases, the human makes a decision to either *accept* or *reject* recommendations. These AI-assisted systems have led to more accurate medical diagnoses (Lundberg et al. 2018), optimized crowdsourcing efforts (Kamar, Hacker, and Horvitz 2012), and creative multiagent game-playing (Jadeberg et al. 2019). Here, we refer to human-AI together as a *hybrid* system.

As hybrid systems are increasingly deployed, it is important to understand their impact on human decision processes. Many factors, such as the human's ability to create a mental picture of the model (Bansal et al. 2019a), their implicit trust in the model (Yin, Wortman Vaughan, and Wallach 2019; Zhang, Liao, and Bellamy 2020), how they are impacted by updates (Bansal et al. 2019b), the representational display of recommendations (Peng et al. 2019), and the interpretability of the model (Gilpin et al. 2018) have all been demonstrated to greatly impact humans. However, to our knowledge, there



(a) Human-only condition.

(b) Hybrid condition.

Figure 2: An example task where the true occupation is *teacher* and confused occupation *professor*. The *interface remains unchanged* across all candidate slates and conditions. Additionally, recommendations do not provide any additional evidence or signal of the underlying model behaviour (e.g. confidence, architecture, explanation for the decision, etc.).

exists no work that studies how both AI predictive performance and bias transfer to humans.

Experimental Setup

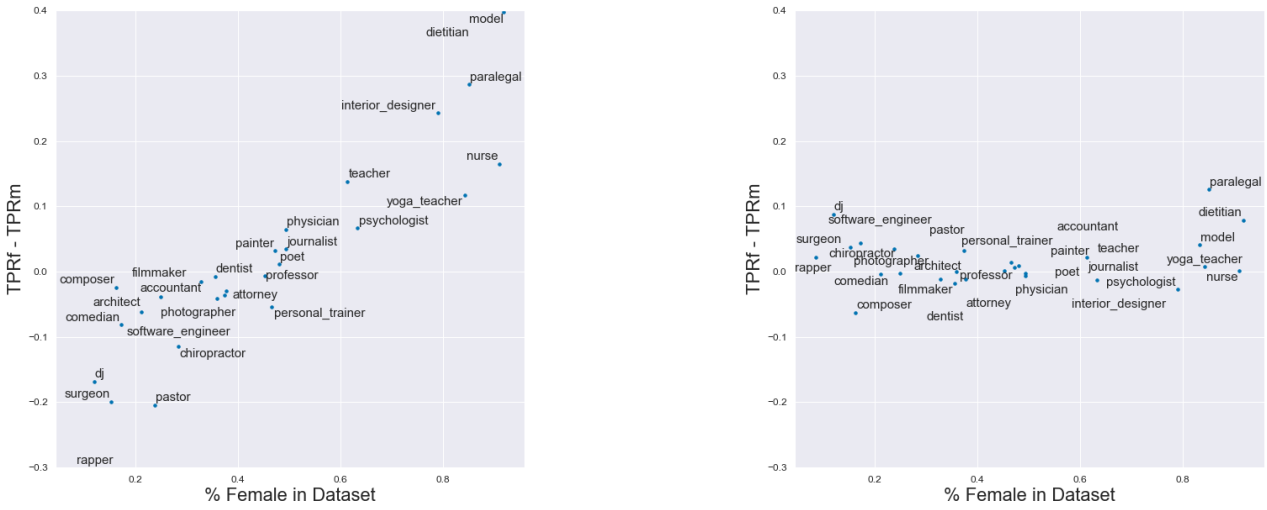
Motivation Our work is motivated by the desire to understand how bias in algorithmic models transfer to hybrid decision-making in realistic deployed settings where both users of trained models and their real-world stakeholders are impacted. Often, it is assumed that a higher-performing model will help a human make more-accurate and less-biased decisions, or conversely, that a human will recognize model mistakes and exert agency in correcting them. Yet, we have very little understanding of how these metrics trickle down through a hybrid decision pipeline. In this work, we evaluate how different models trained on real-world data, when integrated within common hiring pipeline under constraints, alter final system predictive performance and bias. Studying this allows us to better understand the impact of this increasingly-common workflow as well as unearth which types of algorithmic advancements can actually be transferred to a human-in-the-loop system.

