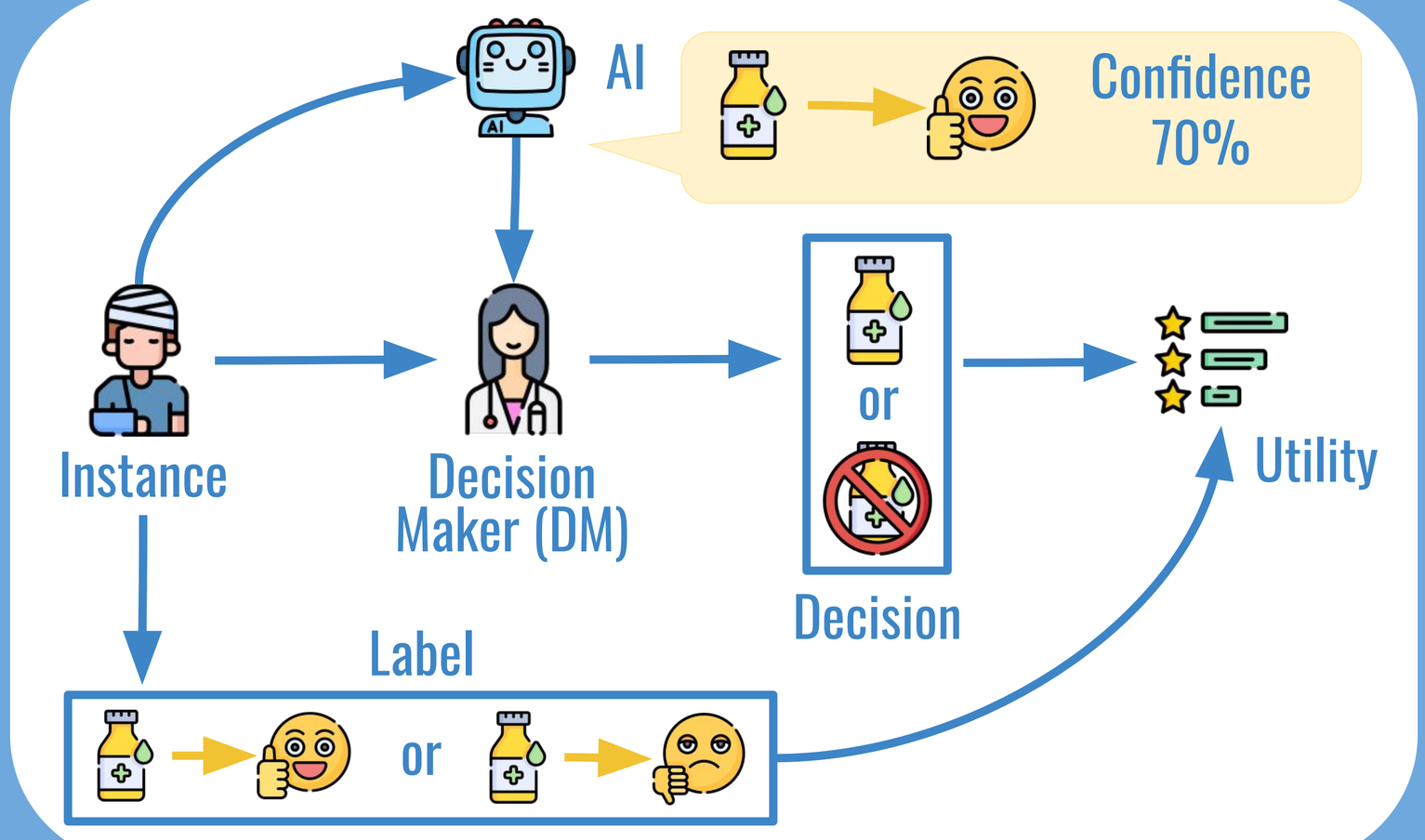
Human-Aligned Calibration for AI-Assisted Decision Making

