Recurrent circuit mechanisms for reward learning in multidimensional environments

Michael Chong Wang (chong.wang.gr@dartmouth.edu)
Psychological and Brain Sciences, Dartmouth College, 3 Maynard Street
Hanover, NH 03755 USA

Alireza Soltani (alireza.soltani@dartmouth.edu)
Psychological and Brain Sciences, Dartmouth College, 3 Maynard Street
Hanover, NH 03755 USA

Abstract

Decision making and learning in naturalistic environments involve choice options with multiple features, whereas usually only a few features and/or their conjunctions are predictive of their associated reward outcomes. It has been shown that humans deploy attention to selectively learn about the predictive values of features and feature conjunctions and generalize those values to similar stimuli/objects. This behavior can be captured by reinforcement learning models with explicit value representations. But how are such representations learned and used for decision making in neural circuits with mixed selectivity, and how does attention modulate these processes? To address these questions, we trained multi-area recurrent neural networks endowed with reward-dependent Hebbian plasticity on a multidimensional reward learning task. After training the networks to perform the task across diverse reward schedules, we tested them on the reward schedule used in a recent human study. The networks exhibited similar attentional biases as those observed experimentally. Despite their distinct topographies, we found that different networks shared an interpretable latent circuit organization that resembled the architecture of attractor network models. Specifically, distributed but orthogonal subspaces were used to encode and communicate information about different features and conjunctions within and across network areas, enabling the simultaneous learning of feature and conjunction values through reward-dependent Hebbian learning. Finally, we discuss how this structure gives rise to value-based selective attention, providing insight into how the underlying mechanisms can be validated in future experiments.

Keywords: recurrent neural networks; Hebbian plasticity; reinforcement learning; selective attention

Introduction

Reward learning in the real world requires learning the values of objects with multiple features from sparse and noisy feedback. The number of objects to learn about increases exponentially as the number of features grows, a problem referred to as the curse of dimensionality (Sutton & Barto, 2018). Previous studies showed that when faced with such challenges, humans deploy attention to prioritize learning about the values of reward-predictive features and conjunctions of features, and then combine those values to approximate the value of each object. While this behavior can be explained by reinforcement learning models (Farashahi & Soltani, 2021; M. C. Wang & Soltani, 2023; Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017), it is unclear how it can be implemented by neural circuits with mixed selectivity using biologically plausible learning rules. To address this question, we trained multi-area recurrent neural networks endowed with reward-dependent Hebbian plasticity on a multidimensional reward learning task (Farashahi & Soltani, 2021). We analyzed the behavior of these networks to verify that they exhibited the same attentional biases as human participants. To further identify the mechanisms underlying this behavior, we utilized dimensionality reduction methods to reduce the high-dimensional networks into low-dimensional interpretable circuits (Barbosa et al., 2023; Haxby, Guntupalli, Nastase, & Feilong, 2020; Langdon & Engel, 2022; Dubreuil, Valente, Beiran, Mastrogiuseppe, & Ostojic, 2022).

Methods

The multidimensional reward learning task involved learning about the values (reward probabilities) associated with multi-featured stimuli through probabilistic binary reward feedback. For the current study, each stimulus consisted of three feature dimensions where each feature dimension had three possible values, leading to 27 stimuli (objects) in total. We trained eight instances of two-area excitatory-inhibitory recurrent neural networks (Kleinman, Chandrasekaran, & Kao, 2021) (Fig. 1A). In each trial, the network's first area received inputs about the two available choice options. The probability of choosing each option was read out from the second area. The second area then received feedback about the chosen option. Finally, a reward outcome was delivered, and recurrent weights within each network were updated according to reward-dependent Hebbian plasticity (Farashahi & Soltani, 2021) to accumulate value information across trials. The network's dynamics are described by the equations below. The state of each unit x_t is a leaky integration of external input z_t and recurrent input. Each recurrent connection has a fixed component W_f and a plastic component W_t^p, which evolved through a Hebbian learning rule modulated by the reward r_t, with learning rates A. ε_t and ξ_t are white noise.