Data Collection We select the task of language-based *occupation classification* due to its direct relevance to real-world hiring scenarios (Peng et al. 2019). To a human, predicting an individual's true occupation from a brief text description remains a common and often high-stakes decision made in professional settings daily. We compile a corpus of public professional bios using the same methodology as De-Arteaga et al. by scraping online bios using the Common Crawl to re-create a dataset where all observations begin with the following sequence: [*name* is a *title*] and subsequently describe a professional background (De-Arteaga et al. 2019). We extract the ground truth occupation and gender of each observation and to the best of our ability, mask out names. We select the 28 most frequently-occurring occupations, resulting in 397,907 observations of which *professor* is the most frequent occupation and *rapper* the least. This dataset represents a publicly-available online pool of candidates that may be screened by a real model.

Model Training The objective is to, without access to the first sentence of a bio which identifies occupation, predict the ground truth using the candidate's self-provided description. To isolate the impact of model architecture on hybrid performance, we elect to train a single-layer fully-connected *deep neural network* (DNN) as well as a simpler *bag of words* (BOW) (De-Arteaga et al. 2019; Bolukbasi et al. 2016). For our BOW, we use a one-versus-all logistic regression with L_2 regularization similar with prior work (De-Arteaga et al. 2019; Romanov et al. 2019). DNN represents a more *black-box* architecture due to its non-linear nature and deeply nested structures whereas BOW remains a good baseline due to its general *interpretability* (Gilpin et al. 2018).

Because some occupations exhibit an uneven skew of either male or female bios and we wish to de-link existing data pipeline biases from our analysis, we create validation and test splits such that both gender and occupation are sufficiently represented. In accordance with prior work (De-Arteaga et al. 2019; Romanov et al. 2019), we use stratified-by-occupation splits, with 65% of the bios (258,639) designated for training, 10% (39,790 bios) designated for validation, and 25% (99,476 bios) designated for testing. This isolates the differences in model performance to their varying architectures, and allows for an equivalent apples-to-apples comparison on resulting hybrid performance and bias.

Human Task Design We construct a constrained decision task by presenting 8 bios, 4 of which belong to the occupation of interest and ask humans to identify the correct 4 out of the 8 that belongs to that occupation. We are in effect simulating a realistic scenario where, say, a recruiter operating under resource constraints is tasked with selecting a subset of candidates for interviewing and may make implicit judgements based on gender (The Guardian 2018). To ensure that our slates are non-trivially difficult for humans, we generate confusion matrices for predictions made by our



(a) BOW classification bias.

(b) DNN classification bias.

Figure 3: DNN and BOW gender bias on the dataset test split as quantified by TPR gender gap (ΔTPR) relative to true proportion of female candidates in the dataset. While both models exhibit biases, DNN's ΔTPRs across occupations do not appear as extreme as BOW's. Note that our candidate slates are generated from bios sampled from this distribution.

models and select the following 3 pairs of highly-confused professions by gender: *attorney* and *paralegal*, *surgeon* and *physician*, and *professor* and *teacher*. Then, to assess the potentially bi-directional nature of bias (for example, a female lawyer being misclassified as a paralegal implies something very different than a male paralegal being misclassified as a lawyer), we create 6 tasks from these 3 occupation pairs (i.e. one type of slate is an attorney misclassified as a paralegal and its counterpart a paralegal misclassified as an attorney).

For each occupation, we design candidate slates where 8 bios are randomly selected from our test split (4 from the true occupation and 4 from the confused occupation), with the additional constraint that gender representation remain equal in both. This is done to enforce the opportunity to select equal subsets of "qualified" candidates, irrespective of how they are actually represented in the world. Altogether, we generate 200 unique slates, randomly ordered, for each occupation to total 9,600 samples from our original dataset (6*200*8 = 9,600 bios total to be classified by each control group).

