Some reserchers ‘attack’ object recognition by adding pixles to an image. If people are aware of what exactly makes an AI fail, does that decrease trust in the AI wholly, or selectively only in contexts that cause AI failures

Confidnece Heuristics vs calibration hypothesis

Manipulate confidence independent of accuracy

Price and Stone 2004: Advisors predicted if a stock would go up or down.

Moderate condition: target occurred to half events, not occurred for other half. Extreme set: added 15% confidence to judgements above 50%, and subtracted 15% to judgements below 50%.

Tenney: was a suspect absent or present

Sah: Estimate weight

Conflicts:

Both in agreement:

Both in conflict:

Weight of Advice’ measure used by Yaniv (2004),

Intro:

Two branches of metacognition: appraisals of self-accuracy, and appraisals of other’s accuracy. Relevant in a number of contexts: seeking medical or financial advice, actuaries, trustworthiness of politicians etc. Recently emerging paradigm looks specifically at how people incorporate decisions from AI algorithms to guide decisions. Such decisions entail two parts: accurately choosing the right outcome, and when possible, expressing an appropriate level of confidence in that decision.

A key phenomenon is this overconfidence heuristic and how it’s modulated by calibration hypothesis. As a primer to overconfidence, we’ll start with a paper by Prince & Stone

Key idea, is people’s decision making on a per trial basis influence by calibration