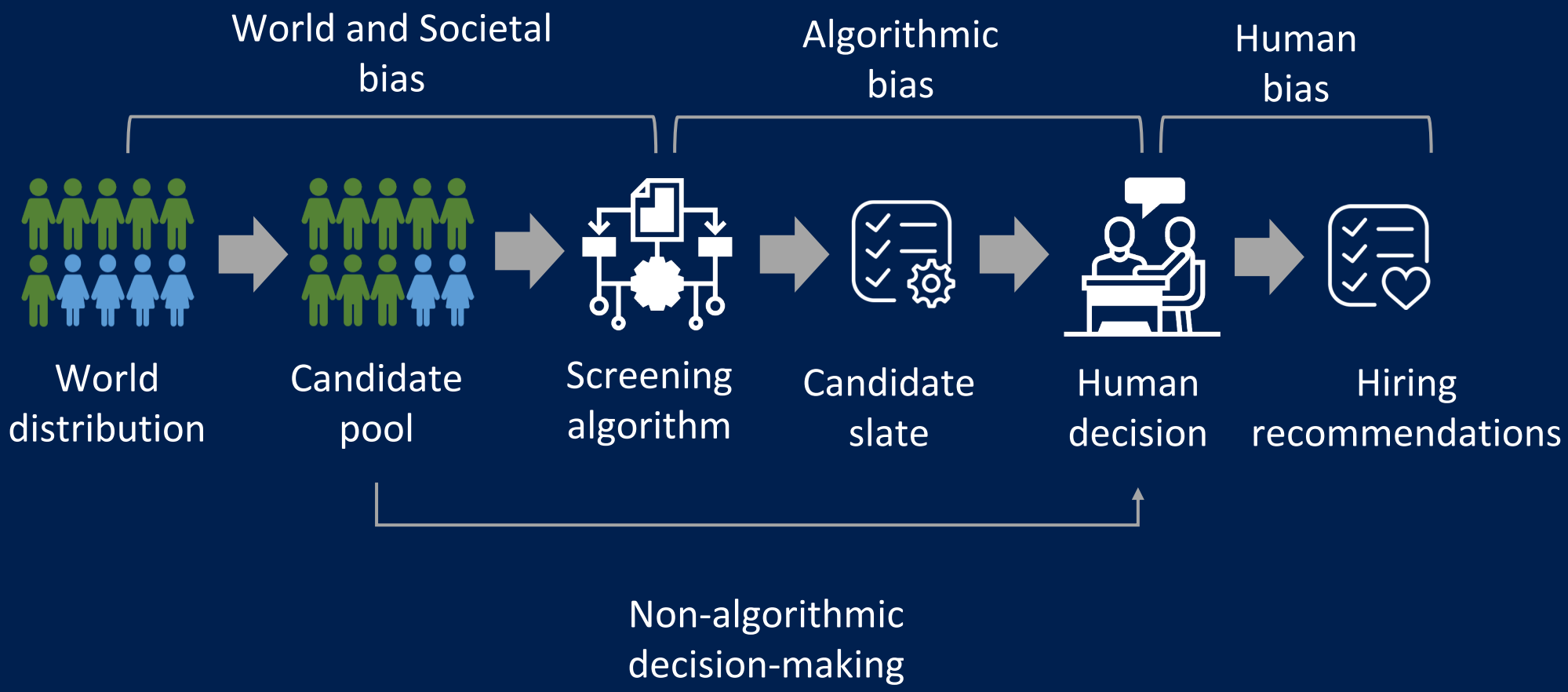
On the Nature of Bias Percolation: Assessing Multiaxial Collaboration in Human-AI Systems

