Learning the value of control with Deep RL

Kai Sandbrink, Laurence Hunt, Christopher Summerfield

Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom

1 Sensing control is critical for interaction with the environment

Control in Behavior and Brains:

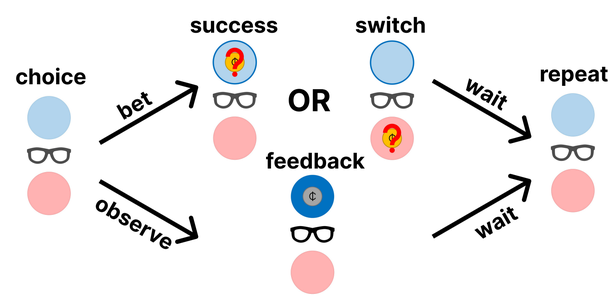
- We adopt different strategies when playing tennis in windy and clear conditions ⇒ our ability to execute actions changes our choices
- This "sense of agency" has been posited to arise from Wolpertian forward models or fluidity of action selection (Haggard and Chambon, 2012)

Reinforcement Learning :

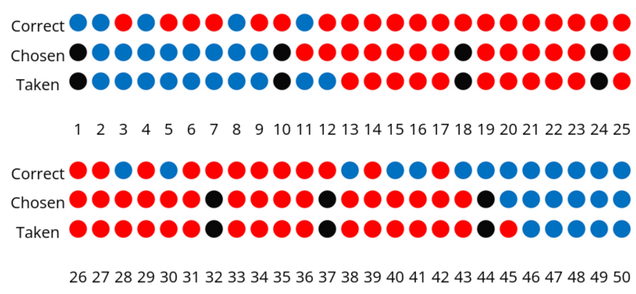
- Control-based measures have primarily been treated as a source for intrinsic motivation treated like curiosity (Oudeyer and Kaplan, 2009)
- ***Can we instead integrate a sense of control into RL without perturbing optimal reward-based solutions?*** Recent ML studies have begun studying control-based solutions to avoiding interference (Yang et al. 2023)

2 Observe-Bet-Efficacy: A novel control integration paradigm

Trial Struture:



Feedback (After Episode):



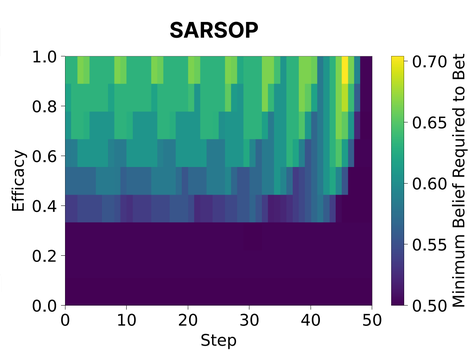
Task 1:

OBE as above

Task 2:

Participants have an additional sleep action that increases ξ by 0.1

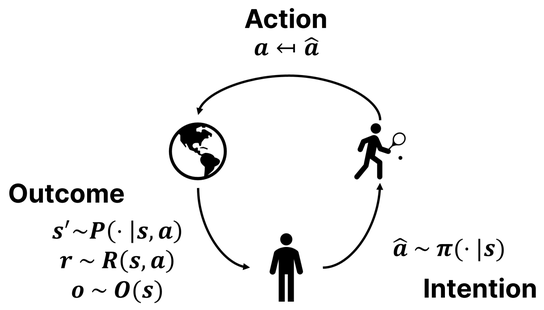
3 The optimal explore-exploit-(sleep) balance depends on control



Tversky et al. *J Exp Psychol* (1966)
Navarro et al. *Cogn Psychol* (2016)
Huys and Dayan *Cognition* (2009)

4 How do we sense control?

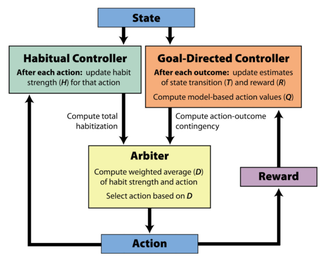
Credit assignment when action selection is noisy:



Selected Action

Feedback	Intended	Other
Good	$\Delta \gg 0$	$\Delta < 0$
Bad	$\Delta \ll 0$	$\Delta > 0$

Diverse roles of dopamine in RL:



Reproduced from: Miller et al. *Psych Rev* (2019)

5 A Deep RL model for the cognitive role of action error signals

Action-Prediction Errors (APEs):

