Learning the value of control with Deep RL

Kai Sandbrink (kai.sandbrink@psy.ox.ac.uk)
University of Oxford Department of Experimental Psychology,
Anna Watts Building, Woodstock Rd, Oxford OX2 6GG, United Kingdom

Christopher Summerfield (christopher.summerfield@psy.ox.ac.uk)
University of Oxford Department of Experimental Psychology,
Anna Watts Building, Woodstock Rd, Oxford OX2 6GG, United Kingdom

Abstract

Agents need to make decisions based on both the value of specific actions and their probability of success. In a reinforcement learning setting, these two quantities are often conflated into a single value estimate. Here, we train deep RL models on an ”Observe-v-Bet” task in which the agent explicitly chooses on every trial whether to explore or exploit. We vary the probability that chosen actions are successfully translated into their desired consequences (i.e. the level of control, with failure resulting in the selection of a random action). In contrast with agents trained solely using a reward signal, those additionally trained using action-prediction error (APE) signals to structure their internal hidden representations successfully learn adaptive policies in which they spend more time exploring in high-control settings and increasing their control in low-control settings. We further show that providing fallacious APE signals results in behavioral pathologies associated with a number of psychiatric conditions. These results provide a model for how control can be sensed in a reinforcement-learning setting and a neural network account of psychiatric diseases.

Keywords: RL; auxiliary teaching signals; control; agency; computational psychiatry

Introduction

Consider playing a game of tennis: In calm weather, it pays to play more aggressively; on a windy day, more conservatively. In order to make optimal decisions, we thus need to constantly consider our level of control over the environment. Theories of motor control posit the existence of a forward model of motor learning, in which predictions are formed based on selected actions and these then retroactively compared to observed outcomes (Miall & Wolpert, 1996; Blakemore, Frith, & Wolpert, 2001; Franklin & Wolpert, 2011). Meanwhile, experimental work has located action-prediction errors (APEs) in reinforcement learning (RL) circuitry in mice (Greenstreet et al., 2022), striatal dopaminergic signals that fire when an action is taken different from the expected one for a state (Bogacz, 2020). In this study, we investigate how these APEs could help agents learn optimal strategies for trading off exploration and exploitation, or learning a ”self-care” policy. Endowing agents with control (or ”agency,” see Haggard & Chambon, 2012; Chambon, Sidarus, & Haggard, 2014) in this way supports cognitive decision-making in an RL framework.

Methods

Observe-Bet-Efficacy

In the classic Observe-v.-Bet task, participants choose between two bandits, exactly one of which pays out on every time step (Tversky & Edwards, 1966; Navarro, Newell, & Schulze, 2016; Blanchard & Gershman, 2018). Participants choose between ”betting” on one of two lights (bandit arms) or ”observing.” On bet trials, if they placed their bet on the correct light, participants earn a reward that pays out at the end of the episode, but receive no immediate feedback. On observe trials, participants see which light would have paid out in that round, but receive no reward. We created a variant of this task in which each bet action is flipped (to the other option) with probability $p(\text{random}) = 1 - \xi$, governed by a parameter efficacy ξ that is constant in an episode but changes between them. In task 1, participants need to trade off exploration with exploitation. In task 2, we introduce a ”sleep action” which participants can take to increase the efficacy by 10% for the remainder of the episode. In both tasks, episodes have 50 steps, lights pay out with a bias (0.4 for task 1 and 0.49 for task 2), and the volatility (payout reversal probability) is 10%.

Integrating APEs into reward-based learning

We implement the REINFORCE algorithm (Sutton & Barto, 2018) as a recurrent neural network with a 46-unit LSTM layer, a 24-unit fully-connected layer, a 4-unit linear action readout, and a softmax that determines the final agent policy. The agent receives the action a_t it took at time t as well as the action that it initially chose before the possible randomization call $\hat{a}_t$ as input. Critically, we also introduce a single additional linear unit that reads out from the LSTM to predict $p(\text{switch}) = \frac{1}{2} \cdot p(\text{random})$. This node is trained using a learning algorithm we call *temporal action-difference learning* because it uses APEs $\left(\delta_{\text{APE},t} = \begin{cases} 0, & \text{if } \hat{a}_t = a_t \\ 1, & \text{otherwise} \end{cases}\right)$ to learn a sense of efficacy in the same way that temporal-difference learning uses reward-prediction errors contribute to build an estimate of value (Schultz, 1998; Glimcher, 2011), i.e.:

$$\hat{\xi}'_{t+1} \leftarrow \hat{\xi}'_t + \alpha \cdot ((1 - \delta_{\text{APE},t}) - \hat{\xi}'_t) \qquad (1)$$

where α is a decaying learning rate. In the limit, the learned efficacy-sense $\hat{\xi}'$ will converge to the switch probability and is therefore directly related to the original efficacy since $\frac{1-\xi}{2} = p(\delta_{\text{APE},t}) = p(\text{switch})$.