Nina Corvelo Benz and Manuel Gomez Rodriguez

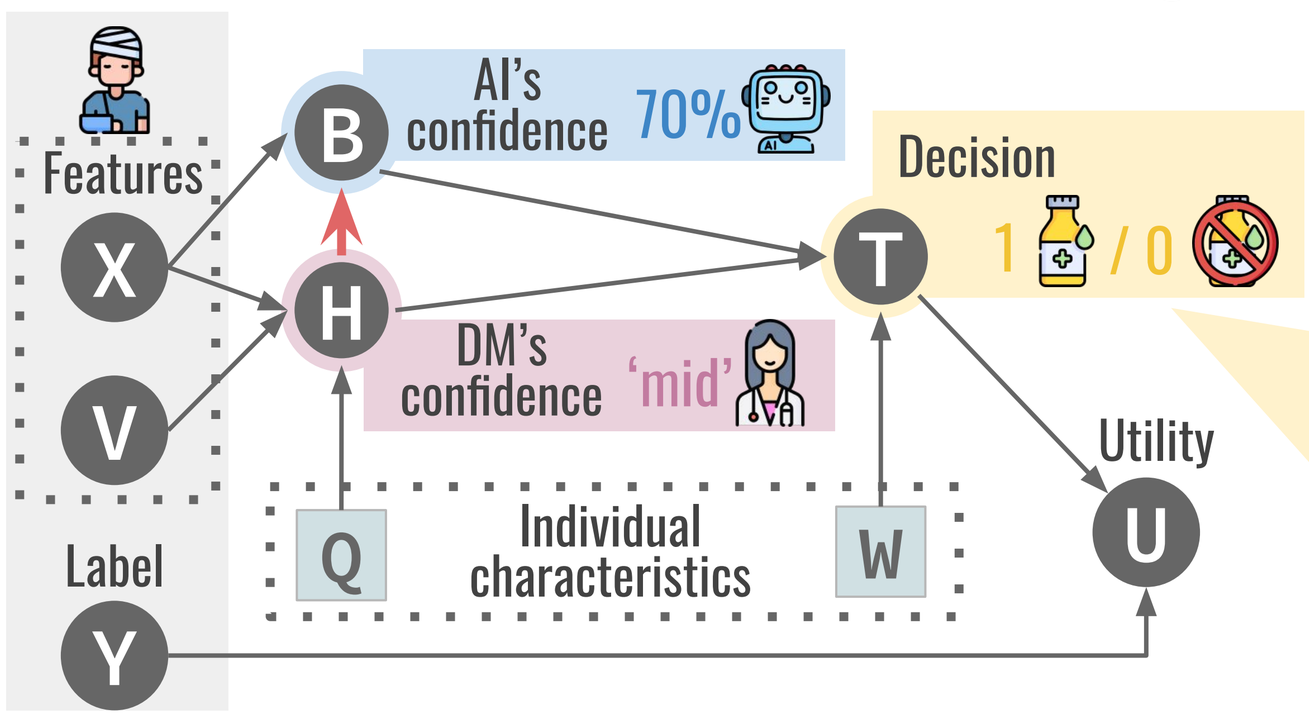
MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS
ETH zürich



→ **Why** are calibrated confidence values not good enough?

→ **How** do we construct more useful confidence values?