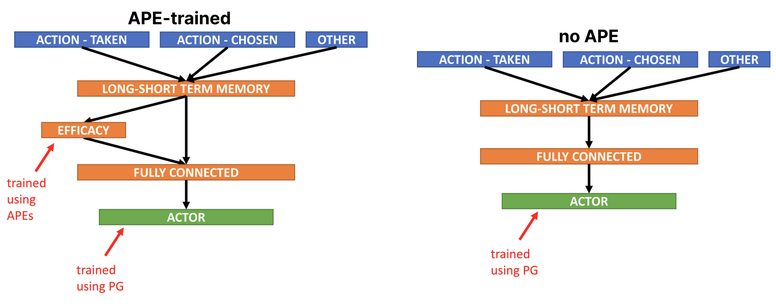
$$\delta_{\mathrm{APE,t}} = \begin{cases} 0 \text{ if } a_t = \widehat{a_t} \\ 1 \text{ otherwise} \end{cases}$$

Greenstreet et al. *BioRXiv* (2022)
Bogacz et al. *eLife* (2020)

Temporal Action-Difference Learning:

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

Model Architectures:

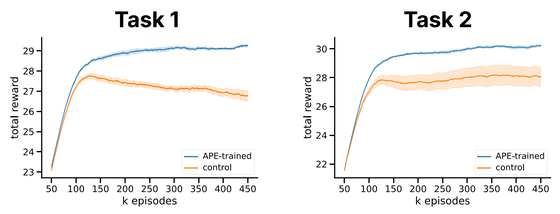


Technical Details:

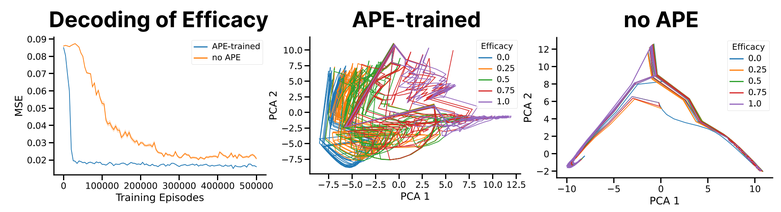
- Inputs: Intended and chosen action, time left in trial, feedback (if observe), reward tally and flag (at start of new episode)
- Architecture: 48 units in LSTM, 24 units in fully-connected, 1 linear efficacy readout unit (if APE-trained), softmax action selection
- Training: REINFORCE with baseline, 500k training episodes, entropy regularization annealed over 150k

6 APE-trained networks outperform the ones without APE by learning a policy that adapts to the different levels of efficacy

