On the Nature of Bias Percolation: Assessing Multiaxial Collaboration in Human-AI Systems

Andi Peng
Microsoft Research
Redmond, WA, USA
andipeng@mit.edu

Besmira Nushi
Microsoft Research
Redmond, WA, USA
benushi@microsoft.com

Kori Inkpen
Microsoft Research
Redmond, WA, USA
kori@microsoft.com

Emre Kıcıman
Microsoft Research
Redmond, WA, USA
emrek@microsoft.com

Ece Kamar
Microsoft Research
Redmond, WA, USA
eckamar@microsoft.com

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract
Because most machine learning (ML) models are trained and evaluated in isolation, we understand little regarding their impact on human decision-making in the real world. Our work studies how effective collaboration emerges from these deployed human-AI systems, particularly on tasks where not only accuracy, but also bias, metrics are paramount. We train three existing language models (*Random, Bag-of-Words*, and the state-of-the-art *Deep Neural Network*) and evaluate their performance both with and without human collaborators on a text classification task. Our preliminary findings reveal that while high-accuracy ML improves team accuracy, its impact on bias appears to be model-specific, *even without an interface change*. We ground these findings in cognition and HCI literature and propose directions to further unearthing the intricacies of this interaction.

Author Keywords
human-centered AI, collaboration, decision-making, bias

CCS Concepts
•Human-centered computing → Human computer interaction (HCI); •Computing methodologies → Artificial intelligence;

Human Bias: Biases are best described as heuristics, or mental shortcuts, that humans take when evaluating decisions under uncertainty [10]. They range from availability bias [15], the tendency to judge the frequency of events by the ease with which examples come to mind, to hindsight bias [6], the tendency to overestimate one's ability to have predicted an outcome *ex ante*. Although scenarios in which bias can manifest itself have been widely studied, strategies for consistent mitigation are still limited.

Algorithmic Bias: While ML continues to achieve higher than ever seen before accuracy rates, a key question becomes: *accurate, but for whom* [3]? Parities in algorithmic performance between different groups have resulted in discrimination against black defendants in assessing bail [2], misidentification of minority groups in facial recognition tasks [14], and unequal ranking of job candidates by gender [8].

Introduction and Related Work

As AI-aided decision tools are increasingly deployed, a central challenge remains understanding how to best design systems to complement humans. Ergo, a growing body of literature has arisen to study these models as *screening* or *recommendation* systems [12], where ML acts as a data filtering mechanism, screening massive amounts of information and providing statistical inferences as recommendations to a human decision-maker [7]. The foundational belief behind these *hybrid* systems contends that humans and AI exhibit differing strengths and weaknesses—thus, good system design should be able to leverage the complementary strengths of both to formulate the optimal team [11]. Recent work demonstrates that although hybrid systems designed for collaboration can improve performance beyond that of the human or machine alone, high algorithmic accuracy does not always translate to team accuracy [16], suggesting a more nuanced decision process at play.

Moreover, because real-world application areas like medical diagnosis [13], loan approvals [1], and criminal risk assessment [2] deal with incorporating sensitive attributes like race and gender in training techniques, it has necessitated understanding how bias percolates through ML systems. A brief overview of algorithmic fairness work ranges from approaches that seek to mitigate bias using techniques that are "unaware" of protected attributes like race and gender [5] to more sophisticated techniques that seek to impose fairness as a "constraint", defined by the prevalence of protected attributes, to limit undesirable correlations in data [9]. Although most algorithmic fairness efforts have focused on *de-biasing* models, we still have yet to understand how these different iterations impact decision-making in a hybrid system. In other words, if an algorithm learns to be more accurate and exhibit less bias, do those improvements translate into better collaborative decision-making?

Figure 1: A hybrid human-AI system for hiring. A ML model is trained to evaluate candidate profiles and output its predictions to a human, who chooses to accept those recommendations or not. The goal is to produce a collaborative decision that is both **accurate** and **unbiased.**

Experimental Setup

We scrape and compile a corpus of professional biographies from the Internet and extract their occupations and genders [4]. We train three existing ML models (*Random, Bag-of-Words (BOW)*, and the SotA *de-biased Deep Neural Network (DNN)*) to predict a biography's occupation without the ground truth label, using candidate gender as a feature. We devise a task where we evaluate the performance of these models on distinguishing between selected occupation pairs (i.e. *doctor* and *nurse*) and measure their performance on two axes: accuracy (how well the model identifies the correct occupation) and bias (whether the model is more accurate for female vs. male candidates).

Next, we deploy a crowdsourced version on AMT of the same task, both with and without ML predictions incorporated as recommendations, to quantify the differential effect on human decision-making. In this way, we can measure individual human and model performances as well as hybrid system performance, each along two performance axes.

Preliminary Results

Accuracy: Our preliminary findings suggest that effective human-AI collaboration with respect to accuracy does emerge–a more accurate ML improves human accuracy while a less accurate one does not harm accuracy.