Evaluation To study the impact of AI recommendations on human decision-making, we conduct a crowdsourced study across three conditions (model-only, human-only, and hybrid) and evaluate the following two metrics:

1. Predictive performance (true positive rate (TPR))
2. Bias (differential TPR in classifying female vs. male candidates (ΔTPR, or TPR_f - TPR_m))

Note, these two axes are not the same: a system may classify candidates successfully at a higher rate but also exhibit bias in being more accurate for male vs. female candidates. The ideal system is one that maximizes TPR without exhibiting significant ΔTPR. We report TPR rather than accuracy since we are studying constrained decision-making where the candidate slate size is fixed and if one classification is correct, this necessitates that another was incorrect. This helps our evaluation of bias (ΔTPR), which is calculated as the difference in TPRs between binary gender candidates of each occupation (De-Arteaga et al. 2019). A positive ΔTPR indicates a bias towards female candidates and negative ΔTPR) towards male. In line with previous work (Peng et al. 2019), we formulate the task as a filtering rather than a classification task, which allows for us to observe bias to a greater extent since a budget is allocated for selection and not all candidates can be prioritized (as is the case in real-world settings). A biased system will exhibit a statistically significant ΔTPR (i.e. $\text{TPR}_f \neq \text{TPR}_m$) across slates.

Model-Only Condition For each of our generated candidate slates, AI recommendations are created by selecting the top 4 bios that our trained DNN and BOW models have the highest confidence in their predictions as belonging to the ground truth occupation. This forces the same constrained decision task that our subsequent conditions will face. In addition, we also test a "random" model, which selects its 4 bios via coin flip to serve as a non-intelligent baseline. Because we are enforcing the same subset criteria on the exact same candidate slates, we can attribute any arising performance differences to model type and not the task itself.

Human-Only Condition For our human-only condition, we deploy slates as HIT tasks on mTurk (Figure 3). We show each participant a unique slate, present a description of the

Table 1: TPR on the same candidates slates across conditions. Pairwise comparisons are made between the human (base condition) and each corresponding model to assess the performance differential. 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	0.60	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.

ground truth occupation, and ask them to select 4/8 bios that they believe to best fit that description. We programmatically enforce that each participant picks the correct number of selections and each bio must be user-clicked as *Selected* or *Not Selected*. Bios are randomly ordered per slate to remove possible confounding factors such as rank ordering preference and recency bias (although final generated slates are kept consistent between conditions). Altogether, we deploy 1,600 uniquely-generated HITs across six tested occupations.

Hybrid Condition For our hybrid condition, we follow the same methodology as for our human-only condition but additionally provide predictions made by our three models. Participants are explicitly instructed that these predictions are "recommendations" from an "AI" that they may choose to disregard and override. For this condition, we deploy 4,800 unique HITs in total (1,600 each for human+DNN, human+BOW, and human+Random). Note: irrespective of the model tested, the interface remained the same and participants could not participate in HITs across conditions.

To increase reproducibility confidence, we run all 200 slates per occupation in two batches of 100 across unique study participant pools, each with a mix of human-only and hybrid conditions: the first between August 23-27, 2019 and the second between September 1-4, 2019. This is done to ensure that demographic skews in crowdsourcing may be mitigated across worker pools. We compensate all participants at a wage of $15 per hour. Participants are additionally screened according to the following qualifications: hold above a 95% approval rating, unique ID per condition, and based in the United States to control for English being the primary spoken language.

Data Ethics and Privacy For all experiments and collected data, we conduct both institutional IRB and data privacy review. We also anonymize all bios (by stripping out names and other identifying features) and participant data (we collect no no personal or private information).

Statistical Testing In evaluating significance across conditions, we are interested in seeing whether a condition (i.e. a specific model) produces changes in hybrid performance when compared to a baseline. We use the human-only condition as our baseline for all comparisons since we are interested in studying the impacts of AI on humans in this work. We utilize Friedman and Wilcoxon signed ranks tests to study the effect of each candidate slate across conditions in pairwise comparisons to the human-only (base) condition.

Results

First, we examine performance of our model-only condition. We see that different models exhibit different TPRs and biases, with BOW and DNN architectures indeed making varied selections on the same task. Second, we turn to the human-only condition and find that humans exhibit their own set of biases that do not parallel either trained model. Third, we assess the impact of recommendations on human decision-making in our hybrid condition and find that although a higher-TPR model consistently produces higher-TPR hybrid teamwork, the impact on bias is model-specific, with DNN mitigating human bias while BOW seemingly inducing it. Last, we assess these results through the lens of human-AI conformity and discover that high-TPR performance from our tested non-linear model does not necessarily increase human-model agreement, resulting in ultimately lower hybrid performance but less biased decisions.

