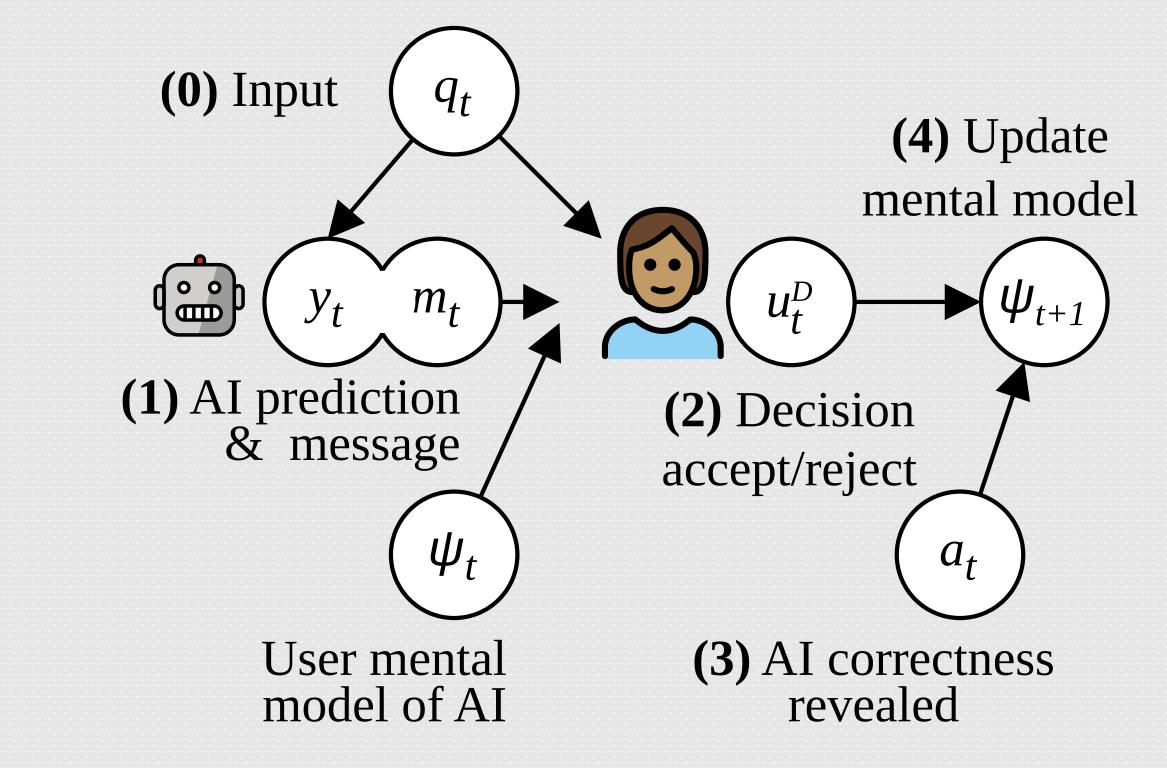
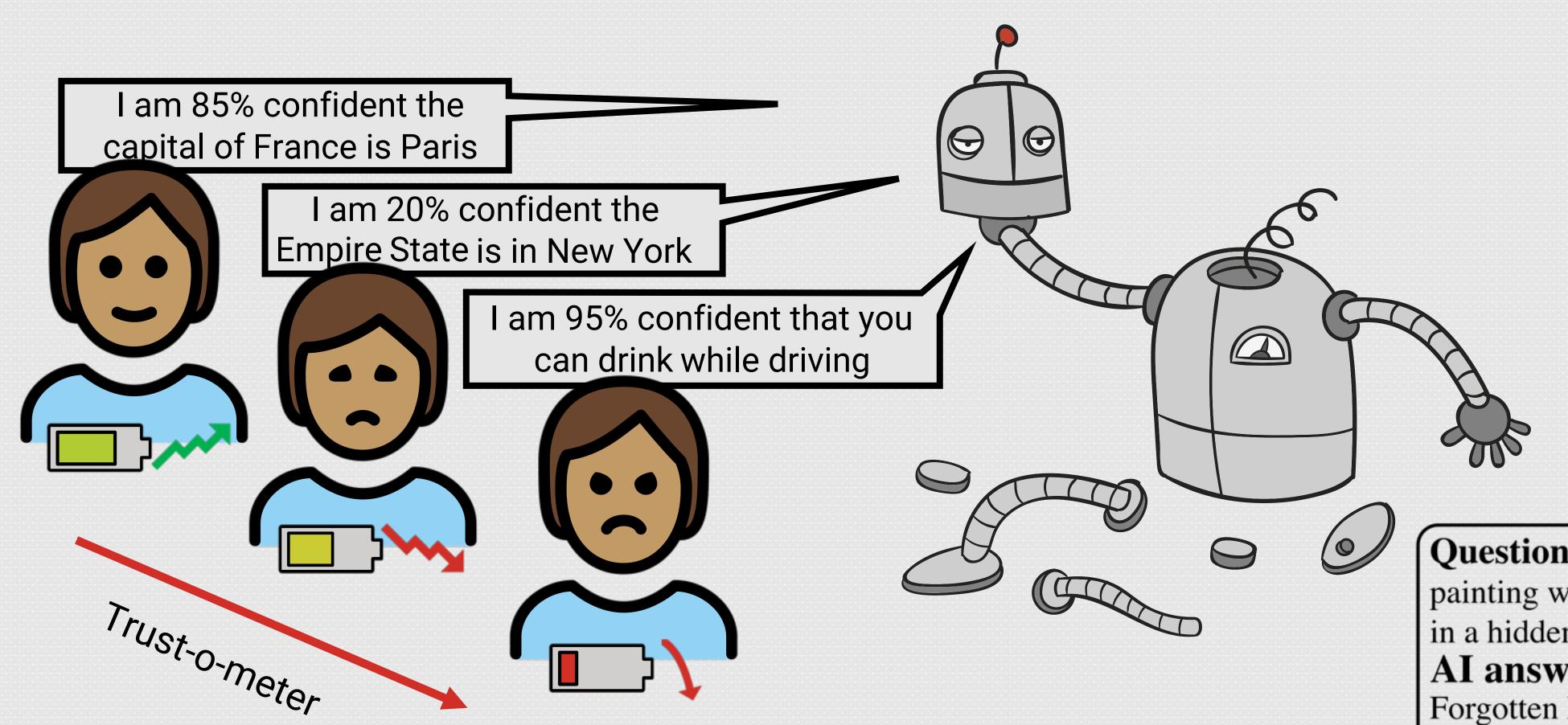
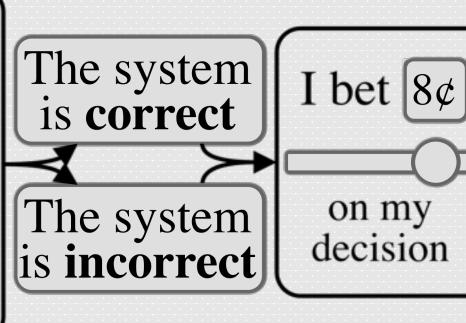
WHENTO NOT TRUST AI?





