An Algorithmic Theory of Metacognition in Minds and Machines

Rylan Schaeffer [1]

[1]Department of Computer Science, Stanford University

Introduction

- Humans sometimes choose actions that they themselves can identify as sub-optimal, or wrong, even in the absence of additional information - how?
- **Contribution 1**: We propose an algorithmic theory of metacognition based on a trade-off in reinforcement learning (RL) between value-based RL and policy-based RL.
- **Contribution 2**: We implement a deep Metacognitive Actor Critic (MAC) and show it can detect (some of) its own suboptimal actions without external information
- **Contribution 3**: We establish a novel connection between RL and Bayesian Optimization

Background: Metacognition

- Common experimental paradigm: participants complete task while subjectively evaluating own performance
- Three well-reproduced experimental findings (Fig. 1) include:
- **Hypermetacognitive Sensitivity**: Participants' performance at evaluating themselves can exceed their performance in doing the task
- **Response-Locked Error Related Negativity (ERN)**: Event-related potential distinguishes correct from incorrect actions, too soon to be driven by external input or feedback
- **Dissociability of Decision-Making and Self-Evaluation**: Interventions (pharmacological, lesion, age, TMS) affect subjects' self-evaluation without affecting decision-making and vice versa

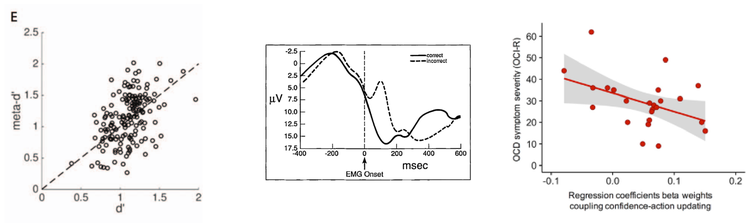


Figure 1. Left: Hypermetacognitive Sensitivity: self-evaluation performance vs task performance . Center: Response-locked ERN in speeded response. Right: OCD severity correlates with action-confidence disagreement.

Background: Reinforcement Learning

- **Value-based** scales poorly to high-dimensional/continuous action spaces:

$$\arg\max_A Q(S_n, A)$$

- **Policy-based** learns slowly due to gradient estimator's high variance:

$$\nabla_\theta \mathbb{E}_{p(\tau;\theta)}[G(\tau)] = \mathbb{E}_{p(\tau;\theta)}\Big[G(\tau)\nabla_\theta \sum_n \log p(A_n|S_n;\theta)\Big]$$

- **Actor-Critic** combines both, stabilizing policy-based Actor's learning with value-based Critic

$$\widetilde{G}(\tau) \stackrel{\text{def}}{=} G(\tau) - \beta\big(Q(\tau) - \mathbb{E}[Q(\tau)]\big)$$

$$\nabla_\theta \mathbb{E}_{p(\tau;\theta)}[G(\tau)] = \mathbb{E}_{p(\tau;\theta)}\Big[\widetilde{G}(\tau)\nabla_\theta \sum_n \log p(A_n|S_n;\theta)\Big]$$

Metacognitive Actor Critic (MAC)

- In Actor-Critic, Critic stabilizes Actor's learning, but does nothing during action selection
- **Why this is bad**: Actor samples action "Drive off the cliff." Critic responds "Driving off the cliff would be bad." Actor-Critic then drives off the cliff.
- **Idea**: Use Critic to help Actor select better actions
- **Approach**: Allow Actor and Critic to interact multiple times within each environment step. Actor iteratively samples hypothetical actions, queries the Critic for its advice, then repeats until (a) satisfied or (b) forced to act by time pressure

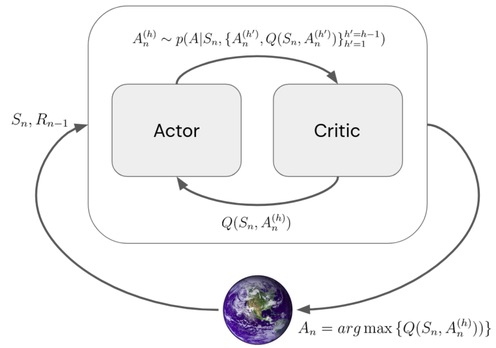


Figure 2. Metacognitive Actor Critic

Algorithm 1: Metacognitive Actor Critic (MAC)