Model-Only Performance Table 1 highlights the TPRs of human and model-only conditions. We see that DNN and BOW do not make identical predictions across candidate slates, with DNN generally outperforming BOW (as evidenced by the difference in TPRs, particularly on *paralegal* and *surgeon* tasks). To probe this further, we analyze the original classifications made by both models and find that, as shown in Figure 3, DNN and BOW exhibit different biases (ΔTPRs) across occupations. For example, BOW ΔTPR of *paralegals* (top right of Figure 3a) indicates both a true high proportion of female paralegals in the dataset as well as model bias in classifying them as such.

Human-Only Performance We next ask the question: do human predictions resemble that of either model? Across both TPR and ΔTPR evaluations, we find that human-only decisions do not overlap with those from either BOW or

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.

	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

* $\text{TPR}_f \neq \text{TPR}_m$, significant at $p < 0.01$. Also in pink.

DNN-only predictions at different rates, thus removing the possible confounder that one model aligned with original human decisions more than the other (details can be found in Appendix). Table 1 shows that the human-only condition significantly under-performs both DNN and BOW models on all occupation slates, although in most cases does perform higher than Random. Moreover, Table 2 illustrates different biases across different conditions, with DNN not exhibiting any significant bias across all occupations, BOW biased towards female *paralegals* and male *surgeons*, and humans biased towards female *paralegals* and *teachers*.

Model-Specific Impact On Hybrid TPR When assessing the impact of model TPR on hybrid decision-making, we find that human decision-makers collaborating with a higher TPR model (DNN and BOW) results in a consistently significant *improvement* across all occupations. This is in accordance with previous work, which has observed that higher-accuracy models generally help lower-accuracy humans (Bansal et al. 2019a, 2021), although this is still far from achieving optimal complementarity. Interestingly, when humans collaborate with a lower TPR model (Random), their own performance is *not impeded* (Table 1).

Model-Specific Impact On Hybrid ΔTPR A different story emerges when evaluating the impact of model ΔTPR on hybrid decision-making, with different models impacting resulting biases differently. When humans collaborate with DNN, the resulting system (irrespective of any human biases at play) becomes unbiased. Table 2 illustrates how the originally biased occupations of *paralegal* and *teacher* become both mitigated by an unbiased DNN. However, an opposite effect can be seen in humans collaborating with BOW, with the resulting system seemingly reflecting both human-only and model-only biases. For example, despite the original human being unbiased in the *surgeon* task, the resulting hybrid system is pulled towards a significant bias towards male candidates. Figure 4 analyzes this result in greater detail using the *surgeon* task as an illustration. Note that the key point is not only that the DNN-hybrid system is ultimately less biased than the BOW-hybrid (lower hybrid ΔTPR), but that the resulting system is pulled below the interpolated expected (blue) line between Human and DNN performance gains towards the fully unbiased (grey) line, whereas the BOW-hybrid is pulled above the interpolated (red) line to-

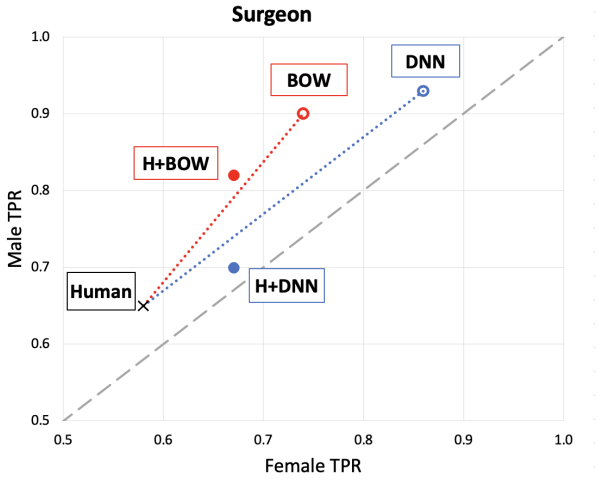


Figure 4: A visual of bias within the *surgeon* task, plotted again female (x-axis) and male (y-axis) TPRs. The center (grey) line represents an unbiased model. The bottom left represents a less accurate model, and the top right more accurate. Interpolation (dotted) lines are drawn to represent the expected 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 actually induce bias (resulting in a hybrid ΔTPR farther from the line).

wards a more biased direction. Visually, this helps differentiate between bias mitigation that may result from performance gains of a higher-performing model and highlights differences between how bias percolates differently from a DNN vs. BOW model down to a human.