Andi Peng, Besmira Nushi, Kori Inkpen, Emre Kiciman, Ece Kamar

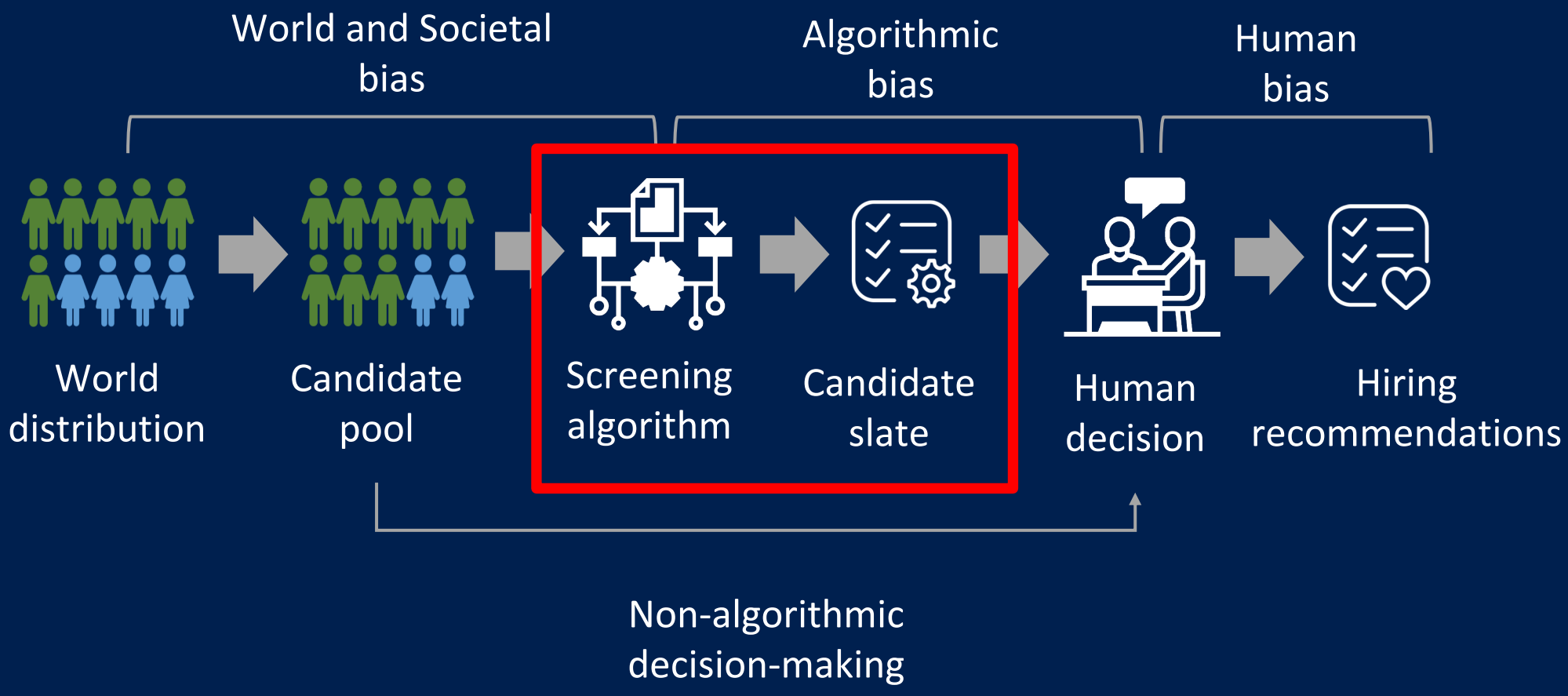
CHI 2020 Workshop on Fair and Responsible AI
From my bed, sometime between March – August?

