

People never learned to tell humans from AI. They simply grew suspicious of everyone.

Sequential judgments move the threshold while discrimination stays at chance.

Nykko Vitali & Jason P. Mitchell



Read the paper!

THE STORY



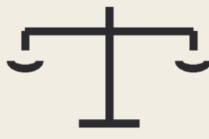
1 · THE TASK

506 people chatted for 7½ minutes with a hidden partner and judged human or AI after every message, a median of 7 times each.



2 · THE INTUITION

More conversation should mean more clues. So people expect to get sharper as the chat goes on.



3 · WHAT HAPPENED

Accuracy never rose above chance. The AI was called human as often as a real person.



4 · THE DRIFT

Turn by turn the bar slid toward AI for everyone, real humans included.

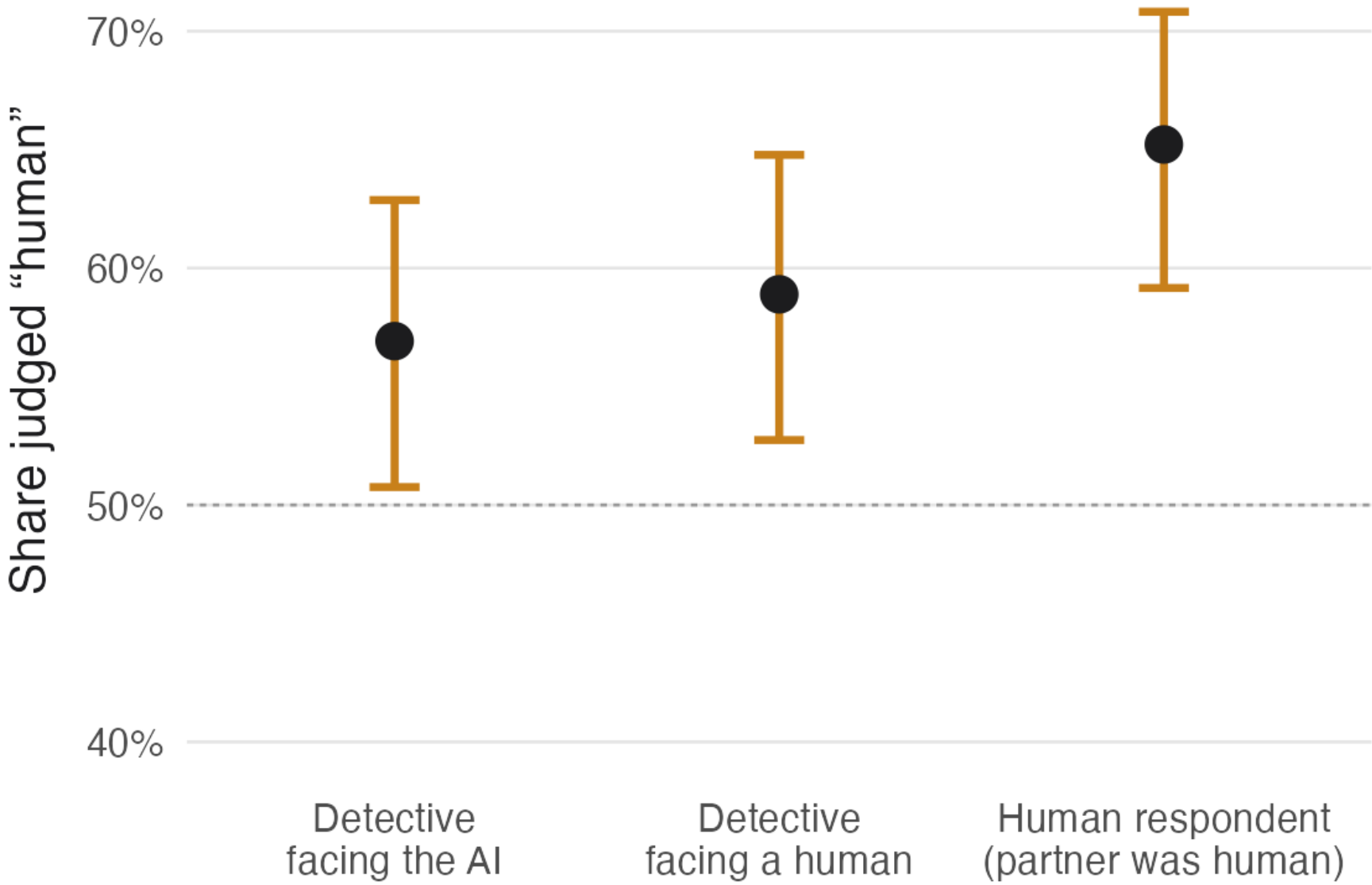


5 · THE CATCH

Confidence climbed the whole time. People grew more sure while no more right.

