

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

ANDI PENG\*, Massachusetts Institute of Technology, USA

BESMIRA NUSHI, Microsoft Research, USA

EMRE KICIMAN, Microsoft Research, USA

KORI INKPEN, Microsoft Research, USA

ECE KAMAR, Microsoft Research, USA

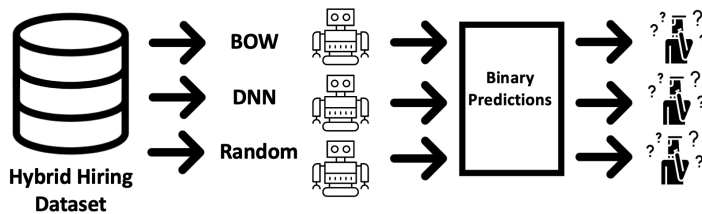


Fig. 1. An example hybrid hiring workflow. A candidate dataset is used to train three NLP classifiers, which each outputs recommendations to human decision-makers. We evaluate accuracy and bias of the resulting system.

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, 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 bios. 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 bias while others accentuate it. We examine these findings through the lens of decision conformity and observe that model architecture choices have an impact on human-AI conformity and bias, motivating the explicit need to assess these complex dynamics prior to deployment.

CCS Concepts: • **Human-centered computing** → *Collaborative and social computing*; • **Computing methodologies** → **Artificial intelligence**;

## ACM Reference Format:

Andi Peng\*, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. 2022. Investigations of Performance and Bias in Human-AI Teamwork in Hiring. In *CHI 2022: Workshop on Trust and Reliance in AI-Human Teams, April 30, 2022, New Orleans, LA*. ACM, New York, NY, USA, 12 pages. <https://doi.org/XXXXXXX.XXXXXXX>

---

Work done while an AI Resident at Microsoft Research.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2022 Association for Computing Machinery.

Manuscript submitted to ACM

## 1 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 [21]. Ergo, a growing body of literature has arisen to study these *screening* or *recommendation* systems [22], where a ML model acts as a data filtering mechanism to provide inferences as recommendations for a human decision-maker [12]. 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?*) [4, 29]. Both axes (*performance* and *bias*) are important for real-world deployment and exhibit different social implications in practice [4].

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 [1, 11, 15, 24], 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 [39]. 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 two in combination (predictive performance and bias) result in complex dynamics that may alter how both percolate down to a human decision-maker.

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 [8] 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 *unchanged* in all conditions and display only the final model recommendation as an aid. 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. 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.

## 2 RELATED WORK

**2.0.1 Algorithmic Bias.** It is now more important than ever to quantify and understand model biases that reinforce the disadvantaged status of different groups [30, 36]. While ML achieves higher-still accuracy, a key question becomes: *accurate, but for whom* [5]? Hiring, long a discriminatory practice [17], has received specific renewed interest due to a rise in automated decision systems deployed with alarmingly detrimental effects towards female candidates [33, 38].

Spurred by such concerns, the ML community has responded with a growing body of literature on algorithmic fairness. A brief overview ranges from approaches that seek to mitigate bias using techniques that are “unaware” of protected attributes like race and gender [9] to more sophisticated techniques that seek to impose fairness as a “constraint” [16]. 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 [10, 14].

**2.0.2 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 [7, 19, 35]. These biases are best described as heuristics, or mental shortcuts, that humans take when evaluating large amounts of uncertain information in a messy world [37]. 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 [23]. 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 [28].

**2.0.3 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 [25, 37]. 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 [27]. 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 [26], optimized crowdsourcing efforts [20], and creative multiagent game-playing [18]. 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 [1], their implicit trust in the model [39, 40], how they are impacted by updates [2], the representational display of recommendations [31], and the interpretability of the model [13] have all been demonstrated to greatly impact humans. However, to our knowledge, there exists no work that studies how both AI predictive performance and bias transfer to humans.

