Neural Prioritisation of Past Solutions Supports Generalisation

Sam Hall-McMaster (sam_hall-mcmaster@fas.harvard.edu)
Dept. of Psychology and Center for Brain Science, Harvard; Max Planck Institute for Human Development

Momchil S. Tomov (mtomov@g.harvard.edu)
Dept. of Psychology and Center for Brain Science, Harvard; Motional AD LLC

Samuel J. Gershman* (gershman@fas.harvard.edu)
Dept. of Psychology and Center for Brain Science, Harvard; Center for Brains, Minds and Machines, MIT

Nicolas W. Schuck* (nicolas.schuck@uni-hamburg.de)
Institute of Psychology, University of Hamburg; Max Planck Institute for Human Development; Max Planck UCL Center for Computational Psychiatry and Ageing Research

*Equal contribution

Abstract:

How do we decide what to do in new situations? One way to solve this dilemma is to reuse solutions developed for other situations. There is now some evidence that a computational process capturing this idea – called successor features & generalised policy improvement – can account for how humans transfer prior solutions to new situations. Here we asked whether a simple formulation of this idea could explain human brain activity in response to new tasks. Participants completed a multi-task learning experiment during fMRI (n=40). The experiment included training tasks that participants could use to learn about their environment, and test tasks to probe their generalisation strategy. Behavioural results showed that people learned optimal solutions (policies) to the training tasks, and reused them on test tasks in a reward-selective manner. Neural results showed that optimal solutions from the training tasks received prioritised processing during test tasks in occipitotemporal cortex and dorsolateral prefrontal cortex. These findings suggest that humans evaluate and generalise successful past solutions when solving new tasks.

Keywords: Decision-making; Generalisation; fMRI; Decoding; Reinforcement Learning

Introduction

The ability to flexibly generalise from past experience to new situations is central to human intelligence, but how humans decide what aspects of their experiences to generalise is still a mystery. Recent behavioural evidence suggests that humans are capable of reusing experiences (Tomov et al., 2021), with a sophisticated algorithm called successor features and generalised policy improvement (SF&GPI). The main idea is that – when a new situation is presented – an agent evaluates decision policies that have been successful during earlier learning experiences. To perform the evaluation and determine a course of action, it uses predictions about the task features that succeed each policy (Barreto et al. 2017, 2018, 2020).

If people generalise using SF&GPI, it should be possible to detect its components in their brain activity. We developed two neural predictions based on this premise. First, we predicted that successful past policies would be represented in brain activity when people are exposed to new tasks. Second, we predicted that these policies would be prioritised, showing stronger encoding than unsuccessful past policies.

Paradigm

To test these predictions, participants completed a multi-task learning experiment during fMRI (Fig. 1). On each trial, participants saw a cue that determined their current task, and then made a choice between four options. Our design had two main trial types. Training trials involved one set of cues and presented feedback after each choice (Fig. 1A). Test tasks – which were used to assess participants' generalisation strategy – involved different cues and did not show feedback (Fig 1B).

Training and test cues were constructed in a specific way to test the SF&GPI algorithm. Across training trials, two options lead to high reward and two options lead to marginal reward or losses. During the test trials, this pattern changed. The two options associated losses or marginal reward during training were now the most rewarding. Based on this design, an SF&GPI-abiding agent was expected to continue choosing the more rewarding option among the optimal training policies for each test task, but would not enact policies that had been unrewarding during training. Participants completed six blocks, with 48 training trials and 20 test trials per block.

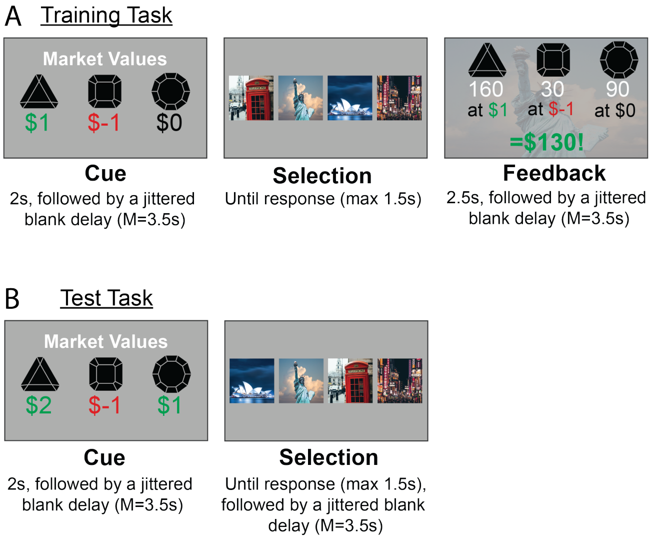
A
Training Task
Market Values
$1
$-1
$0
160
30
90
at $1
at $-1
at $0
=$130!
Cue
2s, followed by a jittered blank delay (M=3.5s)
Selection
Until response (max 1.5s)
Feedback
2.5s, followed by a jittered blank delay (M=3.5s)
B
Test Task
Market Values
$2
$-1
$1
Cue
2s, followed by a jittered blank delay (M=3.5s)
Selection
Until response (max 1.5s), followed by a jittered blank delay (M=3.5s)

