ARTIFICIAL INTELLIGENCE

Toward high-performance, memory-efficient, and fast reinforcement learning—Lessons from decision neuroscience

Jee Hang Lee[1,2]*, Ben Seymour[3,4,5]*†, Joel Z. Leibo[6], Su Jin An[1], Sang Wan Lee[1,2,7]†

Recent insights from decision neuroscience raise hope for the development of intelligent brain-inspired solutions to robot learning in real dynamic environments full of noise and unpredictability.

Recent successes in building agents with superhuman performance have led to reinforcement learning (RL), becoming a dominant theoretical framework to understand decision-making through interaction with the world (*1*). However, recent RL algorithms still have major limitations, such as lack of the ability to develop goal-directed policies or reliance on large amounts of experience to learn (*2*). These limits impede the ability to rapidly adapt in dynamic environments where tasks or contexts frequently change.

In contrast, humans have a remarkable ability to rapidly adapt to environmental changes with limited experience. Recent findings in decision neuroscience suggest that the brain uses not only multiple control systems for RL but also a flexible metacontrol mechanism to select among control options, each different trait associated with prediction performance, cognitive load, and learning speed (*3*). Understanding how the brain implements these options could lead to brain-inspired RL algorithms that can work in real control problems for robots (*4*). Here, we discuss recent findings on human RL that may address several key challenges in robotics: performance-efficiency-speed trade-offs, conflicting demands in multirobot settings, and the exploration-exploitation dilemma.

First, accumulating evidence in decision neuroscience indicates that humans take advantage of two different behavior control strategies: (i) stimulus-driven habitual and (ii) goal-directed cognitive control (*3*). Habitual control is automatic and fast, despite being fragile in a volatile environment, and is well accounted for by model-free RL, which incrementally learns the values of actions through trial and error without a model of the environment. Conversely, goal-directed control can rapidly adapt to changes in the environment, but it is cognitively demanding. It guides actions by learning a model of the environment and uses this knowledge base to quickly adapt to changes in environmental structure, such as learning latent (hidden) causes within state-action space.

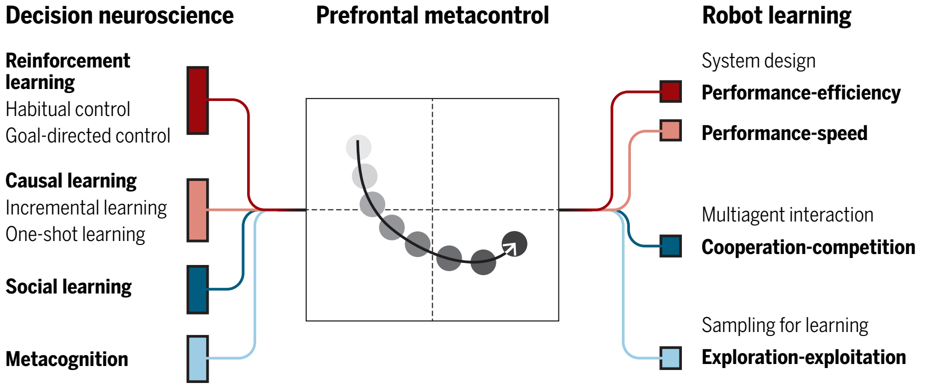
This computational distinction between model-based and model-free RL suggests an inevitable compromise between them. Model-free RL is slow to learn but is fast to achieve a goal once a policy is learned and automatized. Model-based RL provides more accurate predictions than model-free RL in general but is computationally much heavier. Each strategy provides a complementary solution regarding accuracy, speed, and cognitive load, highlighting a trade-off between prediction performance and computational efficiency.

Second, RL algorithms usually require a large amount of experience to adequately learn causal relationships in the presence of different environmental factors (incremental learning). Humans, however, learn fast—often after a single exhibition of an event never experienced before (“one-shot learning”) (*5*). Recent neuroscience studies (*5*, *6*) found that, when interactions with the environment are limited, humans have a strong tendency to increase their learning rates; they strive for quickly making sense of unknown parts of the environment, even when this compromises safety. These results suggest that the brain directly implements computation to find a trade-off between performance and speed.