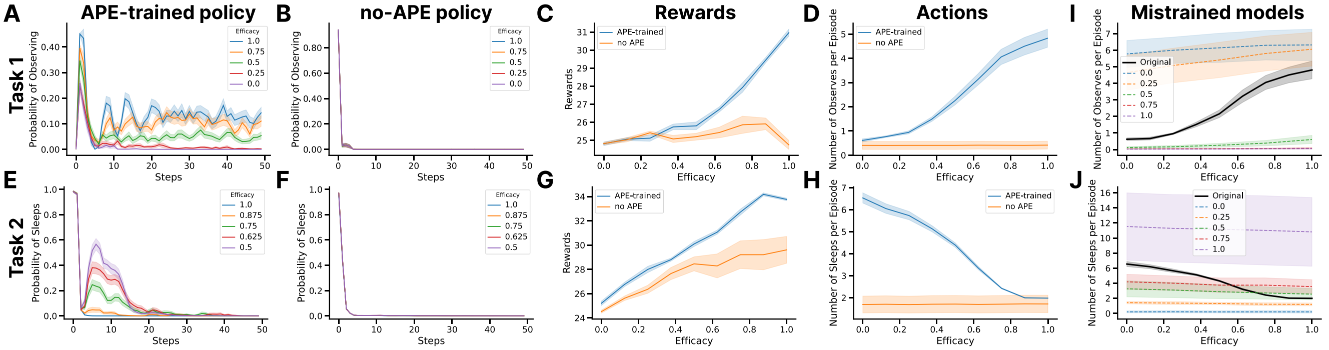

Figure 1: Behavior of the neural network models. For all panels: thick lines mean, shaded area standard error, 100 episodes for each efficacy level. **A.** Policy over an episode for a sample APE-trained model **B.** The same plot for the non-APE-trained models. **C.** Average rewards per efficacy level across five model instantiations. **D.** Same plot as with C but showing number of observe actions. **E-H.** Same plots as A-D, but showing task 2 instead of task 1 and sleep actions instead of observe actions. **I.** Number of observe actions for models mistrained to different efficacy levels. **J.** Same plot but for task 2 and number of sleep actions.

We train both the APE-based and the solely-reward-based networks over a hierarchical task structure over ξ using meta-RL (Wang et al., 2016). We meta-train over the set $\xi \in [0, 0.33] \cup [0.66, 1]$ and test on discrete values across the entire range $\xi \in \{0.125k \,|\, 0 \leq k \leq 8, k \in \mathbb{Z}\}$. We train 5 instantiations per type over 500k episodes of 50 steps, annealing entropy regularization to 0 geometrically over 150k episodes.

Results

APE-trained networks learn adaptive solutions

At test, the APE-trained networks outperform the no-APE networks on both tasks, achieving rewards of 26.83 ± 0.26 compared with 25.36 ± 0.47 averaged across 100 repetitions for each value in the test set for task 1 (mean ± stderr, n=5, greater t(4)=5.50, p=2.85e-4) and 30.09±0.13 compared with 27.59 ± 0.81 for task 2 (t(4)=2.72, p=1.32e-2). The efficacy readout node in the APE-trained networks also learns to represent efficacy well, achieving MSE of 3.41e-2 ± 1e-4 across the test set for task 1 and 3.69e-2 ± 4e-4 for task 2.

To understand the behavioral differences between the APE-trained and the control networks, we examine the policy of a randomly-selected model instantiation for each network type across efficacy levels $\xi \in \{0, 0.25, 0.5, 0.75, 1\}$. Figure 1A illustrates the adaptive policy found by the APE-trained network for task 1, in which the network starts a trial by betting on a random arm and then adjusts its observation rate to the level of efficacy it perceives over time, in accordance with theoretical results that exploration is worth more in high-control settings (Huys & Dayan, 2009). In contrast, as can be seen in Figure 1B, the no-APE network always exhibits a stereotyped pattern which begins with a single observation before betting for the remainder of the episode. These results hold across all model instantiations (Figure 1C) and result for a higher level of reward for the APE-trained network in high efficacy settings (Figure 1D). Similar results hold for task 2 for the amount of time spent sleeping (Figure 1E-H).