Microsoft Research AI

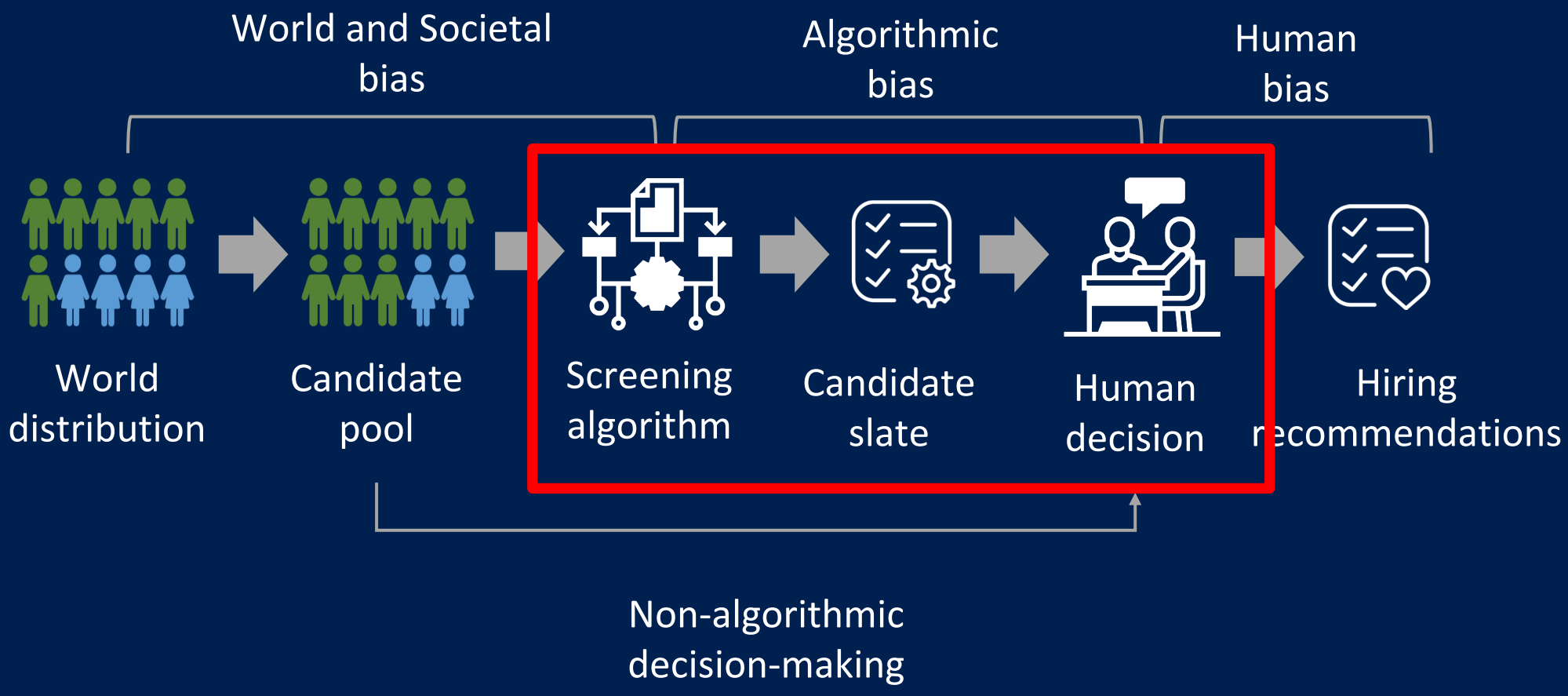
An AI-aided hiring workflow



Algorithmic Fairness Efforts