Figure 1: Gem collector paradigm. On each trial, participants needed to retrieve gems to sell for as much profit as possible. Different cities around the world had different gem numbers and selling prices varied from trial-to-trial. Selling prices were used to define participants' task on the current trial. Training trials involved one set of selling prices. Generalisation was assessed on test tasks that used different selling prices.

Behavioural Results

We first examined how often participants made the optimal choice on training trials. The proportion of optimal choices was significantly above chance (M=0.83, SD=0.06, chance=0.25, *t*(37)=56.16, corrected p<0.001) indicating that participants acquired the optimal training policies. On test tasks, we found participants continued using optimal policies from training ($M_{proportion}$=0.69, SD=0.16, chance=0.5, t(37)=7.44, p<0.001), rather than selecting policies that offered the highest objective rewards. On test trials where participants used either of the two optimal training policies, the more rewarding one was also selected in most cases (M=0.93, SD=0.05, chance=0.5, t(37)=50.74, corrected p<0.001). These results indicate that rather than calculating the very best choice, participants were choosing among the optimal training policies on test trials in a reward-sensitive manner, consistent with the predictions of an SF&GPI algorithm.

Neural Results

To test the neural predictions of SF&GPI, we trained logistic decoders on fMRI data from pre-defined regions of interest, to distinguish choice stimuli seen during feedback on the training trials. We then applied the trained decoders to each measurement timepoint in the test trials and extracted decoding probabilities for stimuli corresponding the optimal training policies.

Dorsolateral PFC (DLPFC) was included based on research implicating it in policy encoding (Botvinick & An, 2008; Fine & Hayden, 2022) and context-dependent action (Badre & Nee, 2017; Flesch et al., 2022; Jackson et al., 2021). Medial temporal lobe (MTL) and orbitofrontal cortex (OFC) were included based on research implicating them in encoding predictive information that can be used for policy selection (De Cothi & Barry, 2020; Geerts et al. 2020; Muhle-Karbe et al., 2023; Stachenfeld et al. 2017; Wimmer & Büchel, 2019). Occipitotemporal cortex (OTC) was included due to its central role in early research using multivariate decoding and its inclusion in contemporary decoding studies (Haxby et al., 2001; Muhle-Karbe et al., 2023; Wittkuhn et al., 2021).

First, we tested the prediction that the optimal training policies would be encoded during test tasks. Neural results showed significant above chance decoding of the more rewarding training policy during test tasks, in OTC and DLPFC (OTC: M=28.07%, SD=2.17, t(37)=8.57, corrected p<0.001; DLPFC: M=25.85%, SD=1.51, t(37)=3.40, corrected p=0.011, Fig. 2A), but not in MTL or OFC (MTL: M=25.17%, SD=1.43, t(37)=0.72, corrected p=0.744; OFC: M=25.40%, SD=1.37, t(37)=1.79, corrected p=0.405). The 'more rewarding training policy' refers to the choice option (among the optimal training policies) that offered more reward on each test task. Second, we tested the prediction that the optimal training policies would receive prioritised processing during test tasks. Consistent with this prediction, average decoding evidence during the test tasks was significantly higher for the more rewarding training policy than the objective best policy (OTC: M_{diff}=4.09%, SD_{diff}=3.69, t(37)=6.75, corrected p<0.001; DLPFC: M_{diff}=1.61%, SD_{diff}=2.59, t(38)=3.77, corrected p=0.002, Fig. 2B). The magnitude of this neural priortisation in OTC was positively correlated with how often participants reused the optimal training policies during test tasks (OTC: Spearman's *Rho*=0.431, corrected p=0.014; DLPFC: Spearman's *Rho*=0.163, corrected p=0.327).