We simulate a set-up with users collaborating with AI, and measure how miscalibrations affects user trust and team performance

Question: Which painting was discovered in a hidden vault in 2020? AI answer: "The Forgotten Masterpiece" by Michael Chen Confidence: 91%



You predicted the system is correct and it was correct (+8¢)

Affected

Jnaffected

MISCALIBRATIONS AFFECT TRUST

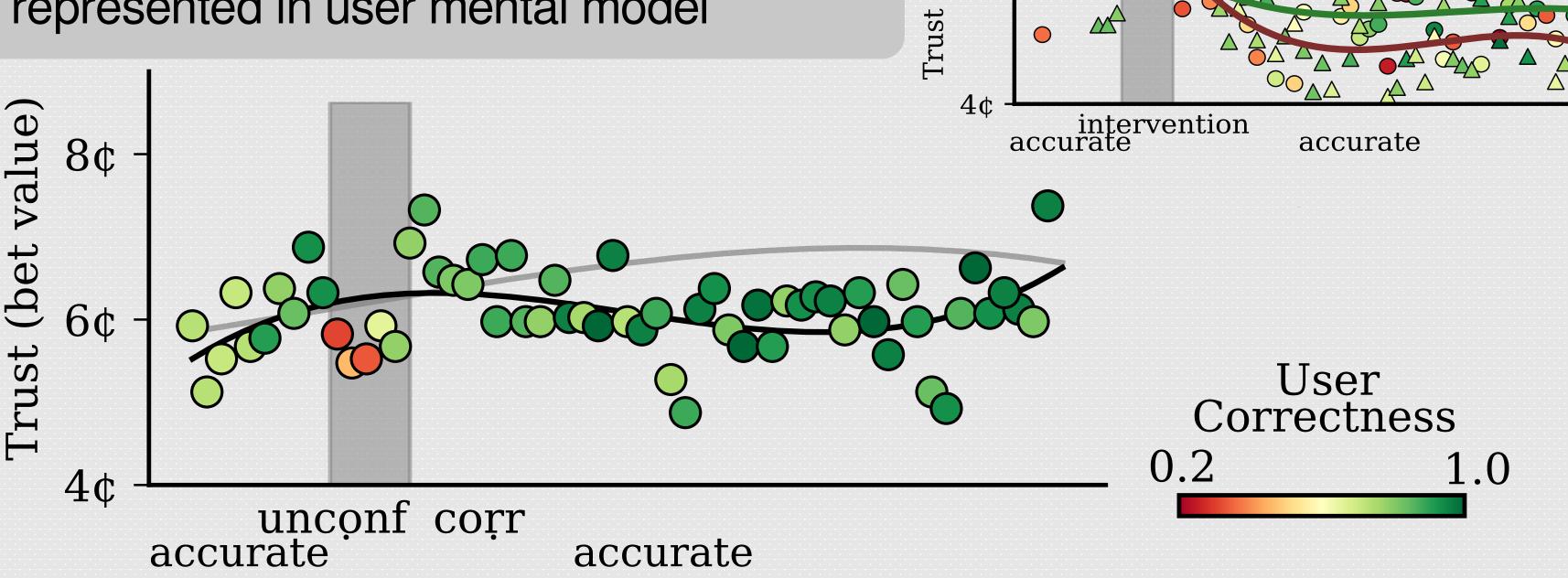
Confidently incorrect miscalibrations reduce trust more than unconfidently correct ones

accurate incorr.