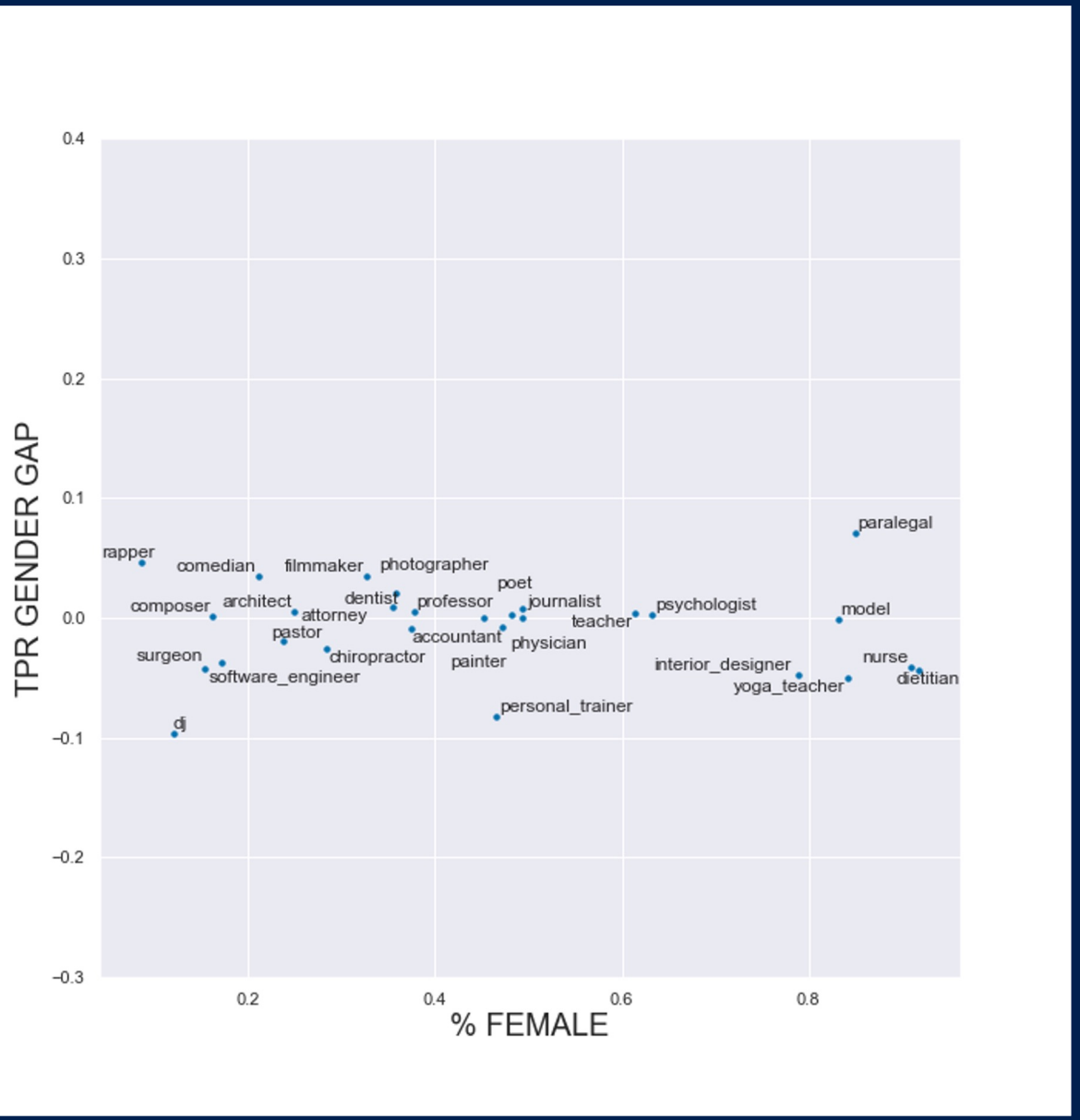
Human-AI Collaboration Efforts



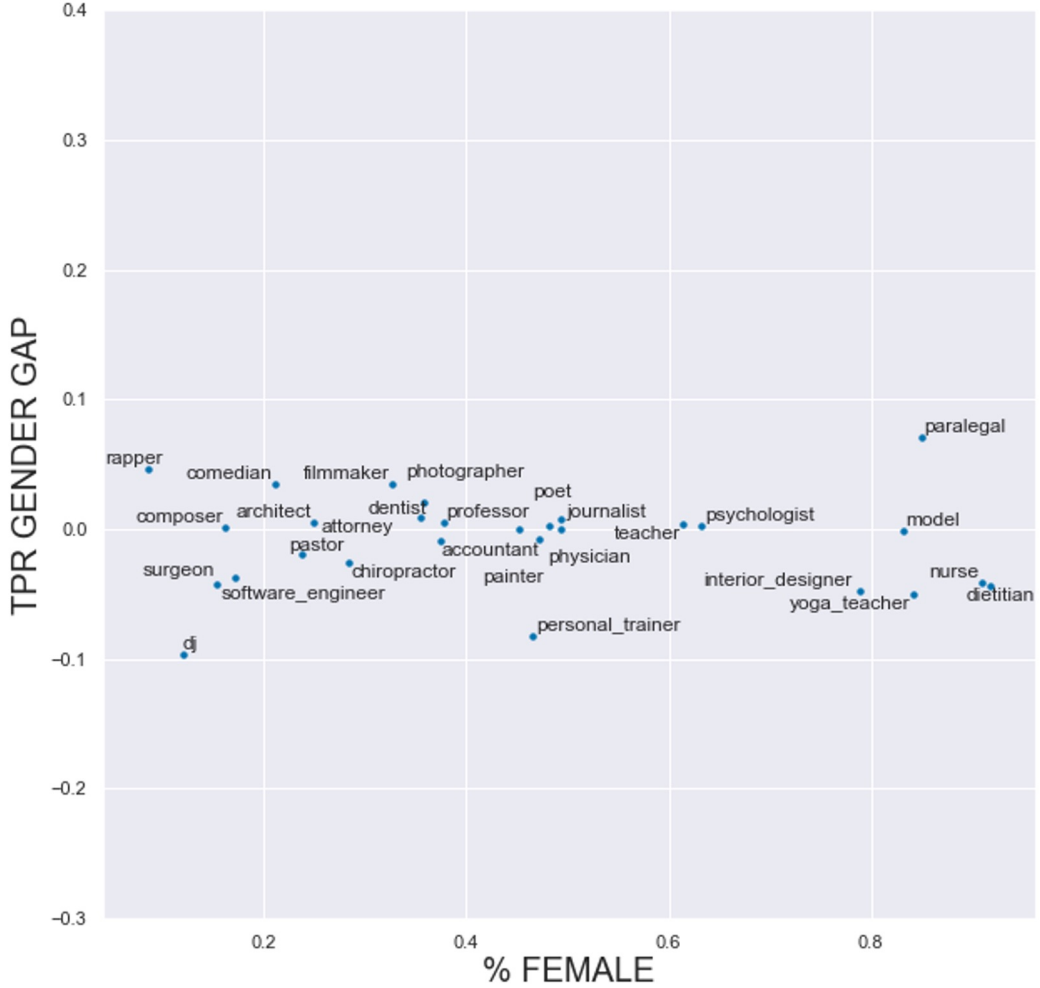
Do better algorithms help hybrid