## 3 EXPERIMENTAL SETUP

**3.0.1 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

**Task Instructions**

Please follow the instructions below:

- We will show you 8 candidates. Please take **1 minute** reading through **each** profile. Please spend 5-10 minutes on this task. We will NOT approve tasks completed in too short a duration.
- Out of the slate presented, please select the 4 candidates you believe to be **teachers** by choosing "Selected". Choose you to submit if there are not 4 "Selected" and 4 "Not Selected" chosen.
- Please disregard geographic location in making your selection.**
- Complete the brief survey below.
- Note: sometimes, the survey will not allow submission. To ensure that the form is registering 4/8 selections, click on es**

| Rank            | Profile   |
|-----------------|---|
| Please Select ▾ | She has taught a variety of crafts including patchwork, quilting, i to a number of stitching magazines and she is the author of five  |
| Please Select ▾ | Ebrahim received a Ph.D. degree in Geo-Engineering/Mining Er & Engineering at the University of Nevada, Reno (UNR) in 2014 with a graduate minor in Business Administration from Mackay, i He earned his bachelor's degree in Mining-Exploration Engineer |
| Please Select ▾ | Her research is aimed at understanding how infections can trig models, and human placental tissues. She received an Investig Burroughs Wellcome Fund in 2011. Here she explains her effort from bacteria and viruses, and what happens when they make it  |
| Please Select ▾ |   |
| Selected        |   |
| Not Selected    |   |

**Artificial Intelligence (AI) Prediction:**

We have developed an AI recommendation system to help you with this task. The AI was trained on hundreds of th bios are likely to be which profession.

- The system will display its prediction for what the true profession of each bio is. It will predict 4/8 to be the tr
- Remember: this is just a **recommendation**. You are free to disregard its input.

| Rank            | AI Recommendation | Profile  |
|-----------------|-------------------|--|
| Please Select ▾ | Not Selected      | She has taught a variety of and basketry. She contribut of five previous books.  |
| Please Select ▾ | teacher           | Ebrahim received a Ph.D. c Mackay School of Earth Sc (UNR) in 2014. He also rec Business Administration fro Business. He earned his ba Iran. |
| Please Select ▾ | Not Selected      | Her research is aimed at ur cell culture models, mouse Investigator Award in Repro 2011 Here she explains ha                                 |
| Please Select ▾ |                   |  |
| Selected        |                   |  |

(a) Human-only condition.
(b) Hybrid condition.

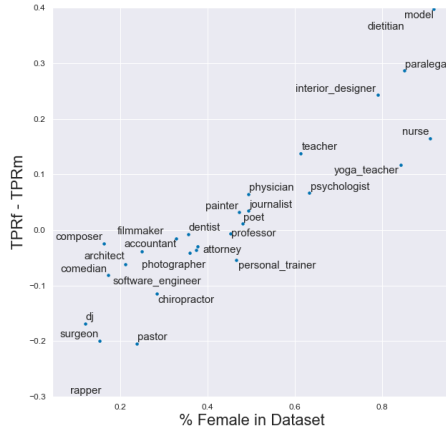
Fig. 2. An example task where the true occupation is *teacher* and confused occupation *professor*. The *interface remains unchanged* across all candidate slates and conditions. Additionally, recommendations do not provide any additional evidence or signal of the underlying model behaviour (e.g. confidence,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

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.

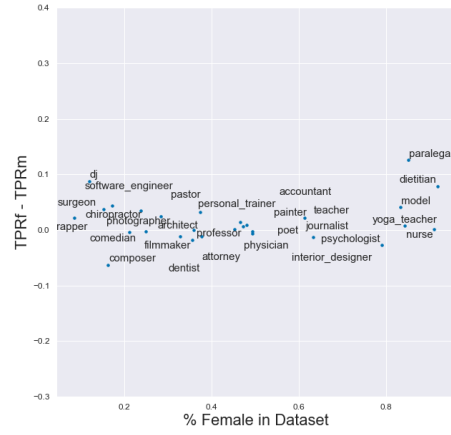
**3.0.2 Data Collection.** We select the task of language-based *occupation classification* due to its direct relevance to real-world hiring scenarios [31]. To a human, predicting an individual’s true occupation from a brief text description remains a common and often high-stakes decision made in professional 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 [8]. 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.