Investigating Conformity Why do we see very different results for model-specific impacts on TPR vs. ΔTPR hybrid decision-making, *even without an interface change*? To better understand a sample-by-sample breakdown, we investigate human-AI conformity, or the rate at which a human appears to follow the model's recommendations in a hybrid system. We compute this by assessing the percentage of hybrid decisions that match those of original model decisions

Table 3: Hybrid decisions that match original model decisions, conditioned on the model being incorrect, i.e. *when does a human accept a wrong prediction?* Here, H+Random serves as a baseline for understanding the additional conformity to a specific architecture beyond blind acceptance of AI recommendations themselves. We observe that humans are significantly more likely to conform to incorrect BOW decisions relative to DNN, which rarely differs from Random.

	H+Random	H+DNN	H+BOW
attorney	0.622	0.663	0.744*
paralegal	0.629	0.634	0.716*
physician	0.673	0.648	0.782*
surgeon	0.561	0.645*	0.809*
professor	0.605	0.504	0.704*
teacher	0.606	0.623	0.804*

* Greater than H+Random when the model is incorrect, significant at $p < 0.01$. Also in blue.

Table 4: Prediction overlap between the human-only and model-only conditions, i.e. *what percentage of the original human decisions matched those of each model?* Although we see higher human overlap with DNN and BOW vs. Random (likely due to Random being a generally lower-performing model that operates by chance), there is no significant difference between DNN vs. BOW. This helps assuage concerns regarding one model resembling human reasoning more than another prior to deployment in the task.

	Random	DNN	BOW
attorney	0.511	0.589*	0.608*
paralegal	0.498	0.583*	0.570*
physician	0.501	0.510	0.526*
surgeon	0.485	0.599*	0.575*
professor	0.510	0.554*	0.570*
teacher	0.516	0.531*	0.526*

* Greater than Random, significant at $p < 0.01$. Also in green.

for each candidate slate (irrespective of whether that classification was the ground truth or not). Figure 5 illustrates that although we see similar conformity rates of the human to DNN, humans conform significantly more to BOW predictions than either DNN or Random. Moreover, this distinction is especially apparent in cases where the model made an incorrect prediction (Table 3). A possible explanation, supported by past work, posits that BOW is a generally more *interpretable* model that humans can understand (and trust) more (De-Arteaga et al. 2019). Because BOW word associations are learned by encoding sparse vectors that map to word vocabularies in a manner that is thought of to be more linear, humans are able to formulate an internal understanding of its recommendations more readily than DNN (a *black-box* non-linear model) or Random (complete chance) (Bansal et al. 2019a; Poursabzi-Sangdeh et al. 2021). In fact, based on Table 2 we observe that despite the lower hybrid performance of the Random model, random recommendations appear to have similar effects to the DNN on mitigating mitigating bias. As a result, humans may be more willing to accept the inferences provided by BOW (even when those recommendations are biased) and conform to its predictions, particularly when operating under resource constraints.

Discussion

Impact on Model Deployment A natural question that arises from these findings is whether DNN and Random (which both appear to be uninterpretable models) help mitigate human biases because they force human decision-makers to self-reflect more, and if so, whether ML deployment should actually prioritize this objective in future system design where minimizing bias may be a priority. To do so would mean an orthogonal departure from current work, where system designers are seeking less biased and more interpretable models. Moreover, our H+BOW was more accurate than our H+DNN, posing a trade-off between high team accuracy vs. low team bias. Our recommendation is that, while our results are somewhat surprising and highlight the importance of studying real-world hybrid decision-making, deploying a less interpretable model serves as a shortcut to true bias mitigation. As a community, we should seek to discover mechanisms that achieve this more explicitly and efficiently to truly leverage the complementary strengths of improved algorithmic design. Examples may include requiring humans to follow explicit forms of self-reflection and decision justification when there exists a a risk of bias.