$$x_t = (1-\alpha_x)x_{t-1} + \alpha_x((W^f + W_{t-1}^p)h_{t-1} + Uz_t + b + \sqrt{2/\alpha_x}\varepsilon_t)$$
$$h_t = \tanh([x_t]_+)$$
$$W_t^p = (1-\alpha_W)W_{t-1}^p + r_t A \odot (h_t h_t^\top + \sqrt{2/\alpha_W}\xi_t)$$

To ensure that each network learned a general strategy for solving the task, we trained them on a large set of random reward schedules (J. X. Wang et al., 2018). We then tested them on a reward schedule utilized in previous human behavioral studies, which was designed such that an agent could learn a good approximation to the stimulus values by learning and integrating the values of one feature dimension (the informative feature) and that of the conjunction of the two other non-informative features (the informative conjunction) (Farashahi & Soltani, 2021; M. C. Wang & Soltani, 2023).

Results & Discussions

Task-optimized plastic recurrent neural networks replicate attentional biases in human behavior

When tested on a reward schedule that was used in prior work (Farashahi & Soltani, 2021; M. C. Wang & Soltani, 2023), the model's performance matched that of human participants. Using similar methods as in a previous study (M. C. Wang

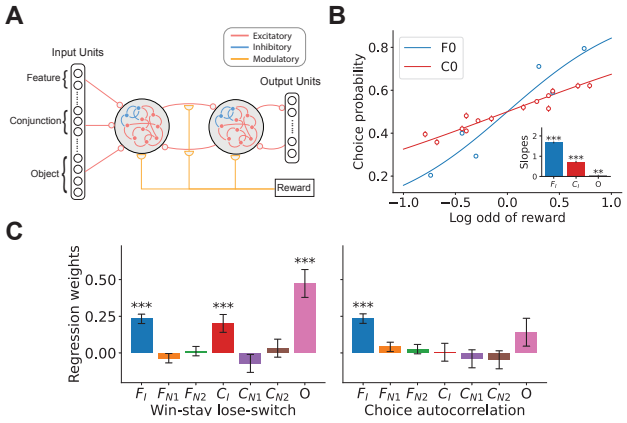

Figure 1: **(A)** Design of recurrent neural networks with reward-dependent Hebbian plasticity. **(B)** Choice behavior of the network as a function of the values of the informative feature, informative conjunction, and the stimulus object. **(C)** Logistic regression analysis of win-stay lose-switch and choice-autocorrelation. F_I, C_I denote the informative feature and conjunction. F_{N1}, F_{N2}, C_{N1}, C_{N2} denote the non-informative features and conjunctions. O denotes the stimulus (object). **: $p < 0.01$, ***: $p < 0.001$.

& Soltani, 2023), we found that the networks used weighted combinations of the values of the informative feature and conjunction to inform their decision making (Fig. 1B). Furthermore, these analyses revealed attentional biases in learning similar to those observed in human participants who preferentially associated reward with the informative feature and conjunction (Fig. 1C).

Latent circuits for multidimensional reward learning

Although the trained networks demonstrated behavioral signatures of attention that resembled those of human participants, different networks exhibited very different and heterogeneous connectivity patterns. To uncover a shared interpretable circuit motif across these networks, we applied demixed principal component analysis to the weights of the network that were fixed across trials (input weights, output weights, choice feedback weights) (Kobak et al., 2016). This provided us with the network-specific subspaces for encoding different stimulus dimensions. For all networks, these subspaces were nearly orthogonal to each other (Fig. 2A). Using these subspaces as a basis set, we performed a change of basis transformation on the within- and between-area connectivity matrices of each network in order to reduce the high-dimensional networks to low-dimensional latent circuits (Langdon & Engel, 2022). Applying this transformation to the within-area recurrent weights, the result showed a consistent pattern of nearly diagonal matrices with high positive diagonal values (Fig. 2B), reflecting within-subspace recurrent excitation, as well as low off-diagonal values, reflecting low interference between subspaces (Dubreuil et al., 2022). Applying this transformation to the between-area connection weights, we found separate communication subspaces that selectively relayed information about different stimulus dimensions across areas (Fig. 2C) (Barbosa et al., 2023). This orthogonal organization further allowed the population representations of previously chosen stimuli to be accurately encoded and retrieved through Hebbian learning (Fig. 2D).