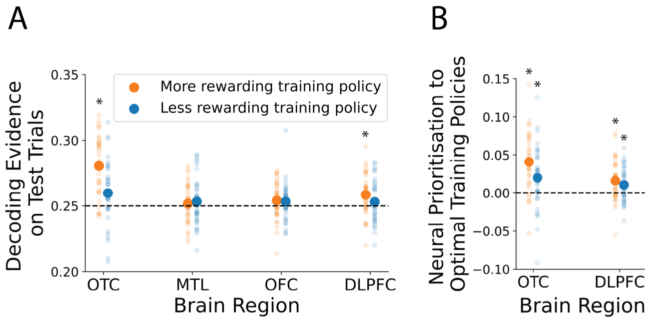

Figure 2: A) Decoding evidence for the optimal training policies during test tasks in each ROI. The dashed line indicates chance. B) Neural prioritisation towards the optimal training policies in OTC and DLPFC. Higher values indicate stronger decoding evidence on test tasks for the optimal training policies over the objective best policy. A-B) Large circles show sample means and small circles show individual participants. *p<0.05.

Conclusion

The present study provides behavioural and neural evidence that generalisation to new tasks was consistent with an SF&GPI-based algorithm. Successful past solutions were prioritised as candidates for decision making on tasks outside the training distribution. This prioritisation provides flexibility when faced with new decisions problems and has lower computational cost than considering all available options. These findings take a step towards illuminating the flexible yet efficient nature of human intelligence.

Acknowledgments

This research was supported by an Alexander von Humboldt Fellowship awarded to S.H.M, an Independent Max Planck Research Group grant and a Starting Grant from the European Union (ERC-2019-StG REPLAY-852669) awarded to N.W.S, and Multidisciplinary University Research Initiatives Grant (W911NF-21-1-0328) awarded to SG.

References

Badre, D., & Nee, D. E. (2018). Frontal cortex and the hierarchical control of behavior. *Trends in Cognitive Sciences*, 22(2), 170-188. https://doi.org/10.1016/j.tics.2017.11.005

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., & Silver, D. (2017). Successor features for transfer in reinforcement learning. *Advances in neural information processing systems,* 30.

Barreto, A., Borsa, D., Quan, J., Schaul, T., Silver, D., Hessel, M., ... & Munos, R. (2018, July). Transfer in deep reinforcement learning using successor features and generalised policy improvement. In *International Conference on Machine Learning* (pp. 501-510). PMLR.

Barreto, A., Hou, S., Borsa, D., Silver, D., & Precup, D. (2020). Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences*, 117(48), 30079-30087. https://doi.org/10.1073/pnas.1907370117

Botvinick, M., & An, J. (2008). Goal-directed decision making in prefrontal cortex: A computational framework. *Advances in Neural Information Processing Systems*, 21.

De Cothi, W., & Barry, C. (2020). Neurobiological successor features for spatial navigation. *Hippocampus*, 30(12), 1347-1355. https://doi.org/10.1002/hipo.23246

Fine, J. M., & Hayden, B. Y. (2022). The whole prefrontal cortex is premotor cortex. *Philosophical Transactions of the Royal Society B*, 377(1844), 20200524. https://doi.org/10.1098/rstb.2020.0524

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 110(7), 1258-1270. https://doi.org/10.1016/j.neuron.2022.01.005

Geerts, J. P., Chersi, F., Stachenfeld, K. L., & Burgess, N. (2020). A general model of hippocampal and dorsal striatal learning and decision making. *Proceedings of the National Academy of Sciences*, 117(49), 31427-31437. https://doi.org/10.1073/pnas.2007981117

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425-2430. https://doi.org/10.1126/science.1063736

Jackson, J. B., Feredoes, E., Rich, A. N., Lindner, M., & Woolgar, A. (2021). Concurrent neuroimaging and neurostimulation reveals a causal role for dlPFC in coding of task-relevant information. *Communications Biology*, 4(1), 588. https://doi.org/10.1038/s42003-021-02109-x

Muhle-Karbe, P. S., Sheahan, H., Pezzulo, G., Spiers, H. J., Chien, S., Schuck, N. W., & Summerfield, C. (2023). Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. *Neuron*, *111*(23), 3885-3899. https://doi.org/10.1016/j.neuron.2023.08.021

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, 20(11), 1643-1653. https://doi.org/10.1038/nn.4650

Tomov, M. S., Schulz, E., & Gershman, S. J. (2021). Multi-task reinforcement learning in humans. *Nature Human Behaviour*, 5(6), 764-773. https://doi.org/10.1038/s41562-020-01035-y

Wimmer, G. E., & Büchel, C. (2019). Learning of distant state predictions by the orbitofrontal cortex in humans. *Nature Communications*, 10(1), 2554. https://doi.org/10.1038/s41467-019-10597-z

Wittkuhn, L., & Schuck, N. W. (2021). Dynamics of fMRI patterns reflect sub-second activation sequences and reveal replay in human visual cortex. *Nature Communications*, 12(1), 1795. https://doi.org/10.1038/s41467-021-21970-2