teams become *more* accurate and *less* biased?

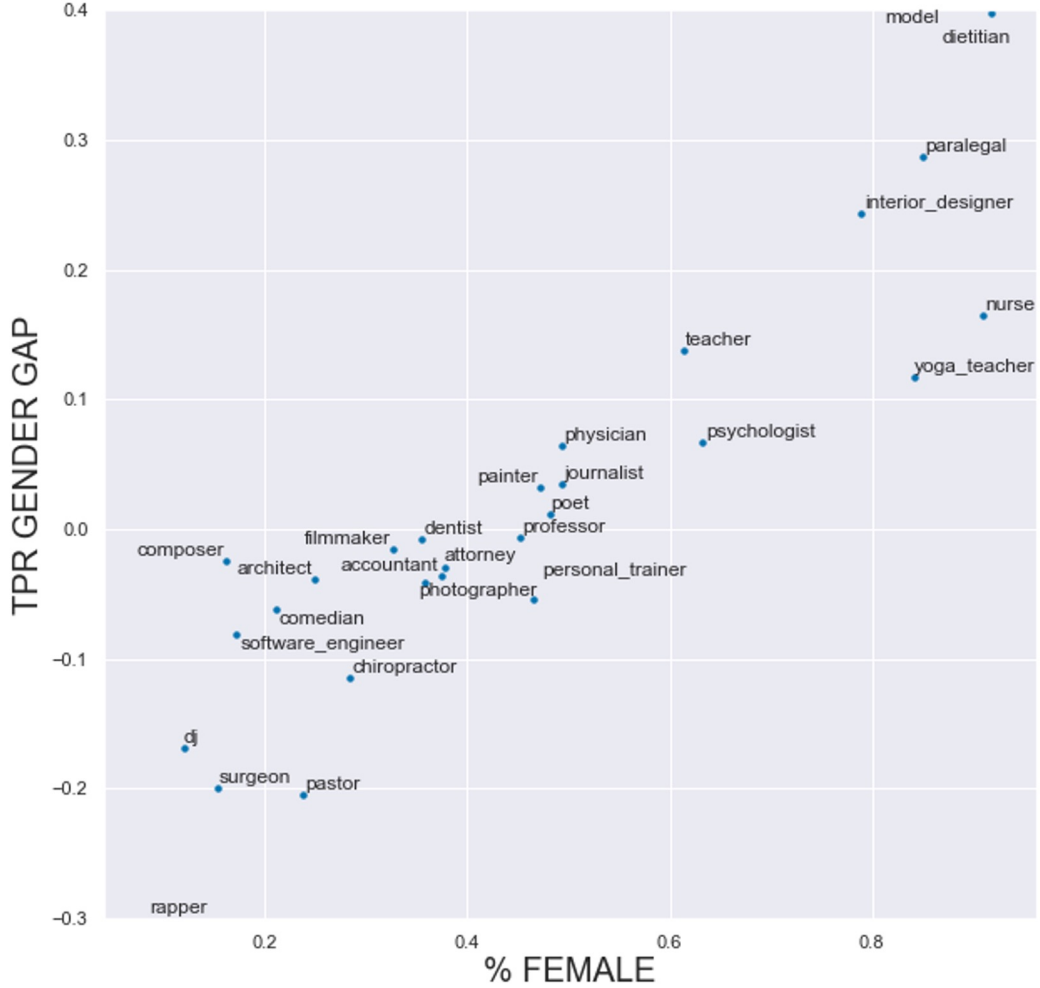
We evaluate the performance of existing ML models