Third, accumulating evidence supports the notion that the prefrontal cortex implements metacontrol to flexibly choose between different learning strategies, such as between model-based and model-free RL (*7*, *8*) and between incremental and one-shot learning (*5*). In a new environment, metacontrol accentuates performance by favoring model-based RL. Because this is computationally expensive, the brain resorts to model-free RL when it finds little benefit from further learning: Either the environment is sufficiently stable to make precise predictions or highly unstable such that predictions from model-based RL become less reliable than those from model-free RL. In other situations, metacontrol prioritizes speed. When the uncertainty in the estimated cause-effect relationships is high, the brain tends to transition to one-shot learning to quickly resolve uncertainty in predicting outcomes. However, when the agent is equally uncertain about all possible causal relationships, it resorts to incremental learning to ensure safe learning. Together, they suggest that brain-like metacontrol can deal with performance-efficiency-speed trade-offs.

Fourth, human RL may account for social phenomena that have been important in human evolution. In human societies where multiple agents are interacting, there are social dilemmas that have partially competitive and partially aligned incentives (*9*). Approaches using model-based RL successfully achieve cooperation in more complex temporally extended settings [e.g., (*10*, *11*)]. These models often work in two stages: First, there is a planning stage where the agent uses its model of the game’s rules to simulate a large number of games with itself and learns separate cooperation and defection policies by

[1] Department of Bio and Brain Engineering, KAIST, Daejeon, Republic of Korea. [2] KAIST Institute for Health Science and Technology, Daejeon, Republic of Korea. [3] Computational and Biological Learning Laboratory, Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK. [4] Brain Information Communication Research Laboratory Group, Advanced Telecommunications Research Institute International, Kyoto, Japan. [5] Center for Information and Neural Networks, National Institute of Information and Communications Technology, 1-4 Yamadaoka, Suita, Osaka 565-0871, Japan. [6] DeepMind, London, UK. [7] KAIST Institute for Artificial Intelligence, Daejeon, Republic of Korea.
*These authors contributed equally to this work.
†Corresponding author. Email: bjs49@cam.ac.uk (B.S.); sangwan@kaist.ac.kr (S.W.L.)



Brain-inspired solutions to robot learning. Neuroscientific views on various aspects of learning and cognition converge and create a new idea called prefrontal metacontrol, which can inspire researchers to design learning agents that can address various key challenges in robotics such as performance-efficiency-speed, cooperation-competition, and exploration-exploitation trade-offs.

CREDIT: A. KITTERMAN/SCIENCE ROBOTICS

independently learning toward both selfish and cooperative objectives. Then, in the execution phase, a tit-for-tat policy is constructed and applied using the previously learned cooperate and defect policies. Other approaches have sought to break down the strict separation between planning and execution stages and instead work in a fully online manner, such as the LOLA (Learning with Opponent-Learning Awareness) algorithm (*12*). In addition to assuming perfect knowledge of the game rules, this model also assumes that agents can differentiate through one another's learning process. This allows agents to learn to teach because they can isolate the effects of their actions on the learning of others.

Last, conventional RL algorithms tend to be optimistic (or overconfident), especially when sampling from a part of the environment they have not sufficiently learned. Learning without an estimate of prediction performance may lead to suboptimal policies (local minima problem), especially in complex and dynamic environments.

Humans appear to get around this problem by using metacognition—the ability to evaluate one's own performance to estimate a level of confidence and/or uncertainty (*13*, *14*). For example, low task difficulty or low environmental noise would make the learning agent confident, leading to more decisive actions, whereas losing confidence would lead to a more cautious and defensive strategy (*15*). Metacognitive learning thus allows for rapid adaptation to the context change while maintaining robustness against environmental noise. Such a strategy has potential for augmenting robot decision-making in several ways—for instance, in resolving exploration-exploitation trade-offs by overseeing how lack of confidence should drive the desire to learn.