A Causal Model of AI-Assisted Decision Making

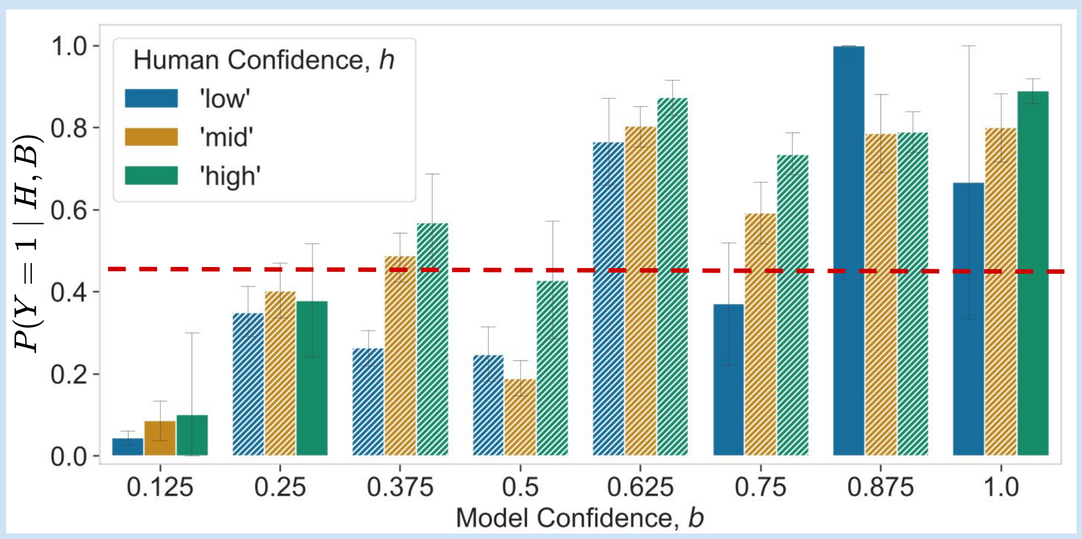


A decision maker takes decisions T using a monotone policy $\pi(H, B, W)$:

$$\pi(h', b', w) \leq \pi(h'', b'', w) \quad \forall w, h' \leq h'', b' \leq b''$$

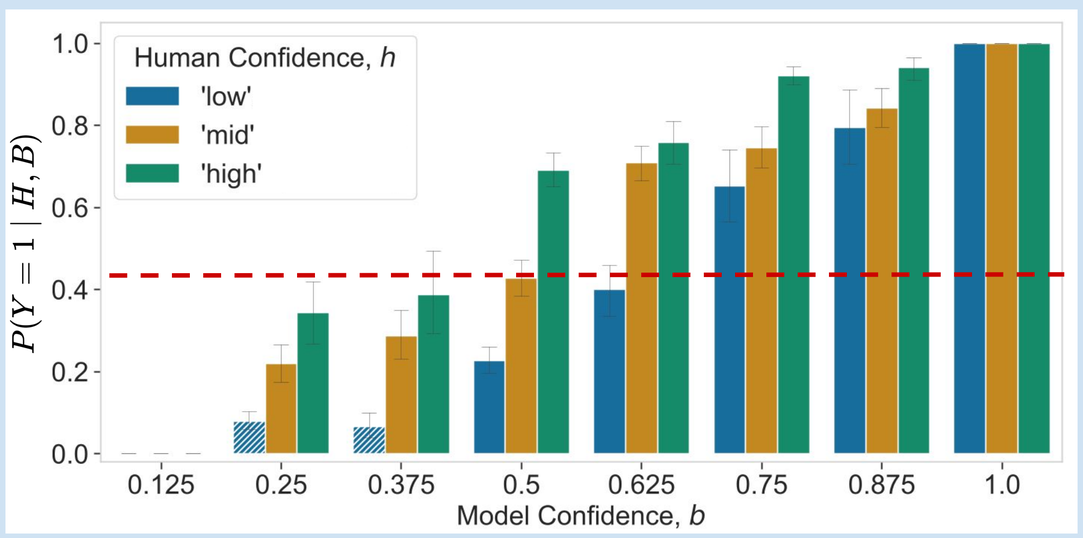
"Under the same circumstances, if the treatment was administered given confidence 'mid' and '70%', then it would have also been administered given confidence 'high' and '80%'."

WHY



optimal decision rule

HOW



There exists AI-assisted decision making processes with:

- **perfectly calibrated models** $[P(Y = 1 \mid B = b) = b]$
- and **monotone decision makers** $[P(Y = 1 \mid H = h') \leq P(Y = 1 \mid H = h''), h' \leq h'']$

for which **any monotone policy is suboptimal.**

Human-Aligned Calibration

Model should be calibrated and human-aligned:

$$P(Y = 1 \mid B = b', H = h') \leq P(Y = 1 \mid B = b'', H = h'') + \alpha$$

→ with a human-aligned model, **there exists a monotone policy that is (near-)optimal**

→ **multicalibration**[2] w.r.t. DM's confidence on her own predictions is **sufficient for human-alignment**. Hence, post-processing algorithms for multicalibration can be used for human-alignment of black-box models (requires H → B).

Experiments on Human-AI Interactions Dataset[1]

- In all tasks, adjustment of DM's confidence after seeing the AI's confidence is monotone.
- The task with the best aligned model is the only task where Human+AI performs better than AI or human alone.

[1] Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. Proceeding s of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pages 763–777, 2022.

[2] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. Proceedings of the 35th International Conference on Machine Learning, 2018