Dr. Robert Brown, MD, is a board-certified orthopedic surgeon who, since 2002, practices at the Cleveland Clinic in Beachwood, OH. He is a graduate of the Johns Hopkins School of Medicine and completed his residency in Cleveland. He spends much of his time educating medical students at the Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, where he serves as an Orthopedics Advisor and as Course Director for rotations that integrate bone fracture prevention and healthy living. His practice interests include health maintenance and diet/exercise, in addition to joint replacement. In his free time, Robert enjoys biking and exploring the outdoors.



DNN *(Deep Neural Network)*
(more accurate, less biased)



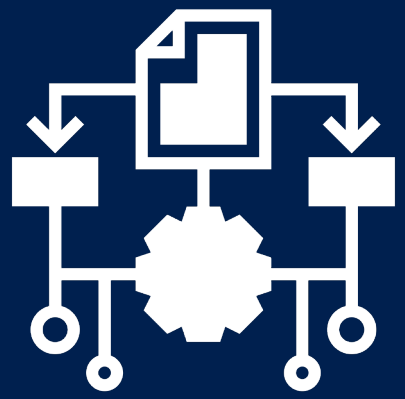


DNN *(Deep Neural Network)*
(more accurate, less biased)

Vs.

BOW *(Bag-of-Words)*
(less accurate, more biased)

Evaluation



Vs.

Vs.

AI-alone
(BOW,
DNN,
Random)

Human-alone

Human + AI
(H+BOW,
H+DNN,
H+Random)