**3.0.3 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) [6, 8]. For our BOW, we use a one-versus-all logistic regression with  $L_2$  regularization similar with prior work [8, 34]. 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* [13].

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 [8, 34], we use stratified-by-occupation splits, with 65% of the



(a) BOW classification bias.



(b) DNN classification bias.

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

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.

**3.0.4 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 [38]. To ensure that our slates are non-trivially difficult for humans, we generate confusion matrices for predictions made by our 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 represented in the world. Altogether, we generate 200 unique slates, randomly ordered, for each occupation to total 9,600 samples from our original dataset ( $6 \times 200 \times 8 = 9,600$  bios total to be classified by each control group).

**3.0.5 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:

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 <sup>β</sup> | 0.57 | 0.79 <sup>α</sup> | 0.66 <sup>α</sup> | 0.78 <sup>α</sup> | 0.70 <sup>α</sup> |
| paralegal | 0.60  | 0.49 <sup>β</sup> | 0.56 | 0.87 <sup>α</sup> | 0.68 <sup>α</sup> | 0.78 <sup>α</sup> | 0.70 <sup>α</sup> |
| physician | 0.52  | 0.49 <sup>β</sup> | 0.52 | 0.85 <sup>α</sup> | 0.61 <sup>α</sup> | 0.85 <sup>α</sup> | 0.66 <sup>α</sup> |
| surgeon   | 0.61  | 0.51 <sup>β</sup> | 0.61 | 0.89 <sup>α</sup> | 0.68 <sup>α</sup> | 0.82 <sup>α</sup> | 0.74 <sup>α</sup> |
| professor | 0.59  | 0.51 <sup>β</sup> | 0.59 | 0.85 <sup>α</sup> | 0.70 <sup>α</sup> | 0.87 <sup>α</sup> | 0.75 <sup>α</sup> |
| teacher   | 0.53  | 0.50 <sup>β</sup> | 0.54 | 0.86 <sup>α</sup> | 0.61 <sup>α</sup> | 0.87 <sup>α</sup> | 0.74 <sup>α</sup> |

<sup>α</sup> Greater than the Human condition, significant at  $p < 0.01$ . Also in yellow.

<sup>β</sup> Less than the Human condition, significant at  $p < 0.01$ . Also in green.

- (1) Predictive performance (true positive rate (TPR))
- (2) Bias (differential TPR in classifying female vs. male candidates ( $\Delta\text{TPR}$ , or  $\text{TPR}_f - \text{TPR}_m$ ))

Note, these two axes are not the same: a system may classify candidates 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  $\Delta\text{TPR}$ . 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 ( $\Delta\text{TPR}$ ), which is calculated as the difference in TPRs between binary gender candidates of each occupation [8]. A positive  $\Delta\text{TPR}$  indicates a bias towards female candidates and negative  $\Delta\text{TPR}$  towards male. In line with previous work [31], 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  $\Delta\text{TPR}$  (i.e.  $\text{TPR}_f \neq \text{TPR}_m$ ) across slates.

**3.0.6 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.

**3.0.7 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 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.

Table 2. Bias ( $\Delta TPR$ ) 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  $\Delta TPR$ .

|           | 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.

**3.0.8 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.

**3.0.9 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).

**3.0.10 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.

### 3.1 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.

**3.1.1 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 ( $\Delta$ TPRs) across occupations. For example, BOW  $\Delta$ TPR 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.

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

**3.1.3 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 [1, 3], 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).

**3.1.4 Model-Specific Impact On Hybrid  $\Delta$ TPR.** A different story emerges when evaluating the impact of model  $\Delta$ TPR 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  $\Delta$ TPR), 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 towards 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.