Dataset Release We introduce our full experimental data as `Hybrid Hiring`, a large-scale dataset for studying human-AI decision-making that is collected and evaluated on real-world candidates. Comprised of 38,400 human judgements over 9,600 unique prediction tasks across seven conditions, our dataset represents a first of its kind released to study human decision-making in the loop utilizing trained ML inferences. Ideally, hiring (and other high-stakes social decisions) should always remain in the purview of human review, and so utilizing datasets and methodologies of this kind will allow the field to investigate the impacts of different research questions on human decision-making in these contexts. Although we specifically investigated hybrid performance of three NLP models, one can easily extend this work to alternate architectures and interfaces.

Limitations While we do our best to simulate a realistic hybrid task by selecting a socially relevant domain where real human data is incorporated in the decision-making of human study participants, we recognize that we are still running a controlled study on mTurk, where transfer of results to real-world deployed systems may be limited. Moreover, we greatly simplify many potential confounders (such as age, presence of non-binary gender, and self-written biography variance) in isolating bias to a single variable. We also do not study state-of-the-art de-biased models due to more complex architectures and leave for future work. We hope that our work moves the needle more in the direction of studying the impacts of ML-aided systems in real-world

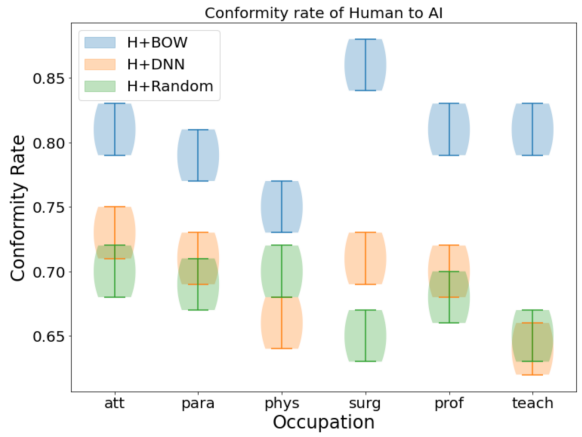


Figure 5: Conformity rate (percentage of hybrid decisions that match those predicted by the model alone) across tested occupations. We see significantly higher conformity to BOW than to DNN and Random predictions, with highlighted bands detailing 95% confidence intervals.

environments and propose that the community jointly invest in producing similar large-scale decision tasks and datasets to further study such intricacies across varied domains.

Conclusion

In asking the question of how model performance impacts human decision-making on two axes, our findings open up additional questions related to the specificity of human responses to different models, *even without an interface change*. Our results motivate the explicit need to further investigate the observed signals regarding differing human intuitions of varied model architectures and how we can best design systems that allow for optimal hybrid collaboration.

Acknowledgements

We would like to thank Adam Kalai and Maria De-Arteaga for helpful discussions on problem formulation, Alexey Romanov for help with data collection and model training, Sarah Jobalia for moral support, and the anonymous reviewers for comments on the draft. Andi Peng is supported by an NSF Graduate Research Fellowship.

References

Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W.; Weld, D.; and Horvitz, E. 2019a. Beyond accuracy: the role of mental models in human-AI team performance. In *Proceedings of the 2019 AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019)*. AAAI.

Bansal, G.; Nushi, B.; Kamar, E.; Weld, D.; and Horvitz, E. 2019b. Updates in human-AI teams: understanding and addressing the performance-compatibility tradeoff. In *Proceedings of the 2019 AAAI Conference on Artificial Intelligence (AAAI 2019)*. AAAI.

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 Conference on Human Factors in Computing Systems (CHI 2021)*, 1–16.

Barocas, S.; Hardt, M.; and Narayanan, A. 2017. Fairness in machine learning. *Nips tutorial*, 1: 2017.

Barocas, S.; and Selbst, A. 2016. Big data's disparate impact. *California Law Review*, 671.

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 2016 Conference on Neural Information Processing Systems (NIPS 2016)*. NIPS.

Cunningham, T. 2013. Biases and implicit knowledge. *Munich Personal RePEc Archive*.

De-Arteaga, M.; Romanov, A.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; and Kalai, A. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* 2019)*. ACM.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 2012 Innovations in Theoretical Computer Science Foundations Conference (ITCS 2012)*. ACM.

Dwork, C.; and Ilvento, C. 2018. Group fairness under composition. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT* 2018)*. ACM.

