Using Tools From Animal Psychology to Measure Metacognition in Artificial Intelligence

Konstantinos Voudouris
Helmholtz Zentrum Munich, Munich, Germany

Alexi Voudouris
The University of Edinburgh, Edinburgh, United Kingdom

Lucy Cheke
University of Cambridge, Cambridge, United Kingdom

Abstract

Metacognition is the ability to monitor one's cognitive processes, including one's own uncertainty when making a decision. Metacognition has been studied in humans and other animals for several decades, and it is of increasing interest to the Artificial Intelligence (AI) research community too. In this work, we implement a well-established experimental procedure to study whether two classes of AI system can monitor their own uncertainty. We ask deep reinforcement learning agents and vision language models to learn to discriminate two stimuli that vary in similarity, where they are rewarded for correct and punished for incorrect discriminations. By measuring the frequency at which they choose not to make a choice, and analysing how that varies with stimulus similarity, we produce a measure of the degree to which their uncertainty informs their decisions. We find some limited evidence that the AI systems we study monitor their own uncertainty when making risky decisions.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*