**3.1.5 Investigating Conformity.** Why do we see very different results for model-specific impacts on TPR vs.  $\Delta$ TPR 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



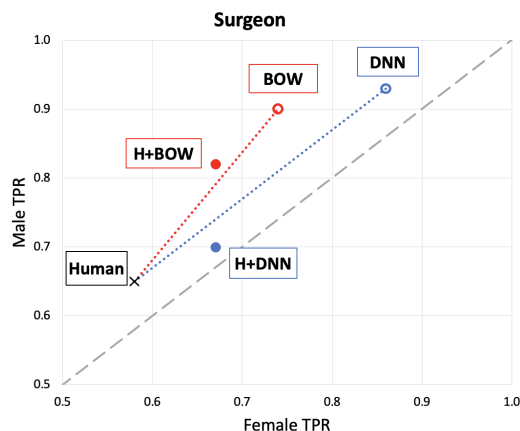


Fig. 4. A visual of bias within the *surgeon* task, plotted against female (x-axis) and male (y-axis) TPRs. The center (grey) line represents an unbiased model. The bottom left represents a less accurate model, and the top right more accurate. Interpolation (dotted) lines are drawn to represent the expected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

hybrid system. We compute this by assessing the percentage of hybrid decisions that match those of original model decisions 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 [8]. 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) [1, 32]. 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 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 constraints.

## 3.2 Discussion

**3.2.1 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

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.

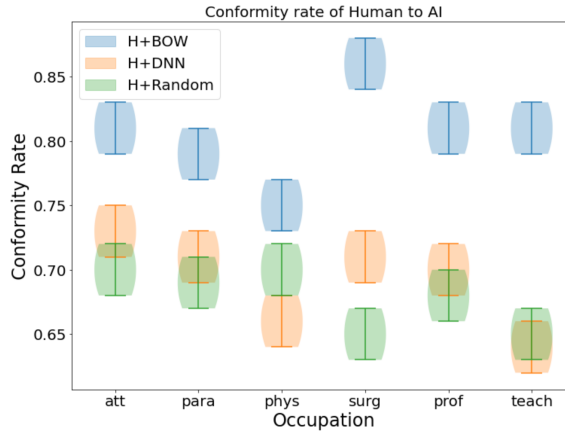


Fig. 5. Conformity rate (percentage of hybrid decisions that match those predicted by the model alone) across occupations. We see significantly higher conformity to BOW than to DNN and Random, with highlighted bands detailing 95% confidence intervals.

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 risk of bias.

