Recurrent neural circuits overcome partial inactivation by compensation and re-learning

Colin Bredenberg[1], Cristina Savin[1], Roozbeh Kiani[1]

[1]Center for Neural Science, NYU

email: cjb617@nyu.edu

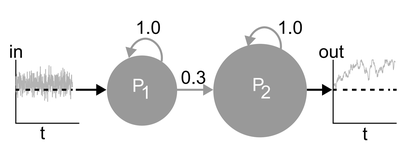
Motivation

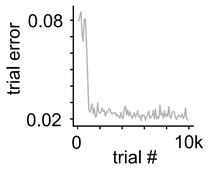
Lesions are an often-used experimental tool for studying the contribution of individual brain circuits to behavior, but their interpretation can be difficult.

What does it mean when a lesion has no effect on behavior?

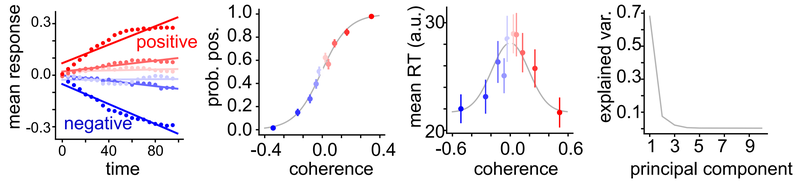
Idea: train different RNNs to solve a task, probe effects of 'lesions'/'inactivation'.

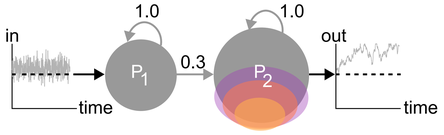
Here we focus on simple evidence integration tasks, a popular target for causal studies.



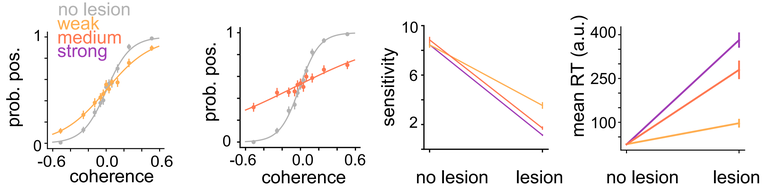


The network successfully learns a low-dimensional approximation of sensory integration.

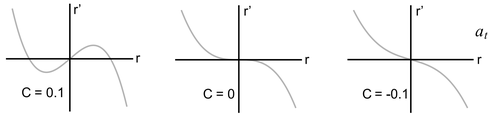




NN lesions decrease sensitivity and increase reaction time, with strong lesions leading to complete loss of function.



Pitchfork bifurcation explains network function and perturbation effects

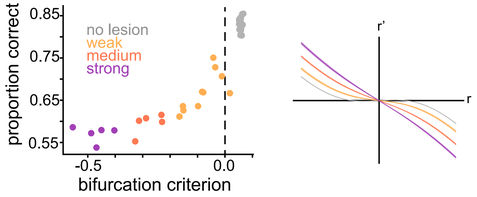


$$a_t - a_{t-1} \approx v_l^T (Wv_l - v_l) a_{t-1} - v_l^T (Wv_l^T)^{\circ 3} a_{t-1}^3$$

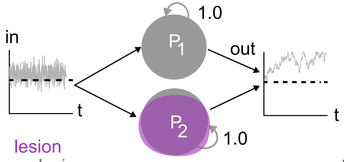
$$v_l^T (W - I) v_l < 0$$

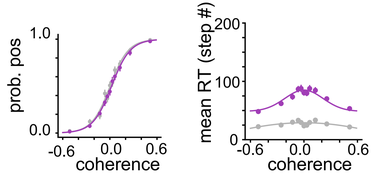
Changes in function can be modeled as a supercritical pitchfork bifurcation.

The bifurcation criterion (C) is monotonically related to task performance.

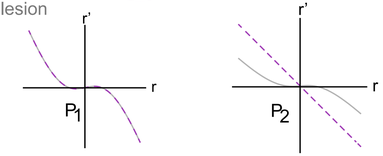


Lesions in a network with parallel architecture can have little to no effect.

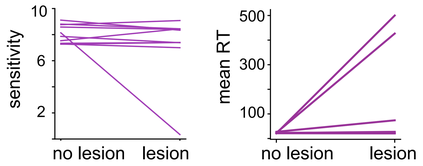




lesion
no lesion

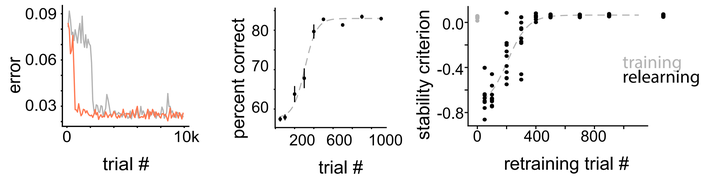


Diverse solutions: 1 or 2 bistable attractors.



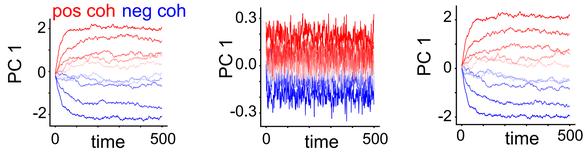
Lesions that target a single functioning attractor cause total loss of function; Lesions that target one of two attractors do not disrupt function.

Relearning Restores Function

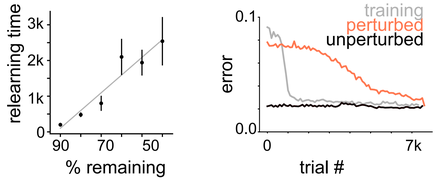


Perturbed networks relearn faster than the training time.

The bifurcation criterion is restored after relearning.



Retrained networks arrive at similar integration behavior.



Relearning time is linearly correlated with lesion magnitude.

Transient inactivation significantly slows relearning.

Conclusions

Results of lesion studies should be interpreted with caution: constantly learning networks compensate for lesions and may not show a clear loss of function after a short period of time. Networks with redundant or parallel architectures may be resistant to lesions as well.

Networks solve sensory integration tasks by learning a low-dimensional bistable attractor. The integrity of the attractor predicts the effect of lesions.

Time course of inactivation effects should be monitored.

Inactivation techniques that allow intermixing perturbed and unperturbed trials slow down relearning, making results interpretable.

References

Shadlen MN, Hanks TD, Churchland AK, Kiani R, Yang T. (2006). The Speed and accuracy of a simple perceptual decision: a mathematical primer. In *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Doya K, Ishii S, Pouget A, Rao RPN (eds.), MIT Press: Cambridge.

Wong, K. F., & Wang, X. J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience, 26*(4), 1314-1328.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual review of neuroscience*, 30.