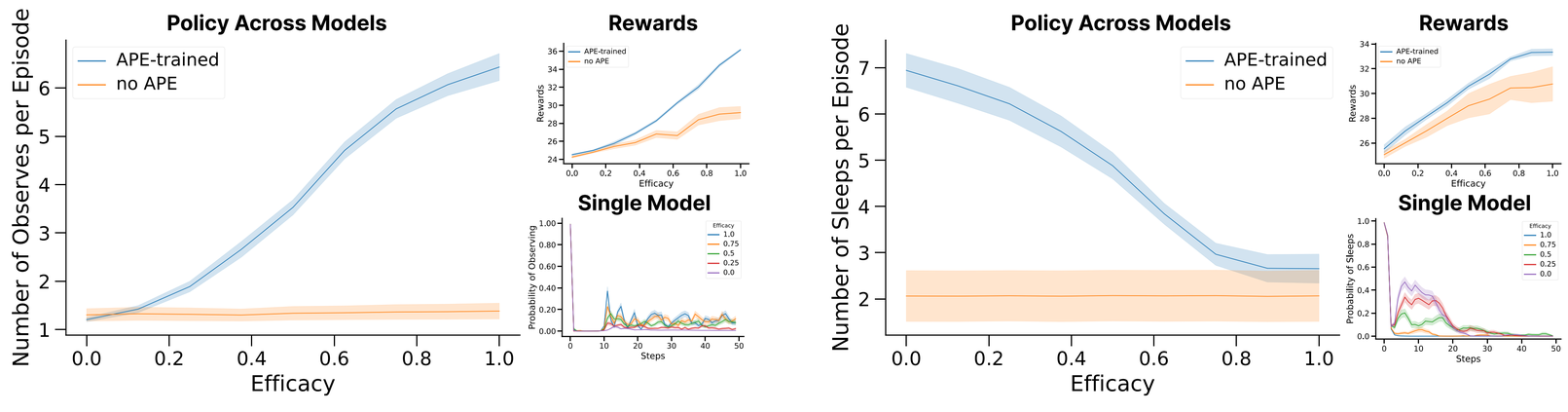
Learning



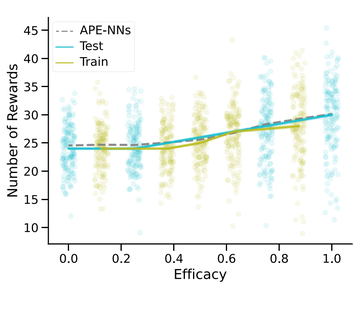
Task 1 Neural Representations



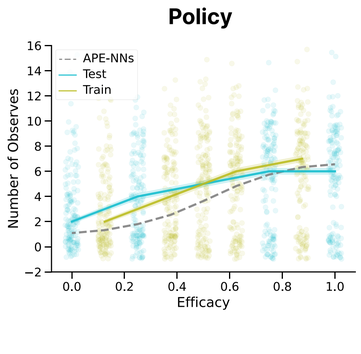
Policy



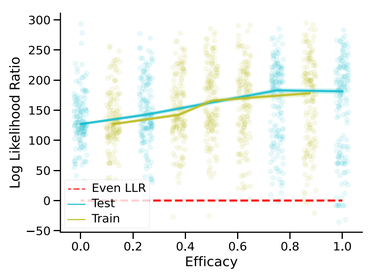
Rewards



Policy



Likelihood APE / no-APE



149 participants

Practice Task 1 (Day 1):

Train
Text and color cue
{0, 0.25, 0.75, 1}

Test
No signal
{0.125, 375, 0.5, 0.725, 0.875}

Practice Task 1 (Day 2):

Train
Color cue
{0.125, 375, 0.5, 0.725, 0.875}

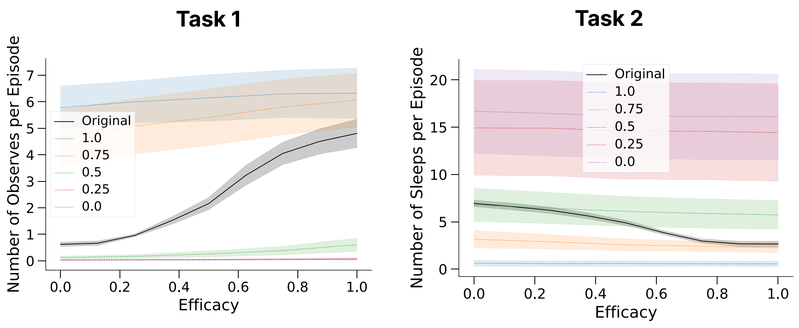
Test
No signal
{0, 0.25, 0.75, 1}

Practice Task 1 (Day 1):

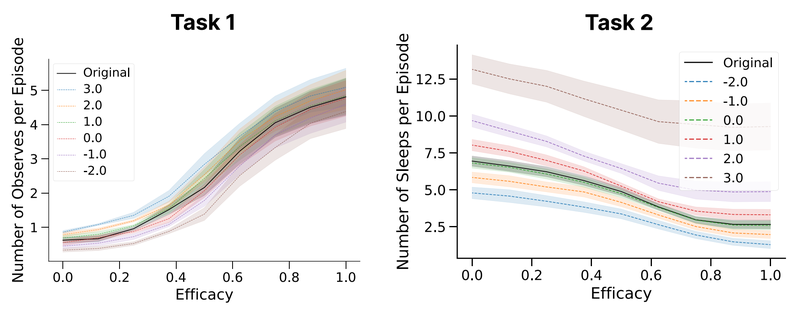
Train
Color cue
{0.125, 375, 0.5, 0.725, 0.875}

Test
No signal
{0, 0.25, 0.75, 1}

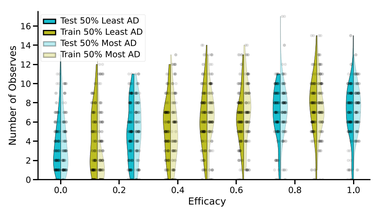
Population Mistraining:



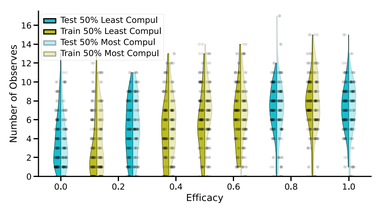
Single-Node Ablation:



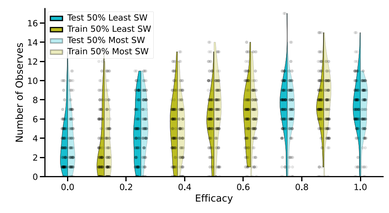
Anxiety & Depression



Compulsivity



Social Withdrawal



7 Human behavior closely matches that of the APE-trained networks

8 Induced errors in the population control representation simulate disorders

9 Outlook: Studying transdiagnostic psychiatric symptoms

Acknowledgments:
We thank Alireza Modirshanechi, Johanni Brea, Jessica Thompson, Brian Christian, Wulfram Gerstner, and Marion Rouault for enlightening discussions.

Funding:
K.J.S.: Cusanuswerk Doctoral Research Fellowship (German Ministry of Education and Research), L.H.: Royal Society/Wellcome Sir Henry Dale Fellowship, C.S.: ERC Consolidator Grant 725937

CCN Paper