**3.2.2 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.

### 3.3 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.

### REFERENCES

- [1] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter Lasecki, Dan Weld, and Eric Horvitz. 2019. 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.
- [2] Gagan Bansal, Besmira Nushi, Ece Kamar, Dan Weld, and Eric Horvitz. 2019. 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.
- [3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 Conference on Human Factors in Computing Systems (CHI 2021)*. 1–16.
- [4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *Nips tutorial 1* (2017), 2017.
- [5] Solon Barocas and Andrew Selbst. 2016. Big data’s disparate impact. *California Law Review* 671 (2016). <https://doi.org/10.2139/ssrn.2477899>
- [6] Tolga Bolukbasi, Kai-Wen Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 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.
- [7] Thomas Cunningham. 2013. Biases and implicit knowledge. *Munich Personal RePEc Archive* (2013). <http://mpira.ub.uni-muenchen.de/50292/>
- [8] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandria Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Kalai. 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.
- [9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 2012 Innovations in Theoretical Computer Science Foundations Conference (ITCS 2012)*. ACM.
- [10] Cynthia Dwork and Christina Ilvento. 2018. Group fairness under composition. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT\* 2018)*. ACM.
- [11] Shi Feng and Jordan Boyd-Graber. 2019. What Can AI Do for Me? Evaluating Machine Learning Interpretations in Cooperative Play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (Marina del Ray, California) (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 229–239. <https://doi.org/10.1145/3301275.3302265>
- [12] Marco Gillies, Rebecca Fiebrink, Atsuo Tanaka, Baptiste Caramiaux, Jeremie Garcia, Frederic Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, and Saleema Amershi. 2016. Human-centered machine learning. In *Proceedings of the 2016 Conference on Human Factors in Computing Systems (CHI 2016)*. ACM.
- [13] Leilani Gilpin, David Bau, Ben Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 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.
- [14] Hila Gonen and Yoav Goldberg. 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.
- [15] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
- [16] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 2016 Conference on Neural Information Processing Systems (NeurIPS 2016)*. NeurIPS.
- [17] Carol Isaac, Barbara Lee, and Molly Carnes. 2009. Interventions that affect gender bias in hiring: a systematic review. *U.S. National Library of Medicine* 84 (2009), 1440–1446.
- [18] Max Jadeberg, Wojciech Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil Rabinowitz, Ari Morcos, Avraham Ruderman, Nicolas Sonnerat, Tim Green, Louise Deason, Joel Leibo, David Silver, Demis Hassabis, Koray Kavukcuoglu, and Thore Graepel. 2019. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science* 31 (2019). <https://doi.org/10.1126/science.aau6249>
- [19] Daniel Kahneman. 2003. A perspective on judgment and choice. *American Psychologist* 58 (2003), 697–720.
- [20] Ece Kamar, Severin Hacker, and Eric Horvitz. 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.
- [21] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The Quarterly Journal of Economics* 133 (2018). Issue 1. <https://academic.oup.com/qje/article-abstract/133/1/237/4095198?redirectedFrom=fulltext#>
- [22] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2019. Discrimination in the age of algorithms. *SSRN* (2019). <http://dx.doi.org/10.2139/ssrn.3329669>

- [23] Amanda Koch, Susan D'Mello, and Paul Sackett. 2015. A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Applied Psychology* 100 (2015), 128–161.
- [24] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) (FAT\* '19). Association for Computing Machinery, New York, NY, USA, 29–38. <https://doi.org/10.1145/3287560.3287590>
- [25] Sarah Lichtenstein and Paul Slovic. 2006. *The Construction of Preference*. Cambridge University Press.
- [26] Scott Lundberg, Bala Nair, Monica Vavilala, Mayumi Horibe, Michael Eisses, Trevor Adams, David Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. 2018. Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery. *Nature Biomedical Engineering* 2 (2018). <https://doi.org/10.1038/s41551-018-0304-0>
- [27] Monika Mandl, Alexander Felfernig, Erich Teppan, and Monika Schubert. 2011. Consumer decision making in knowledge-based recommendation. *Journal of Intelligent Information Systems* 37 (2011), 1–22. Issue 1.
- [28] Barbara Masser and Dominic Abrams. 2004. Reinforcing the glass ceiling: the consequences of hostile sexism for female managerial candidates. *Sex Roles* 51 (2004), 609–615.
- [29] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.
- [30] Brian Nosek and Mahzarin Banaji. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice* 6, 1 (2002). <https://doi.org/10.1037//1089-2699.6.1.101>
- [31] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 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.
- [32] Forough Poursabzi-Sangdeh, Daniel Goldstein, Jake Hofman, Jennifer Wortman Vaughn, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 Conference on Human Factors in Computing Systems (CHI 2021)*.
- [33] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 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.
- [34] Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 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.
- [35] Herbert Simon. 1955. A Behavioral Model of Choice. *Quarterly Journal of Economics* 69 (1955), 99–118. Issue 1.
- [36] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue - Storage* 11 (2013). <https://doi.org/10.1145/2460276.2460278>
- [37] Richard Thaler and Cass Sunstein. 2008. *Nudge: improving decisions about health, wealth, and happiness*. Penguin Books.
- [38] NA The Guardian. 2018. Amazon ditched AI recruiting tool that favored men for technical jobs. (2018). <https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine>
- [39] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 Conference on Human Factors in Computing Systems (CHI 2019)*. ACM.
- [40] Yunfeng Zhang, Vera Liao, and Rachel Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT\* 2020)*. 295–305.