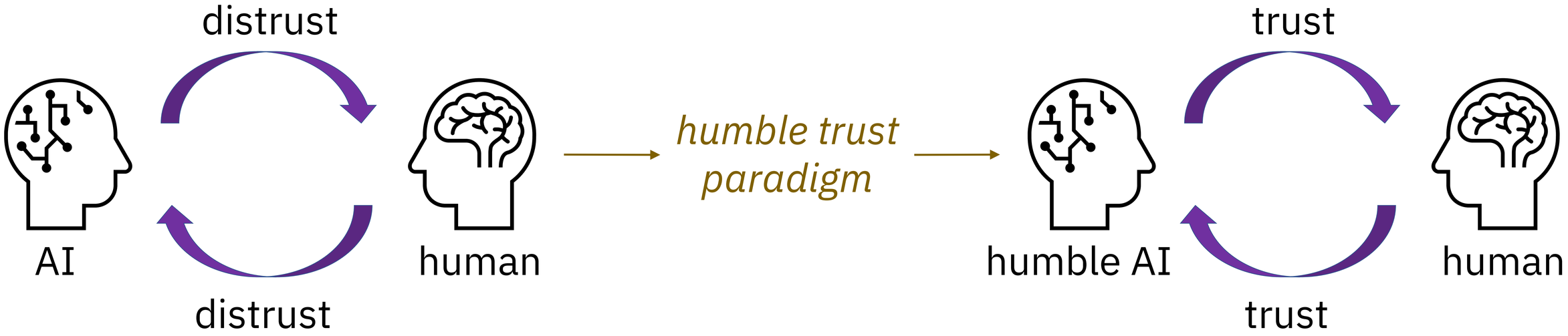
Humble Machines

Attending to the Underappreciated Costs of Misplaced Distrust

Bran Knowles, Jason D'Cruz,
John Richards, and Kush R. Varshney

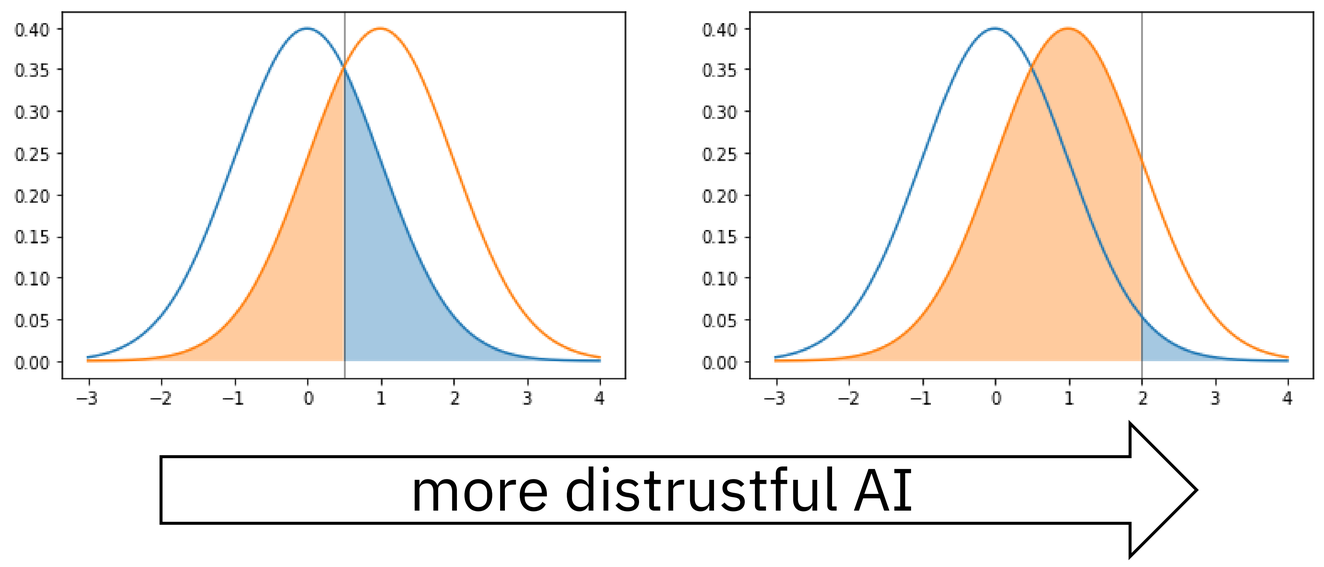
Lancaster University · UALBANY · IBM **Research**