for *Environment Step* $n = 1, 2, ..., N$ **do**
- Environment sends S_n, R_{n-1} to MAC.
- MAC computes $V(S_n)$
- MAC initializes Hypothetical Actions: $\mathcal{A}_n \leftarrow \{\}$
- MAC initializes Hypothetical Actions' Values: $\mathcal{Q}_n \leftarrow \{\}$
- **for** *Hypothetical Evaluation* $h = 1, 2, ..., H$ **do**
  - MAC's Actor constructs a policy: $p(A|S_n, \mathcal{A}_n, \mathcal{Q}_n)$
  - MAC's Actor samples a hypothetical action: $A_n^{(h)} \sim p(A|S_n, \mathcal{A}_n, \mathcal{Q}_n)$
  - MAC's Critic evaluates the hypothetical action: $Q(S_n, A_n^{(h)})$
  - MAC adds hypothetical action to set of hypothetical actions: $\mathcal{A}_n \leftarrow \mathcal{A}_n \cup \{A_n^{(h)}\}$
  - MAC adds hypothetical action's value to set of hypothetical actions' values: $\mathcal{Q}_n \leftarrow \mathcal{Q}_n \cup \{Q(S_n, A_n^{(h)})\}$
- **end**
- MAC chooses real action from hypothetical actions e.g. $A_n \stackrel{\text{def}}{=} \arg\max\Big\{Q(S_n, A_n^{(h)})\Big\}$
- MAC sends real action A_n to Environment

end

- We posit the MAC can explain metacognitive experimental findings
- **Hypermetacognitive Sensitivity**: Critic can advise against, but not advocate for, actions
- **Response-Locked Error Related Negativity**: Assuming time pressure only permits 1 round of interactions, Critic can detect error but not quickly enough for Actor to sample a new action
- **Dissociability of Decision-Making and Self-Evaluation**: Interfering with Actor affects decision-making without affecting self-evaluation, and interfering with Critic affects self-evaluation without affecting decision-making

Experimental Results

- LSTM-based MAC is trained to perform a psychophysics-inspired speeded-response two-alternative forced-choice task
- MAC outputs two quantities: (1) binary action indicating signal side (left or right), and (2) self-evaluation $Q(S_n, A_n) - V(S_n)$

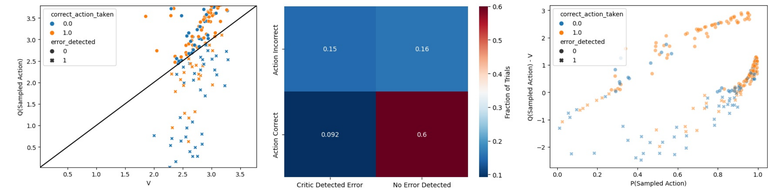
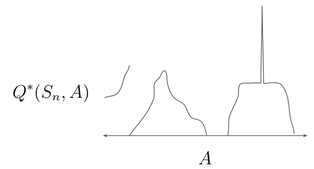


Figure 3. Left: By defining a subjectively-perceived "error" whenever $Q(S_n, A_n) - V(S_n) < 0$, Critic recognizes a significant fraction of actions sampled by the Actor as incorrect. Center: Critic detects suboptimal actions better than chance. Right: Many erroneous actions were sampled with high probability by the Actor, showing that the Critic uses information the Actor has not yet learnt.

Connection to Bayesian Optimization

- Bayesian optimization studies gradient-free global optimization of a black-box scalar function



- Bayesian Optimization uses two components
- 1) a **surrogate function** $Q(S_n, \cdot)$, which learns to emulate true objective function $Q^*(S_n, \cdot)$
- 2) an **acquisition function** $p(A|S_n;\theta)$, which determines where to sample next
- MAC Actor $\Leftrightarrow$ acquisition function and MAC Critic $\Leftrightarrow$ surrogate function

References

[1] Stephen M Fleming and Nathaniel D Daw.
Self-evaluation of decision-making: A general bayesian framework for metacognitive computation.
Psychological review, 124(1):91, 2017.

[2] Peter I Frazier.
A tutorial on bayesian optimization.
arXiv preprint arXiv:1807.02811, 2018.

[3] William J Gehring, Brian Goss, Michael GH Coles, David E Meyer, and Emanuel Donchin.
A neural system for error detection and compensation.
Psychological science, 4(6):385–390, 1993.

[4] Matilde M Vaghi, Fabrice Luyckx, Akeem Sule, Naomi A Fineberg, Trevor W Robbins, and Benedetto De Martino.
Compulsivity reveals a novel dissociation between action and confidence.
Neuron, 96(2):348–354, 2017.