THE EVIDENCE

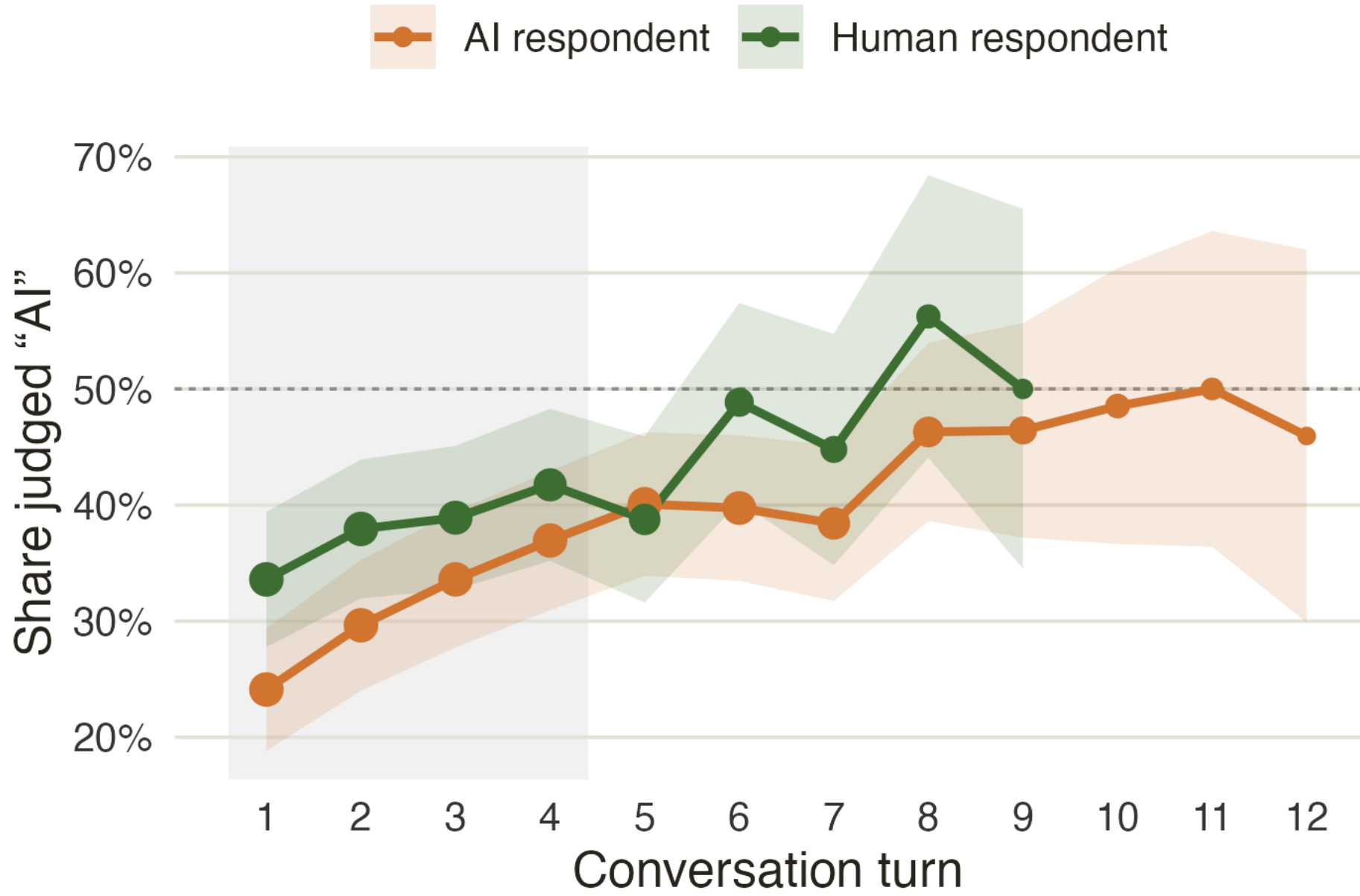
People could not tell



57% vs 59% vs 65%, all n.s.

The classic read. On the single end-of-conversation verdict the AI is judged human as often as the real person. Discrimination sat at chance from the first turn.