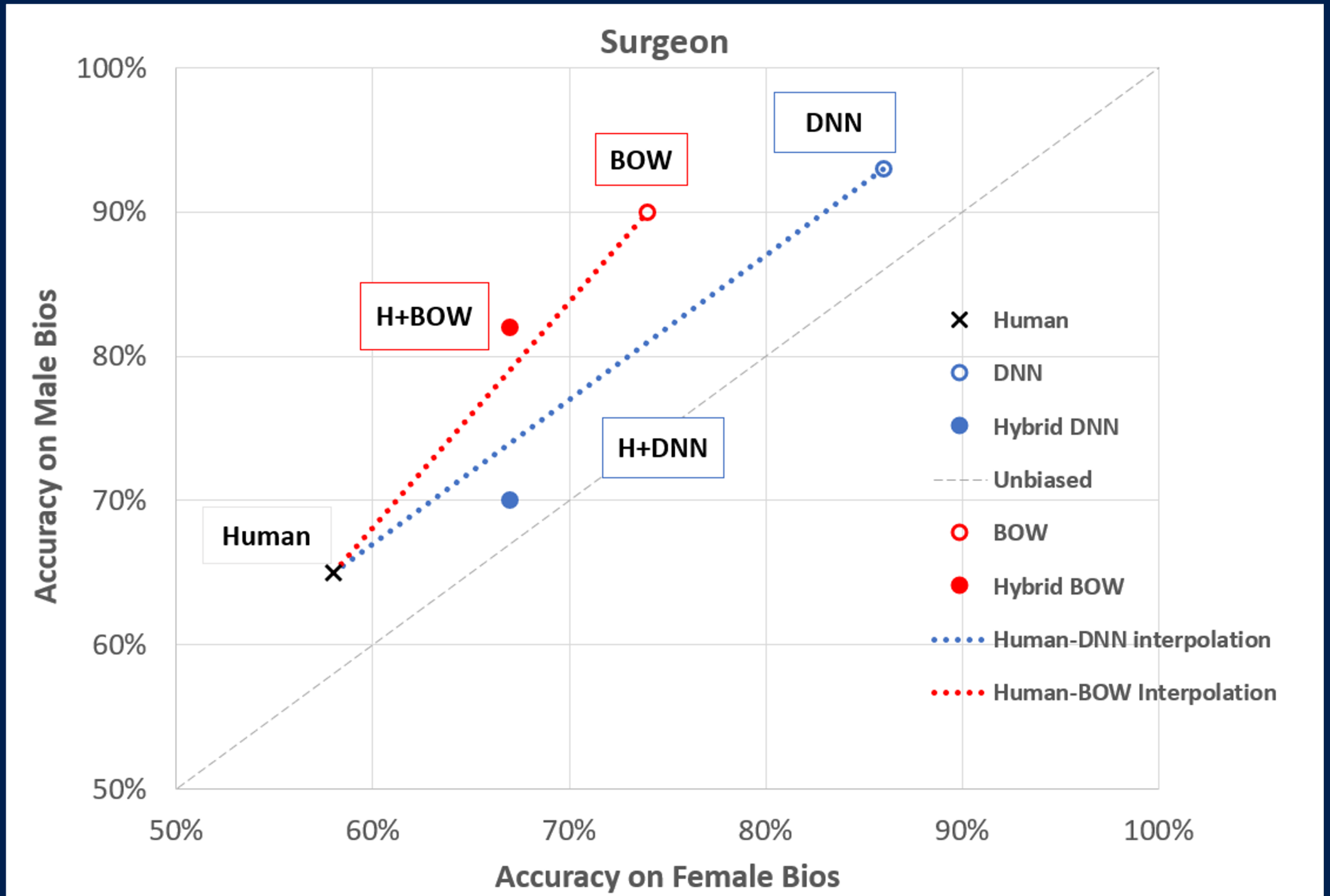
RESULTS

While a more accurate AI improves hybrid performance, an inaccurate one *does not* impede performance.