Feng, S.; and Boyd-Graber, J. 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*, IUI '19, 229–239. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362726.

Gillies, M.; Fiebrink, R.; Tanaka, A.; Caramiaux, B.; Garcia, J.; Bevilacqua, F.; Heloir, A.; Nunnari, F.; Mackay, W.; and Amershi, S. 2016. Human-centered machine learning. In *Proceedings of the 2016 Conference on Human Factors in Computing Systems (CHI 2016)*. ACM.

Gilpin, L.; Bau, D.; Yuan, B.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining explanations: an overview of interpretability in machine learning. In *Proceedings of the 2018 IEEE Conference on Data Science and Advanced Analytics (DSAA 2018)*. IEEE.

Gonen, H.; and Goldberg, Y. 2019. Lipstick on a pig: debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019)*. ACM.

Green, B.; and Chen, Y. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–24.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 2016 Conference on Neural Information Processing Systems (NeurIPS 2016)*. NeurIPS.

Isaac, C.; Lee, B.; and Carnes, M. 2009. Interventions that affect gender bias in hiring: a systematic review. *U.S. National Library of Medicine*, 84: 1440–1446.

Jadeberg, M.; Czarnecki, W.; Dunning, I.; Marris, L.; Lever, G.; Garcia Castaneda, A.; Beattie, C.; Rabinowitz, N.; Morcos, A.; Ruderman, A.; Sonnerat, N.; Green, T.; Deason, L.; Leibo, J.; Silver, D.; Hassabis, D.; Kavukcuoglu, K.; and Graepel, T. 2019. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 31.

Kahneman, D. 2003. A perspective on judgment and choice. *American Psychologist*, 58: 697—-720.

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 2012 International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*. IFAAMAS.

Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2018. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133.

Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Sunstein, C. R. 2019. Discrimination in the age of algorithms. *SSRN*.

Koch, A.; D'Mello, S.; and Sackett, P. 2015. A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Applied Psychology*, 100: 128–161.

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*, FAT* '19, 29–38. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Lichtenstein, S.; and Slovic, P. 2006. *The Construction of Preference*. Cambridge University Press.

Lundberg, S.; Nair, B.; Vavilala, M.; Horibe, M.; Eisses, M.; Adams, T.; Liston, D.; King-Wai Low, D.; Newman, S.-F.; Kim, J.; and Lee, S.-I. 2018. Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery. *Nature Biomedical Engineering*, 2.

Mandl, M.; Felfernig, A.; Teppan, E.; and Schubert, M. 2011. Consumer decision making in knowledge-based recommendation. *Journal of Intelligent Information Systems*, 37: 1–22.

Masser, B.; and Abrams, D. 2004. Reinforcing the glass ceiling: the consequences of hostile sexism for female managerial candidates. *Sex Roles*, 51: 609–615.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Nosek, B.; and Banaji, M. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1).

Peng, A.; Nushi, B.; Kiciman, E.; Inkpen, K.; Suri, S.; and Kamar, E. 2019. What you see is what you get? The Impact of Representation Criteria on Human Bias in Hiring. In *Proceedings of the 2019 AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019)*. AAAI.

Poursabzi-Sangdeh, F.; Goldstein, D.; Hofman, J.; Wortman Vaughn, J.; and Wallach, H. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 Conference on Human Factors in Computing Systems (CHI 2021)*.

Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* 2020)*, 469–481.

Romanov, A.; De-Arteaga, M.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; Rumshisky, A.; and Kalai, A. 2019. What's in a name? Reducing bias in bios without access to protected attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019)*. NAACL.

Simon, H. 1955. A Behavioral Model of Choice. *Quarterly Journal of Economics*, 69: 99–118.

Sweeney, L. 2013. Discrimination in online ad delivery. *Queue - Storage*, 11.

Thaler, R.; and Sunstein, C. 2008. *Nudge: improving decisions about health, wealth, and happiness*. Penguin Books.

The Guardian, N. 2018. Amazon ditched AI recruiting tool that favored men for technical jobs.

Yin, M.; Wortman Vaughan, J.; and Wallach, H. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 Conference on Human Factors in Computing Systems (CHI 2019)*. ACM.

Zhang, Y.; Liao, V.; and Bellamy, R. 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* 2020)*, 295–305.