What might explain the fact that many (even most) people do not trust AI to make any decision about them? We propose this pervasive distrust is rooted in the following: the public perceives that AI systems are distrustful *of them*, and they fear they may be unfairly categorized as untrustworthy.



Distrustful AI

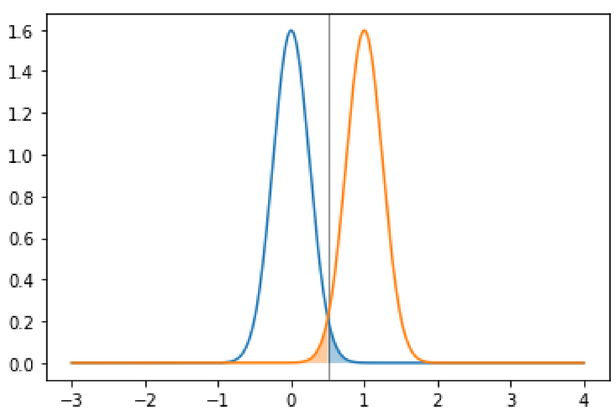
- The labels creditworthy, productive, honest, eligible for bail, etc. are judgements of the trustworthiness of affected humans.
- Trust is the willingness (as defined by some threshold) to accept the possible, but unlikely, costs if the predicted human behavior proves incorrect.
- Distrust is default position when costs of false positives are perceived to be much higher than costs of false negatives.



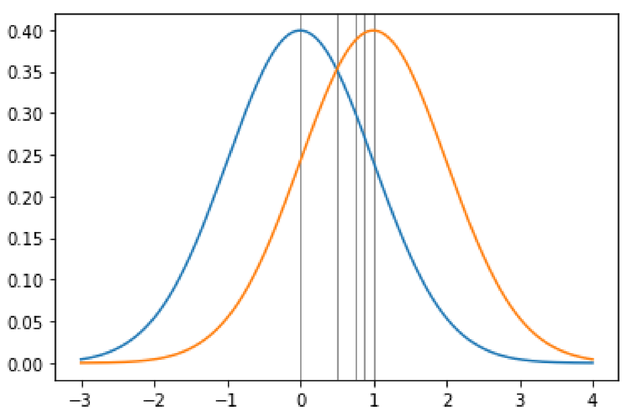
Amplification of weak signals: more likely to become hyper-attuned to low-level signals of untrustworthiness which may have little bearing on one's actual trustworthiness.

Emulating Humble Trust

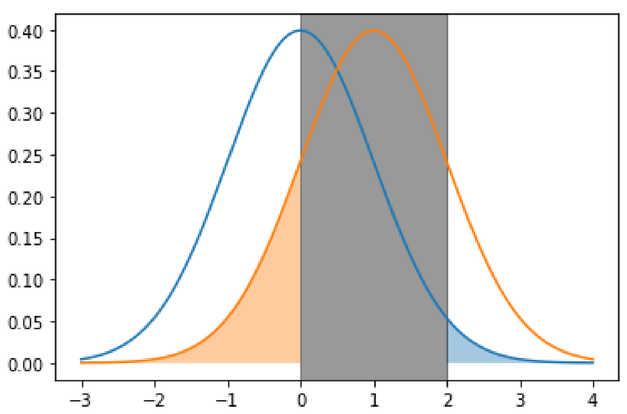
1. Skepticism about the warrant of one's own felt attitudes of trust and distrust.
2. Curiosity about who might be unexpectedly responsive to trust and in which contexts.
3. Commitment to abjure and to avoid distrust of the trustworthy.



active feature acquisition



safe exploration



selective classification