Towards metacognitive agents: integrating confidence in sequential decision-making

Baptiste Pesquet (baptiste.pesquet@inria.fr), Frédéric Alexandre *INRIA Bordeaux*

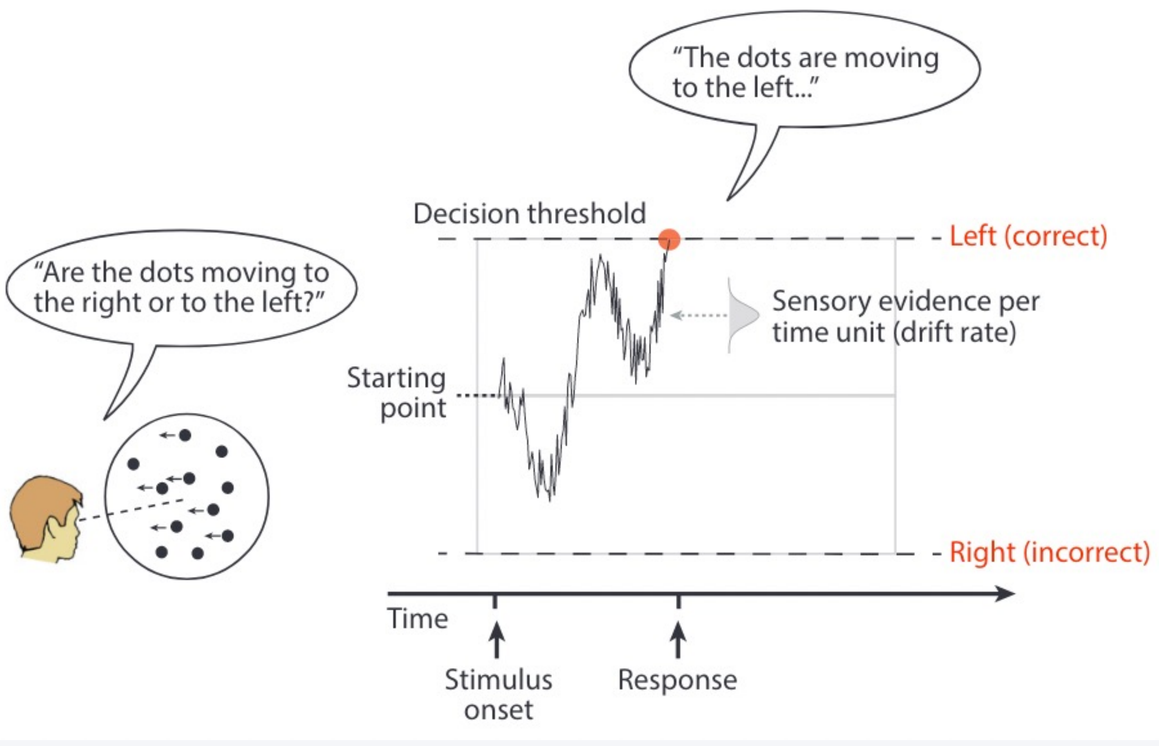
LaBRI

Abstract

In natural cognition, confidence is used to evaluate the quality of decisions and adapt one's behavior to the task at hand. For now, artificial agents lack this kind of metacognitive ability and interact with their environment in a purely reactive way. Inspired by recent findings about the cognitive modeling of confidence, we propose a novel architecture for sequential decision-making. It combines an evidence accumulation model with a metacognitive module that computes and exploits confidence to tune the decision process. The model has been assessed on a perceptual decision-making task, showing promises for more flexible artificial agents and a possible path towards artificial metacognition.