In summary, we found that an interpretable circuit that resembles attractor network models of decision making and learning is embedded in the high-dimensional connectivity of the RNN. Interestingly, value-based selective attention has been demonstrated in attractor networks (Pannunzi et al., 2012). This explains the behavior of RNNs used in the current study and provides a plausible mechanistic explanation for value-based attentional biases in human reward learning.

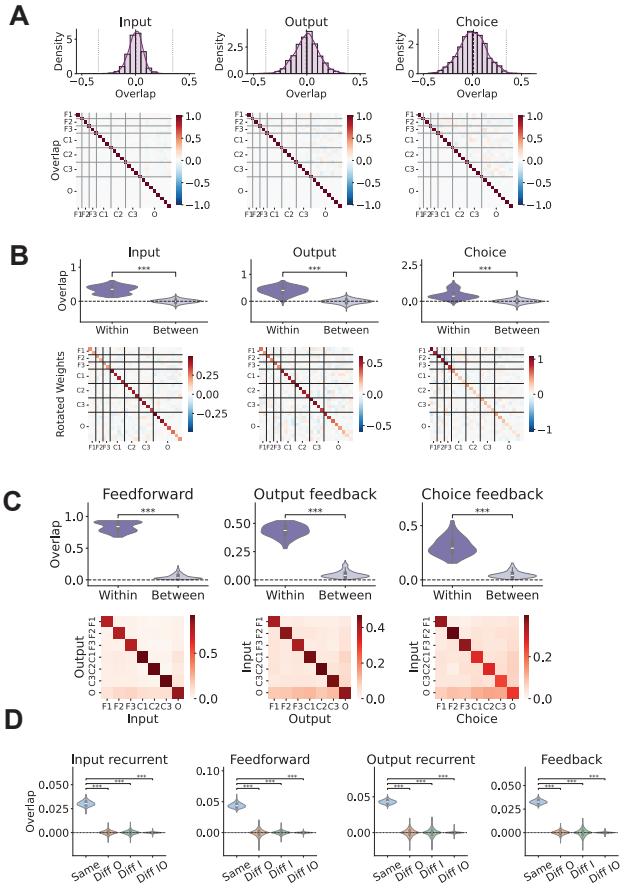

Figure 2: Task-optimized neural networks share a common latent circuit organization. **(A)** Overlap between stimulus feature, conjunction, and object encoding subspaces. **(B)** Within-area connectivity after a change of basis transformation. **(C)** Between-area connectivity after a change of basis transformation. **(D)** Overlap between the retrieved pattern of activity and the true pattern (Same) encoded through Hebbian learning, compared to false (Diff I, Diff O, Diff IO) patterns. All heat maps show averages across networks. All histograms show the distribution of individual values. ***: $p < 0.001$.

Acknowledgments

This work is supported by NSF CAREER Award BCS1943767 to A.S.

References

Barbosa, J., Proville, R., Rodgers, C. C., DeWeese, M. R., Ostojic, S., & Boubenec, Y. (2023). Early selection of task-relevant features through population gating. *Nature Communications*, *14*(1), 6837.

Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F., & Ostojic, S. (2022). The role of population structure in computations through neural dynamics. *Nature neuroscience*, *25*(6), 783–794.

Farashahi, S., & Soltani, A. (2021). Computational mechanisms of distributed value representations and mixed learning strategies. *Nature communications*, *12*(1), 7191.

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *elife*, *9*, e56601.

Kleinman, M., Chandrasekaran, C., & Kao, J. (2021). A mechanistic multi-area recurrent network model of decision-making. *Advances in neural information processing systems*, *34*, 23152–23165.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., ... Machens, C. K. (2016). Demixed principal component analysis of neural population data. *elife*, *5*, e10989.

Langdon, C., & Engel, T. A. (2022). Latent circuit inference from heterogeneous neural responses during cognitive tasks. *BioRxiv*, 2022–01.

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*(2), 451–463.

Pannunzi, M., Gigante, G., Mattia, M., Deco, G., Fusi, S., & Del Giudice, P. (2012). Learning selective top-down control enhances performance in a visual categorization task. *Journal of Neurophysiology*, *108*(11), 3124–3137.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., ... Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, *21*(6), 860–868.

Wang, M. C., & Soltani, A. (2023). Contributions of attention to learning in multi-dimensional reward environments. *bioRxiv*, 2023–04.