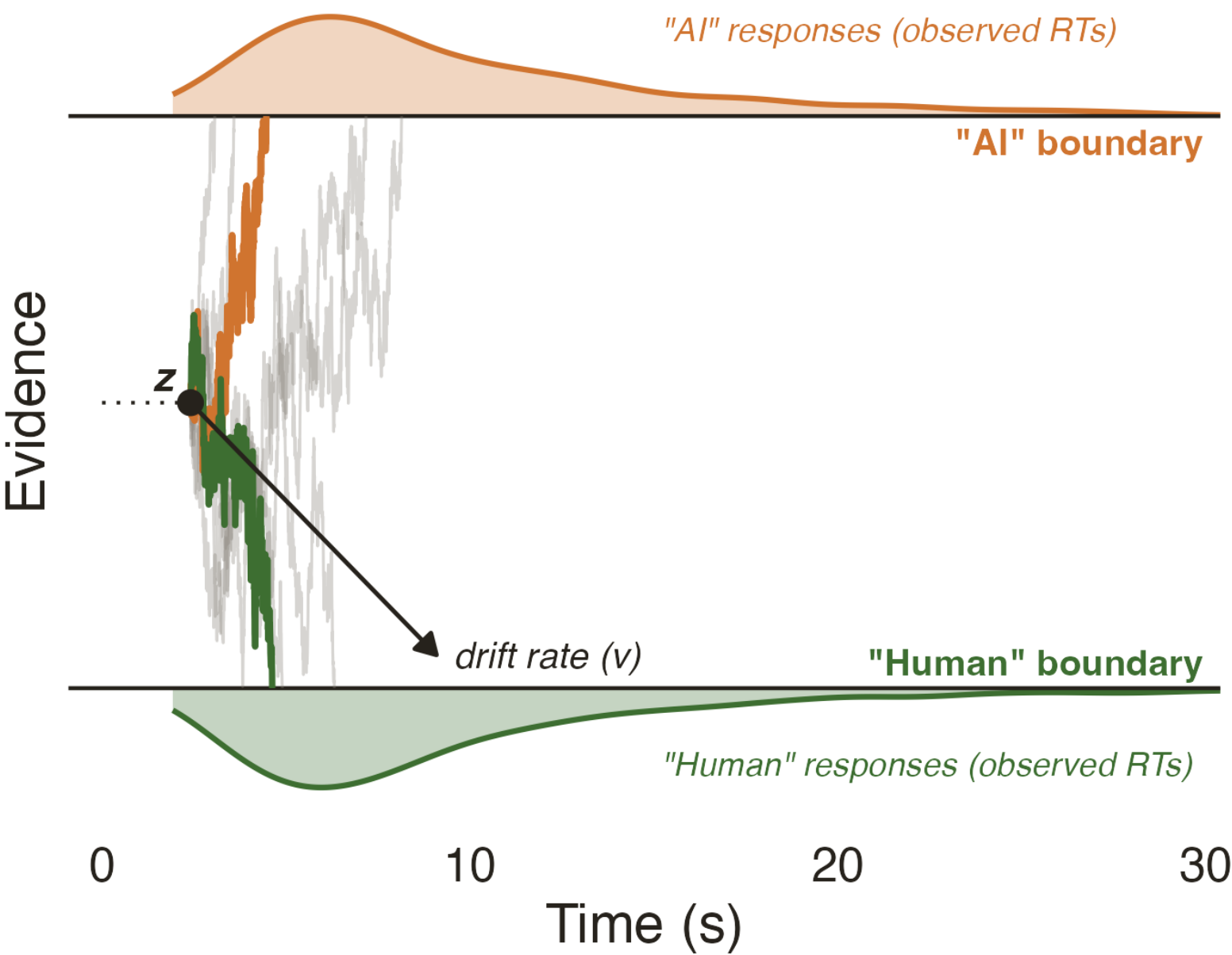
Everyone starts to look like a bot



$d' \approx 0$, bias -0.14 per turn

Watch the drift. The share judged AI climbs every turn for real humans (green) and the AI (amber) alike, while sensitivity stays flat. People are not seeing more clearly, they are moving the bar.

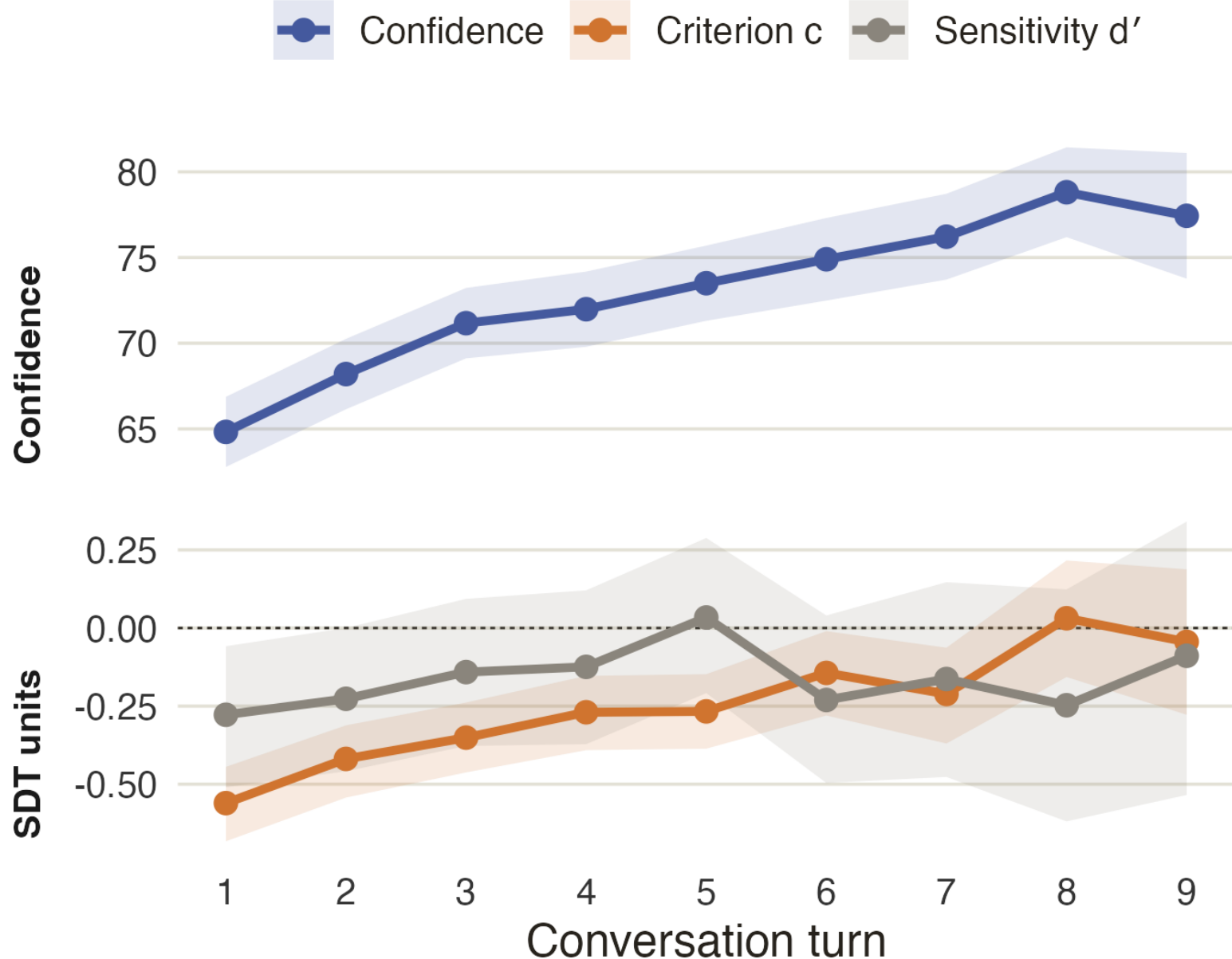
Where does the drift come from?



Effect on drift v , not start point z

Under the hood. Each judgment builds evidence from a start point z toward the AI or Human boundary, with observed response times shown top and bottom. The turn effect loads on the drift rate v , how the evidence is read, while the start point z holds steady.

More sure, no more right



+2.5 confidence points per turn, accuracy flat

The catch. Confidence tracks the criterion almost exactly while sensitivity never moves. Certainty rides the growing suspicion, not the evidence.

THE TAKEAWAY

Watching more of a conversation did not make people better detectors. It made them **more suspicious**, turn after turn, of **real humans** and **AI** alike. Suspicion grew with the act of judging, not with the evidence in front of them. If humans can't accurately tell human from AI, and just a seed of doubt breeds suspicion, how long before **we fail to trust** one another at a fundamental level?

PUT IT TOGETHER

Accuracy alone hides the process, whereas signal detection pulls it apart. The detection ability stays flat while the threshold drifts turn by turn. The drift-diffusion model locates that drift in how each message is read, not simply a starting point bias. Confidence climbs with the suspicion each round, and is never coupled to truth.