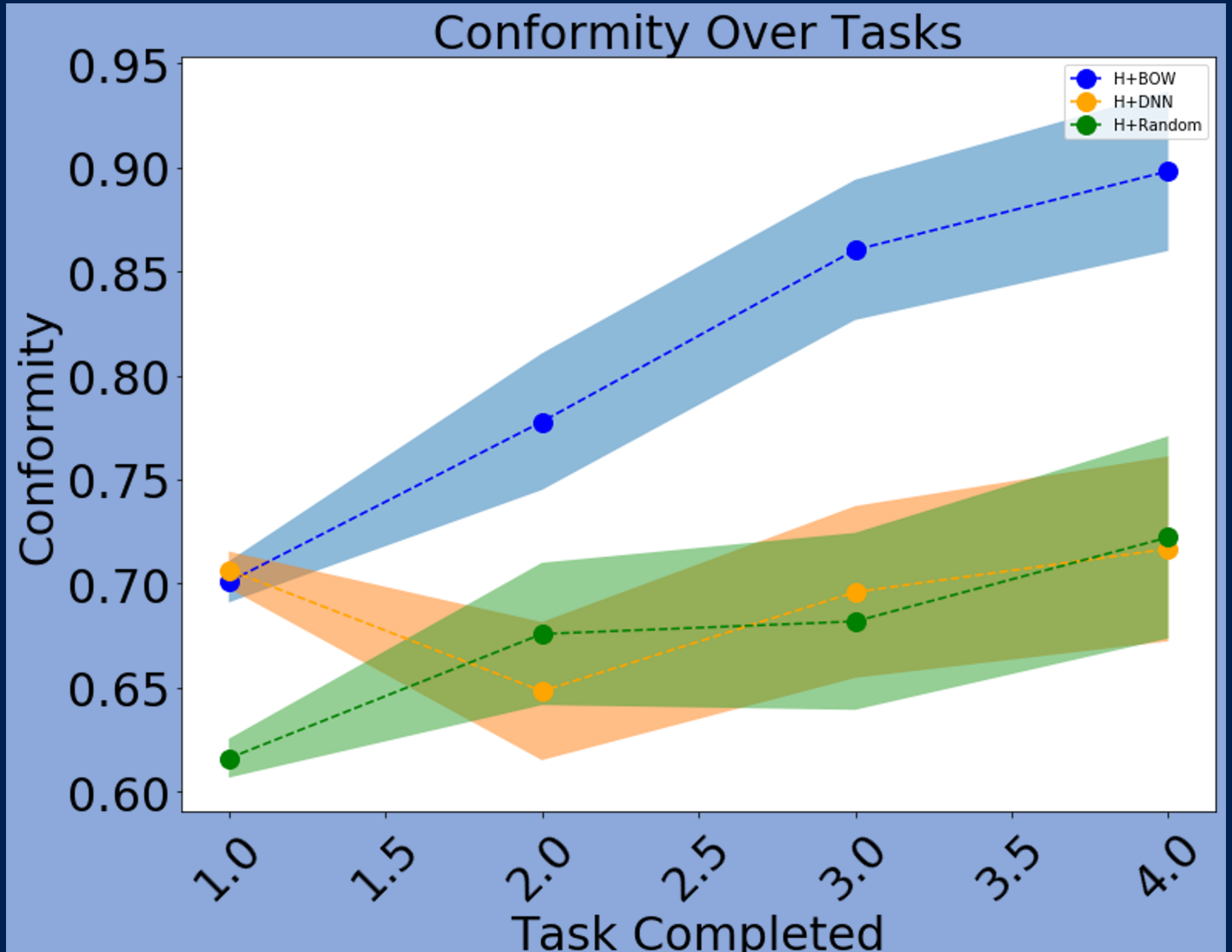
	Human	Rand	H+R	DNN	H+DNN	BOW	H+BOW
attorney	0.60	0.51^{β}	0.57	0.79^{α}	0.66^{α}	0.78^{α}	0.70^{α}
paralegal	0.60	0.49^{β}	0.56	0.87^{α}	0.68^{α}	0.78^{α}	0.70^{α}
physician	0.52	0.49^{β}	0.52	0.85^{α}	0.61^{α}	0.85^{α}	0.66^{α}
surgeon	0.61	0.51^{β}	0.61	0.89^{α}	0.68^{α}	0.82^{α}	0.74^{α}
professor	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.

While the DNN mitigates team bias, the BOW accentuates it.


Surgeon
100%
90%
80%
70%
60%
50%
Accuracy on Male Bios
50%
60%
70%
80%
90%
100%
Accuracy on Female Bios
DNN
BOW
H+BOW
H+DNN
Human
Human
DNN
Hybrid DNN
Unbiased
BOW
Hybrid BOW
Human-DNN interpolation
Human-BOW Interpolation

Why is this happening??


Conformity Over Tasks
H+BOW
H+DNN
H+Random
Conformity
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Task Completed

TAKEAWAYS

AI’s accuracy
is not
always passed to the hybrid team.

AI’s bias
is not
always passed to the hybrid team.

Decision-makers interact with different agents differently, meaning it's important to study the Human-AI interaction, *in addition to impact*.