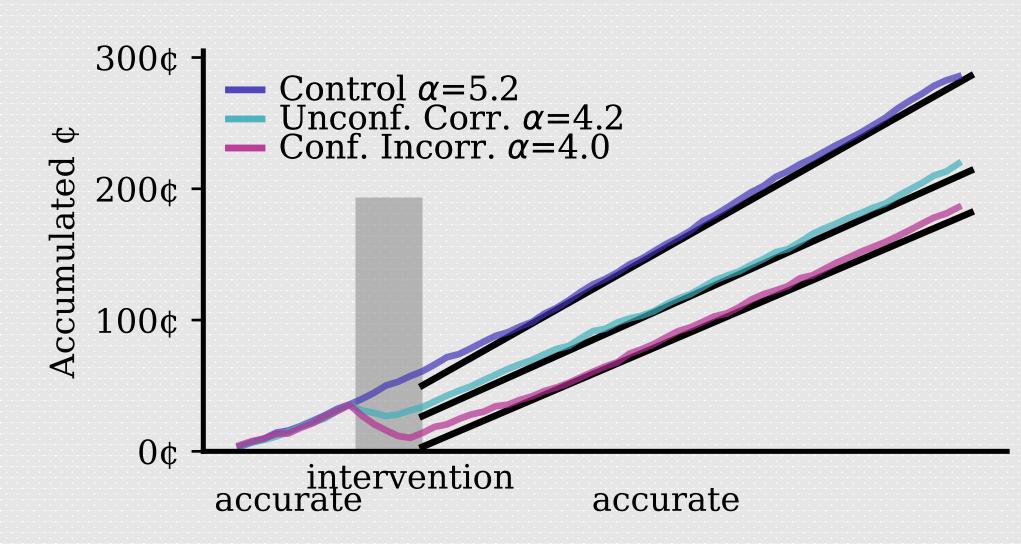


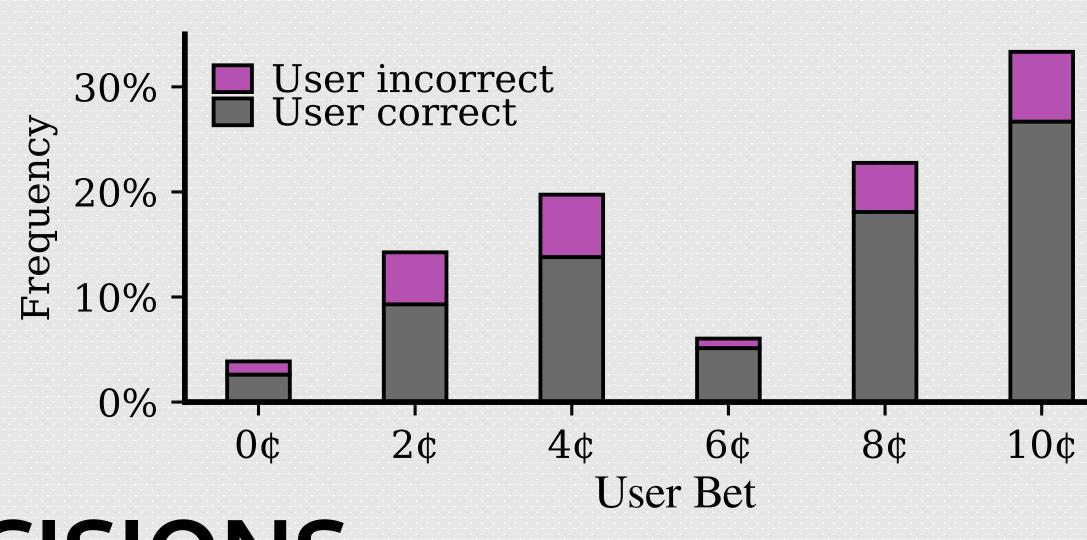
Even 5 miscalibrations are enough to affect user trust long-term

Miscalibration on one domain affect trust also on other domain but mostly on the original one \rightarrow evidence for domain being represented in user mental model



MISCALIBRATIONS AFFECT HUMAN-AI COLLABORATION

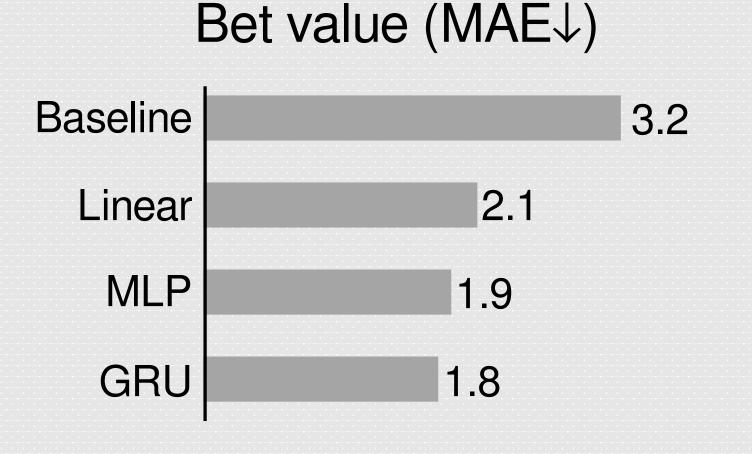