Mistraining efficacy causes behavioral pathologies

We performed manipulations to investigate the importance of the learned efficacy readout to the performance of the APE-trained networks. To manipulate the population-level representation of efficacy that emerges in the LSTM (detected through a decoding analysis and through ablations of the efficacy node), we selectively unfroze the weights from input to LSTM and from LSTM to efficacy readout of converged APE-trained networks and continued training the networks for another 100k episodes on false APEs generated from binomial distributions corresponding to fixed probabilities. Performance drops across mistraining for all efficacy values, with none of the mistrained networks performing better than 26.0 across the test set for task 1 and 27.5 for task 2. The policies of the mistrained networks are similar to, and in some cases more extreme than, the policies the correctly-trained networks exhibit at the level of efficacy that was used in mis-training (see Figure 1I for task 1 and Figure 1J for task 2).

Discussion

Task 1 focuses on explore-and-exploit, normatively impacted by control (Huys & Dayan, 2009), and task 2 on allocating the correct amount of effort to control as in (learning) expected-value of control (Masís, Musslick, & Cohen, 2021; Shenhav, Botvinick, & Cohen, 2013). Failures to make decisions based on one's level of control is a behavioral correlate of mental diseases such as depression (Seligman, 1975). In the future, we hope our study of sensed control's impact can be used to model disease and lay groundwork for potential interventions.

Acknowledgments

We thank Alireza Modirshanechi, Johanni Brea, Jessica Thompson, Wulfram Gerstner, and Marion Rouault for enlightening discussions. Funding: K.J.S.: Cusanuswerk Doctoral Research Fellowship (German Ministry of Education and Research). C.S.: ERC Consolidator Grant 725937.

References

Blakemore, S.-J., Frith, C. D., & Wolpert, D. M. (2001, July). The cerebellum is involved in predicting the sensory consequences of action. *NeuroReport*, *12*(9), 1879.

Blanchard, T. C., & Gershman, S. J. (2018, February). Pure correlates of exploration and exploitation in the human brain. *Cognitive, Affective & Behavioral Neuroscience*, *18*(1), 117–126. doi: 10.3758/s13415-017-0556-2

Bogacz, R. (2020, July). Dopamine role in learning and action inference. *eLife*, *9*, e53262. doi: 10.7554/eLife.53262

Chambon, V., Sidarus, N., & Haggard, P. (2014). From action intentions to action effects: How does the sense of agency come about? *Frontiers in Human Neuroscience*, *8*.

Franklin, D. W., & Wolpert, D. M. (2011). Computational mechanisms of sensorimotor control. *Neuron*, *72*(3), 425–442. doi: 10.1016/j.neuron.2011.10.006

Glimcher, P. W. (2011, September). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*(supplement_3), 15647–15654. doi: 10.1073/pnas.1014269108

Greenstreet, F., Vergara, H. M., Pati, S., Schwarz, L., Wisdom, M., Marbach, F., ... Stephenson-Jones, M. (2022, September). *Action prediction error: A value-free dopaminergic teaching signal that drives stable learning.* bioRxiv. doi: 10.1101/2022.09.12.507572

Haggard, P., & Chambon, V. (2012, May). Sense of agency. *Current Biology*, *22*(10), R390-R392. doi: 10.1016/j.cub.2012.02.040

Huys, Q. J. M., & Dayan, P. (2009, December). A Bayesian formulation of behavioral control. *Cognition*, *113*(3), 314–328. doi: 10.1016/j.cognition.2009.01.008

Masís, J. A., Musslick, S., & Cohen, J. (2021). The Value of Learning and Cognitive Control Allocation. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Miall, R. C., & Wolpert, D. M. (1996, November). Forward Models for Physiological Motor Control. *Neural Networks*, *9*(8), 1265–1279. doi: 10.1016/S0893-6080(96)00035-4

Navarro, D. J., Newell, B. R., & Schulze, C. (2016, March). Learning and choosing in an uncertain world: An investigation of the explore-exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77. doi: 10.1016/j.cogpsych.2016.01.001

Schultz, W. (1998, July). Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology*, *80*(1), 1–27. doi: 10.1152/jn.1998.80.1.1

Seligman, M. E. P. (1975). *Helplessness: On depression, development, and death*. New York, NY, US: W H Freeman/Times Books/ Henry Holt & Co.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013, July). The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function. *Neuron*, *79*(2), 217–240. doi: 10.1016/j.neuron.2013.07.007

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA: A Bradford Book - MIT Press.

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, *71*, 680–683. doi: 10.1037/h0023123

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., ... Botvinick, M. (2016). *Learning to reinforcement learn* (No. arXiv:1611.05763). arXiv.