In conclusion, the integration of findings from human decision neuroscience can offer valuable insights into action control systems for robots, leading to safer, more capable, and more efficient learning. Such an interdisciplinary approach should also yield insights for neuroscience, providing a robust test base for developing new theories of human decision computation.

REFERENCES AND NOTES

1. D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G.van den Driessche, T. Graepel, D. Hassabis, Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
2. B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
3. N. D. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
4. S. Elfwing, B. J. Seymour, Parallel reward and punishment control in humans and robots: Safe reinforcement learning using the MaxPain algorithm, in *7th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EpiRob 2017)* (2018), vol. 2018, pp. 140–147.
5. S. W. Lee, J. P. O'Doherty, S. Shimojo, Neural computations mediating one-shot learning in the human brain. *PLOS Biol.* **13**, e1002137 (2015).
6. T. E. J. Behrens, M. W. Woolrich, M. E. Walton, M. F. S. Rushworth, Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).
7. S. W. Lee, S. Shimojo, J. P. O'Doherty, Neural computations underlying arbitration between model-based and model-free learning. *Neuron* **81**, 687–699 (2014).
8. J. X. Wang, Z. Kurth-Nelson, D. Kumaran, D. Tirumala, H. Soyer, J. Z. Leibo, D. Hassabis, M. Botvinick, Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**, 860–868 (2018).
9. P. Kollock, Social dilemmas: The anatomy of cooperation. *Annu. Rev. Soc.* **24**, 183–214 (1998).
10. M. Kleiman-Weiner, M. K. Ho, J. L. Austerweil, M. L. Littman, J. B. Tenenbaum, Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction, in *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (2016), pp. 1679–1684.
11. A. Lerer, A. Peysakhovich, Maintaining cooperation in complex social dilemmas using deep reinforcement learning, https://arxiv.org/abs/1707.01068 (2017).
12. J. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, I. Mordatch, Learning with opponent-learning awareness, in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (International Foundation for Autonomous Agents and Multiagent Systems, 2018), pp. 122–130.
13. S. M. Fleming, H. C. Lau, How to measure metacognition. *Front. Hum. Neurosci.* **8**, 443 (2014).
14. Y.-L. Boureau, P. Sokol-Hessner, N. D. Daw, Deciding how to decide: Self-control and meta-decision making. *Trends Cogn. Sci.* **19**, 700–710 (2015).
15. P. Domenech, E. Koechlin, Executive control and decision-making in the prefrontal cortex. *Curr. Opin. Behav. Sci.* **1**, 101–106 (2015).

Acknowledgments: B.S. is funded by the Wellcome Trust (no. 097490) and Arthritis Research UK (no. 21357). J.H.L., S.J.A., and S.W.L. are supported by (i) the ICT R&D program of MSIP/IITP (no. 2016-0-00563, Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion), (ii) the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean government (no. 2017-0-00451, Development of BCI based Brain and Cognitive Computing Technology for Recognizing User's Intentions using Deep Learning), (iii) IITP grant funded by the Korean government (MSIT) (no. 2018-0-00677, Development of Robot Hand Manipulation Intelligence to Learn Methods and Procedures for Handling Various Objects with Tactile Robot Hands), (iv) Samsung Research Funding Center of Samsung Electronics under project number SRFC-TC1603-06, and (v) the research fund of the KAIST (Korea Advanced Institute of Science and Technology; grant code: G04150045).

10.1126/scirobotics.aav2975

Citation: J. H. Lee, B. Seymour, J. Z. Leibo, S. J. An, S. W. Lee, Toward high-performance, memory-efficient, and fast reinforcement learning—Lessons from decision neuroscience. *Sci. Robot.* **4**, eaav2975 (2019).

Science Robotics

Toward high-performance, memory-efficient, and fast reinforcement learning––Lessons from decision neuroscience

Jee Hang Lee, Ben Seymour, Joel Z. Leibo, Su Jin An and Sang Wan Lee

Sci. Robotics **4**, eaav2975.
DOI: 10.1126/scirobotics.aav2975

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title *Science Robotics* is a registered trademark of AAAS.