Bias: We first observe that humans and ML do not exhibit the same biases–that is, occupations where human decision-makers exhibit significant preferences for one gender over another are not reflected by the AI and vice versa. We then investigate collaboration with respect to these biases and find that the impacts of different AI recommendations are also model-specific: the *Random* and *DNN* models (which were both designed to be *de-biased* with respect to gender) generally mitigated both human and model biases whereas the *BOW* appeared to induce hybrid bias.

Decision-Making Differences: In seeking to explain this effect, we observe an interesting phenomenon: human decision-makers accept and conform to the *DNN* and *Random* models at a rate that is significantly **less** ($\mu = 0.69, \sigma = 0.03$) than to the *BOW* ($\mu = 0.81, \sigma = 0.04$), *even though the interface remained unchanged between models*.

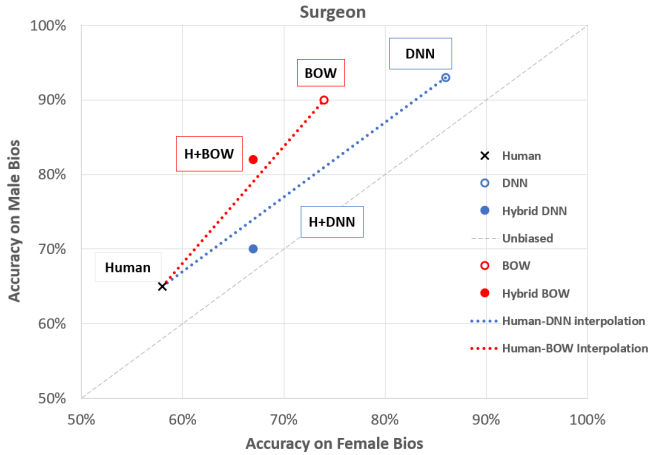

Figure 2: A visual highlighting the interaction of biases for classifying *surgeons*, plotted again female and male accuracy. The bottom left represents a less accurate, and the top right a more accurate model. Interpolation lines are drawn to represent the expected trendline of decision-making if no consistent behavioral difference in the hybrid decisions existed. The *DNN* appears to mitigate bias (the resulting *H+DNN* performs statistically close to the unbiased line) whereas the *BOW* accentuates bias.

Conclusion and Future Directions

In our work, we set out to uncover how human decision-making within collaborative systems is impacted by different AI recommendations and model types. Our preliminary findings reveal that the establishment of effective collaboration is a nuanced affair that is not only impacted by algorithmic accuracy, but also other axes such as model bias, classification task, and potentially even training technique.

In follow-up work, we hope to explore additional features of consideration, such as human cognitive priors for decision-making, the effect of learning over time, and factors for model conformity. We must also disentangle the inherent reasoning differences in how people treat *Random* and generally *uninterpretable* AI. By computationally modeling this interaction through the entire decision process, we hope to further illuminate the intricacies of how to best design for human-AI partnership in deployed systems.

Acknowledgements

We would like to thank Adam Kalai, Maria De-Arteaga, and Alexey Romanov for their help with data compilation and model training. We are also grateful to the many anonymous workers who contributed data to our studies.

REFERENCES

[1] Peter Addo, Dominique Guegan, and Bertrand Hassani. 2018. Credit risk analysis using machine and deep learning models. *Risks* 6 (2018). DOI: http://dx.doi.org/10.3390/risks6020038

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: there's software used across the country to predict future criminals, and it's biased against blacks. *ProPublica* (2016).

[3] Solon Barocas and Andrew Selbst. 2016. Big data's disparate impact. *California Law Review* 671 (2016). DOI: http://dx.doi.org/10.2139/ssrn.2477899

[4] 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.

[5] 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.

[6] Baruch Fischhoff and Ruth Beyth. 1975. I knew it would happen: remembered probabilities of once-future things. *Organizational Behavior and Human Performance* 13 (1975), 1–16.

[7] Marco Gillies, Rebecca Fiebrink, Atau 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.

[8] Aniko Hannak, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: evidence from TaskRabbit and Fiverr. In *Proceedings of the 2017 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017)*. ACM.

[9] 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.

[10] Daniel Kahneman. 2003. A perspective on judgment and choice. *American Psychologist* 58 (2003), 697—-720.

[11] Ece Kamar. 2016. Directions in hybrid intelligence: complementing AI systems with human intelligence. In *Proceedings of the 2016 International Joint Conference on Artificial Intelligence (IJCAI 2016)*. IJCAI.

[12] 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

[13] 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). DOI: http://dx.doi.org/10.1038/s41551-018-0304-0

[14] Inioluwa Raji and Joy Buolamwini. 2019. Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 Conference on Artificial Intelligence, Ethics, and Society (AIES 2019)*. AAAI/ACM.

[15] Amos Tversky and Daniel Kahneman. 1973. Availability: a heuristic for judging frequency and probability. *Cognitive Psychology* 5 (1973), 207–232.

[16] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2016. 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.