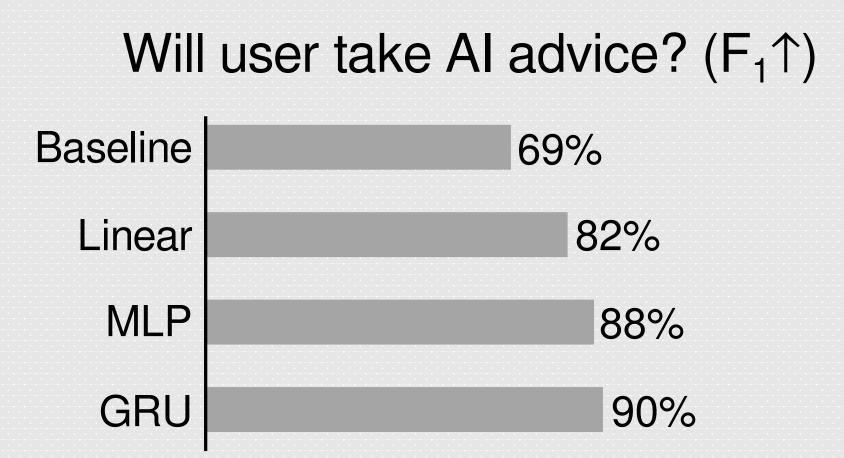


With lower trust users do not accept Al predictions and lower Human-Al team success

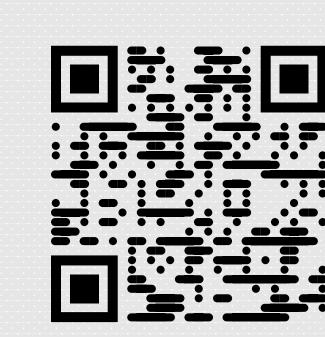
PREDICTING USER DECISIONS

accurate





Predicting user trust and decisions allows AI to adapt its output to improve collaboration; GRU learns how user trust evolves with respect to trust-eroding events.



value)

8¢

6¢