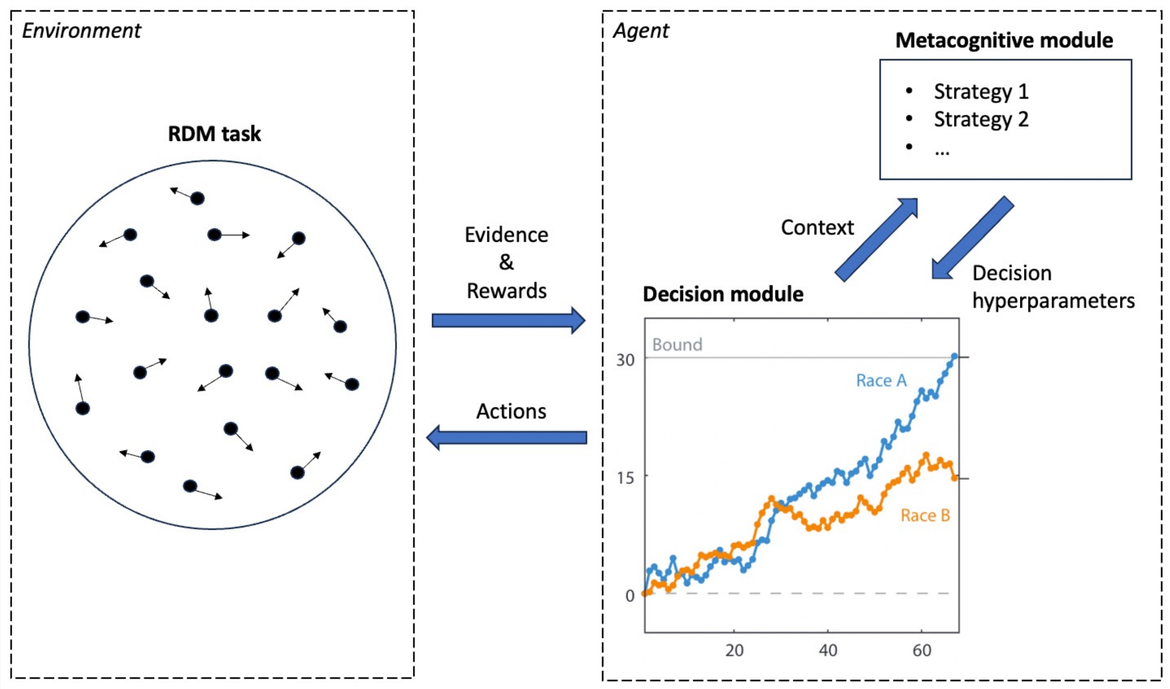
Sequential decision-making

- Evidence Accumulation Models : DDM, LBM, etc
- Speed-Accuracy Tradeoff
- Confidence as subjective estimate of decision quality

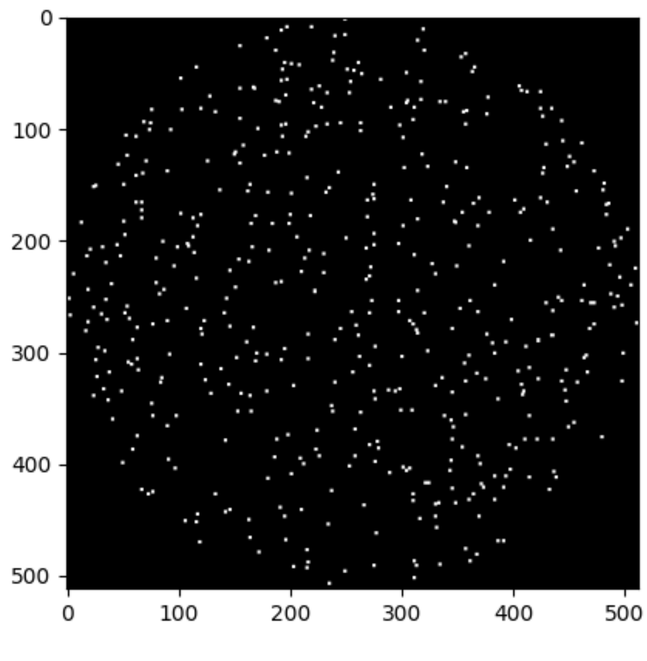


Proposed agent architecture

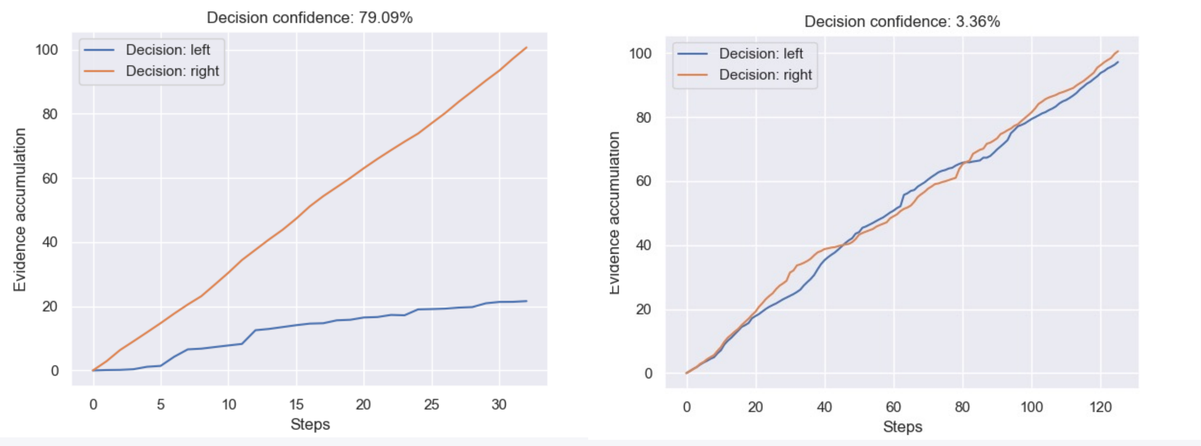
- **Decision module** (EAM) to select the chosen alternative.
- **Metacognitive module to t**une the decision process through adjusting the decision hyperparameters.



Testbed: Random Dot Motion task



Preliminary results



- **Left**: high confidence, short decision process
- **Right**: low confidence, longer decision process

Reference paper

B. Pesquet, F. Alexandre (2025). Towards metacognitive agents: integrating confidence in sequential decision-making, *2025 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium