# Brain mechanism of foraging: reward-dependent synaptic plasticity or neural integration of values?

**Ulises Pereira-Obilinovic**[1], **Han Hou**[2], **Karel Svoboda**[2], **and Xiao-Jing Wang**[1,*]

[1]Center for Neural Science, New York University, New York, NY 10003, USA; [2]Allen Institute for Neural Dynamics, Seattle, WA 98109, USA; [*]xjwang@nyu.edu

This manuscript was compiled on September 26, 2022

**In foraging, behavior is guided by action values that are stored in memory and updated depending on the history of choices and rewards. What is the neural mechanism for action value maintenance and reward-dependent update? Here we explore two contrasting network models: neural integration and synaptic learning of action value. We show that both mechanisms reproduce existing experimental data, but they depend on distinct and falsifiable assumptions about the biological neural circuits. We demonstrate neural dynamical signatures and perturbations that could be carried out experimentally to differentiate the two contrasting scenarios.**

Foraging | Value-based decision-making | Reinforcement learning | Neural integrator

During foraging, action values are stored in short-term memory over times of several behavioral trials, much longer than intrinsic neural time scales. The stored action values bias actions to maximize reward. In the cortex, single-neuron activity of the monkey posterior parietal cortex (1, 2), the mice medial prefrontal (mPFC) (3) and retrosplenial cortex (4), is correlated with estimated action values. Such value coding has traditionally been explained by reward-dependent synaptic plasticity (5, 6). However, recent findings of graded neural activity in proportion to newly updated action values during the inter-trial intervals (3, 4) have raise an alternative account by virtue of neural integration (7). In this work, we investigated these two contrasting scenarios (Fig. 1A, B).

According to the neural integrator model (Fig. 1A, top row), action values are stored in directions of slow change in population activity or neural integrator. The neural integrator integrates reward events by updating action values at each trial by an input proportional to the reward prediction error (RPE) in the direction of action value encoding (see Fig. 1A top row). Importantly, in this mechanism, changes in the synaptic efficacy are not needed for action value update, similar to a theoretical proposal of mPFC as a meta-reinforcement learning system (8). On the other hand, according to the synaptic model (Fig. 1A, bottom row), dopamine release triggered by reward enables Hebbian plasticity on the recurrent connections of the valuation network, in turn modifying in a reward-dependent manner the synaptic efficacy on the network recurrent connections (5, 6). Since synaptic inputs drive spiking activity, action values modulate neuronal activity.

## Results

**Foraging dynamics.** We explore these two contrasting hypotheses for the neural mechanisms underlying action value maintenance and its reward-dependent update during foraging. We develop two experimentally testable circuit models. These models were guided by anatomical and physiological evidence of the neural circuits involved in foraging behavior. We aim for these models to be simple enough to be suitable for mechanistic understanding and mathematical analysis, but to include key details to the extent their predictions can be interpreted in actual brain circuits.

Here we focus in the dynamic foraging task (DFT) (2–4, 9, 10) (see SI).

In our models, the network architecture consists of a mesoscopic network of two inter-connected areas, modeled as recurrent networks (Fig. 1A, second column). Action values for the two contingencies are maintained and updated depending on reward, in a network we refer to as a valuation network. Action is selected depending on action values in what we refer to as the action selection network. This architecture is inspired in the separation between the limbic and motor streams of information during decision-making (11).

In our neural integrator model, action values for the two contingencies are maintained by the population activity in a two-dimensional manifold embedded in a high-dimensional neural space (the space spanned by the activity of all neurons in the network) (Fig. 1A, first row). After the action is selected and the reward is delivered, an RPE signal updates the corresponding action value represented in the action value manifold (Fig. 1B, first row). This manifold correspond to overlapping populations of neurons. Unlike classic neural integrators where noise gets integrated, in our model, random perturbations hardly produce significant changes in the encoding of action values. This is because, in high dimensions, random perturbations are, on average orthogonal to the action value encoding directions (see SI). However, this comes with a price to pay. Namely, our model predicts that RPE signals should be aligned to the action value encoding axis corresponding to the chosen contingency. Therefore, such alignment should be contingency-dependent, with RPE signals projecting to different action value axes depending on the selected action (Fig. 1A, top row). This model predicts different RPE-correlated activity patterns in the valuation network for different selected contingencies.

In the synaptic model, feedback from the action selection network carrying choice information will transiently elevate the activity of the corresponding selective populations in the valuation network after the action is selected (Fig. 1B, bottom row). In contrast, little to no feedback activity is produced in the corresponding unchosen population. After reward delivery, the transient elevated activity in the corresponding chosen populations will drive Hebbian learning gated by a global

U.P.-O., K.S., and X.-J.W. designed research; U.P.-O. performed the research; H.H. performed the behavioral experiments; H.H. fitted the behavioral models with the help of U.P.-O.; U.P.-O. wrote the paper with the contributions of all authors.

The authors declare no competing interest.

[2]To whom correspondence should be addressed. E-mail: xjwang@nyu.edu

RPE signal using a three-factor learning rule (5, 6) (see the bottom row in Fig. 1 A). Thus, only the recurrent synapses of the corresponding chosen populations will modify their synaptic efficacy. Importantly, these selective feedback signals after action selection are not observed in the neural integrator (Fig. 1B).

During foraging, both models display value-dependent persistent activity during the inter-trial interval (ITI) (Fig. 1B) (3, 4). After the action selection network makes a choice, the activity is modified in both models, and action value representations are persistent during the next ITI (3, 4). Both models display matching behavior (2–4, 10) (Fig. 1C). Strikingly, as it has been observed in the cortex (3, 4), in both models, the valuation network presents population activity correlated with baited reward (Fig. 1D).

A non-selective multiplicative go signal to the action selection network is needed for stable foraging in a broad range of reward probabilities for both models (12). This signal is consistent with a non-selective input to the thalamus in a recurrent connected two-network action selection model involving the cortex and thalamus (13) (see SI). Our model predicts that the ITI activity in the action selection network also presents, a rather small, action value correlated persistent firing.

We fit both network models using mouse behavior during a dynamic foraging task (see SI). We found that both models have comparable performance with standard behavioral models for foraging (3, 4, 14, 15).

**Response to perturbations.** A central difference between these models is their response to perturbations.

First, in the synaptic model, the timescale of response to random fluctuations of the selective populations in the valuation network is correlated with action values, whereas in the neural integrator it is not (see Fig. 2A). The reason is that an increase (decrease) in the strength of excitatory recurrent connections increases (decreases) network reverberations and slows down (speeds up) the network's dynamics (see SI). Since action values updates are caused by changes in the synaptic efficacy of the recurrent synapses of the corresponding selective populations (see Fig. 1A, bottom), these changes will modify the valuation network's response to perturbations once the action value is updated (see Fig. 2A, bottom). In contrast, for the neural integrator, action value updates do not necessarily require synaptic efficacy changes (see Fig. 2A, top).

Second, when a perturbation targets action value encoding populations in the valuation network during the ITI the models behave qualitatively differently. In the case of the synaptic model, because action values are stored in the synaptic efficacies, disruptions in the action value encoding due to selective perturbations always rebound to previous values (see Fig. 2B). By contrast, in the case of the neural integrator, perturbations of population activity during the ITI can modify the action value encoding (Fig. 2C). Importantly, however, the effect highly depends on the way of perturbation. While on-manifold perturbation dramatically modifies action value encoding, random perturbations has little effect, because in large networks, random perturbations independently distributed across neurons are orthogonal (on average) to the action value manifold. On the other hand, uniform (spatially correlated) perturbations where all neurons are all perturbed by the same constant amount, reminiscent of optogenetic manipulations, could also disrupt action value encoding as on-manifold perturbations if they are strong enough. In the neural integrator, perturbations can be further engineered by ranking neurons by the strength of their encoding to one of the two contingencies. We found that by perturbing the first 10% of the highest action value encoding neurons produce a 83% of change in action value encoding (see Fig. 2D).

Lastly, one difference between these two models' architecture is that the synaptic model requires selective feedback projections from the action selection to the valuation network (see Fig. 1A). When this feedback signal is disrupted, action value update is not possible (see Fig. 2B)

## Discussion

In light of recent experimental findings, we considered two contrasting scenarios for action value estimation and updating underlying foraging behavior. For the synapse model, we considered a valuation circuit where the recurrent connections undergo reward modulated plasticity, in contrast to most previous models where plasticity occurs only at feedforward input connections. Because of the trial-to-trial changes of recurrent connection strengths, neural activity during inter-trial intervals varies in proportion to action values, consistent with the new experimental observations (3, 4). For the neural integrator model, we assumed a line attractor encoding an action value. Crucially, when embedded in a high dimensional space of neural population activity, such a line attractor becomes robust against random noise. We showed that this model can account for graded changes of neural firing during inter-trial intervals, but requires that value updating is realized by a reward prediction error signal (mediated by dopamine cells) projecting to value-coding neurons in a special way along a specific action value axis (e.g., for action A, not B), in a given trial. The latter implies that dopamine projection targets a selective neural populations in a choice dependent manner, an assumption not supported by presently known evidence but testable in the future. We demonstrate several other ways to differentiate the two scenarios, especially by optogenetic perturbations.

The perturbations proposed here are experimentally feasible using current neurotechnologies. However, reinforcement learning involves multiple interacting brain areas, which should be modeled in order to effectively target brain areas for performing these perturbations. The development of experimental protocols that run online behavioral models could be used, as proposed here, to target neural assembles in order to disprove the neural integrator hypothesis (see Fig. 2D). Our work shows that multi-regional neural recordings and perturbation experiments guided by network modeling are needed to advance understanding of the circuit mechanisms underlying foraging.

## Materials and Methods

A neural integrator and a synaptic network model for foraging behavior were implemented. Numerical simulations and analytical calculations were carried out. Additionally, a reduced version of these models were fit in mouse behavior during foraging. More details of the methods in SI. The code for simulating the network models is available at the GitHub repository https://github.com/ulisespereira/foraging-integrator-vs-synaptic. The code for fitting the reduced network models on mice behavior is available upon request.

**ACKNOWLEDGMENTS.** We thank Nai-Wen Tien and Marton Rozsa for their help with the behavioral training in the pilot experiments. We thank Aldo Battista and Jorge Jaramillo for the discussions and constructive comments on the manuscript. We thank Vishwa Goudar, Nuo Li, and Adam Carter for fruitful discussions. This work was supported by The Swartz Foundation (to U.P.-O.), James Simons Foundation Grant 543057SPI (to X.-J.W.), Janelia Visiting Scientists Program (to X.-J.W. and U.P.-O.), Howard Hughes Medical Institute (to K.S.), National Institute of Health U-19 grant 122-01-002-20 (to H.H. and K.S.), and The Allen Institute for Neural Dynamics (to H.H. and K.S.).

1. ML Platt, PW Glimcher, Neural correlates of decision variables in parietal cortex. *Nature.* **400**, 233–238 (1999).
2. LP Sugrue, GS Corrado, WT Newsome, Matching behavior and the representation of value in the parietal cortex. *science* **304**, 1782–1787 (2004).
3. BA Bari, et al., Stable representations of decision variables for flexible behavior. *Neuron* **103**, 922–933 (2019).
4. R Hattori, B Danskin, Z Babic, N Mlynaryk, T Komiyama, Area-specificity and plasticity of history-dependent value coding during learning. *Cell* **177**, 1858–1872 (2019).
5. A Soltani, XJ Wang, A biophysically based neural model of matching law behavior: melioration by stochastic synapses. *J. Neurosci.* **26**, 3731–3744 (2006).
6. Y Loewenstein, HS Seung, Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. *Proc. Natl. Acad. Sci.* **103**, 15224–15229 (2006).
7. HS Seung, How the brain keeps the eyes still. *Proc. Natl. Acad. Sci.* **93**, 13339–13344 (1996).
8. JX Wang, et al., Prefrontal cortex as a meta-reinforcement learning system. *Nat. neuroscience* **21**, 860–868 (2018).
9. RJ Herrnstein, Relative and absolute strength of response as a function of frequency of reinforcement. *J. experimental analysis behavior* **4**, 267 (1961).
10. B Lau, PW Glimcher, Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. experimental analysis behavior* **84**, 555–579 (2005).
11. J Cox, IB Witten, Striatal circuits for reward learning and decision-making. *Nat. Rev. Neurosci.* **20**, 482–494 (2019).
12. HK Inagaki, et al., A midbrain-thalamus-cortex circuit reorganizes cortical dynamics to initiate movement. *Cell* **185**, 1065–1081 (2022).
13. ZV Guo, et al., Maintenance of persistent activity in a frontal thalamocortical loop. *Nature* **545**, 181–186 (2017).
14. RA Rescorla, A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Curr. research theory* pp. 64–99 (1972).
15. GS Corrado, LP Sugrue, HS Seung, WT Newsome, Linear-nonlinear-poisson models of primate choice dynamics. *J. experimental analysis behavior* **84**, 581–617 (2005).

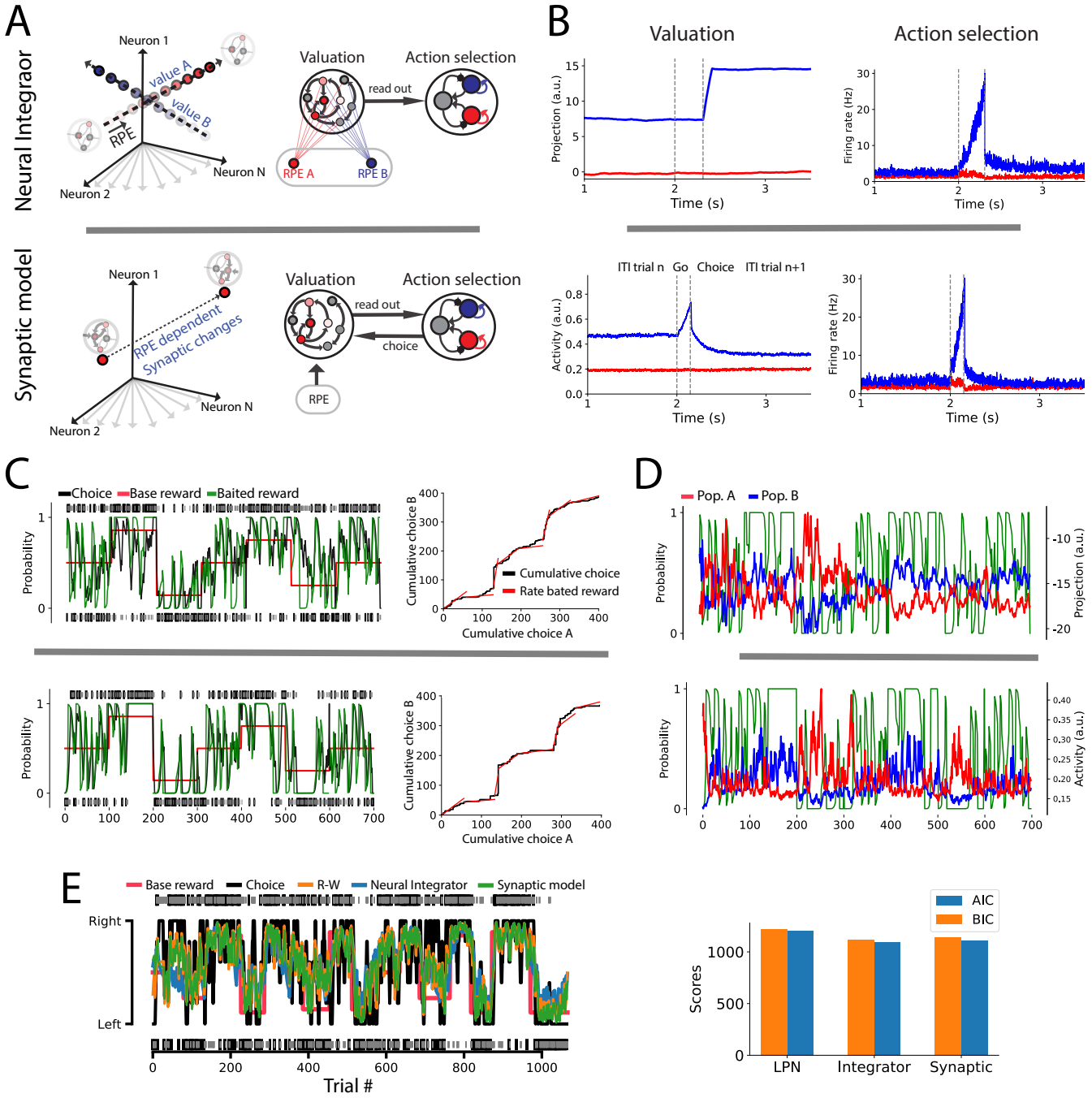


**Fig. 1.** Foraging behavior: Neural integrator vs. synaptic model. (A) Dynamical mechanism during foraging (first column) and network architecture (second column). (B) Single-trial network dynamics. (C) First column: foraging behavior during a DFT session. Second column: Cumulative choice and rate of baited rewards for contingency A vs. B. (D) Baited reward (green) and population activity averaged during the ITI for an example DFT session. (A-D) Top row: Neural integrator; Bottom row: Synaptic model. (E) Left: Foraging behavior of a mouse during an example session fitted using a Rescorla–Wagner model (R-W) (14), Neural integrator, and a synaptic model. Right: Model fitting AIC and BIC scores. Here the contingencies are licking left (A) and right (B).

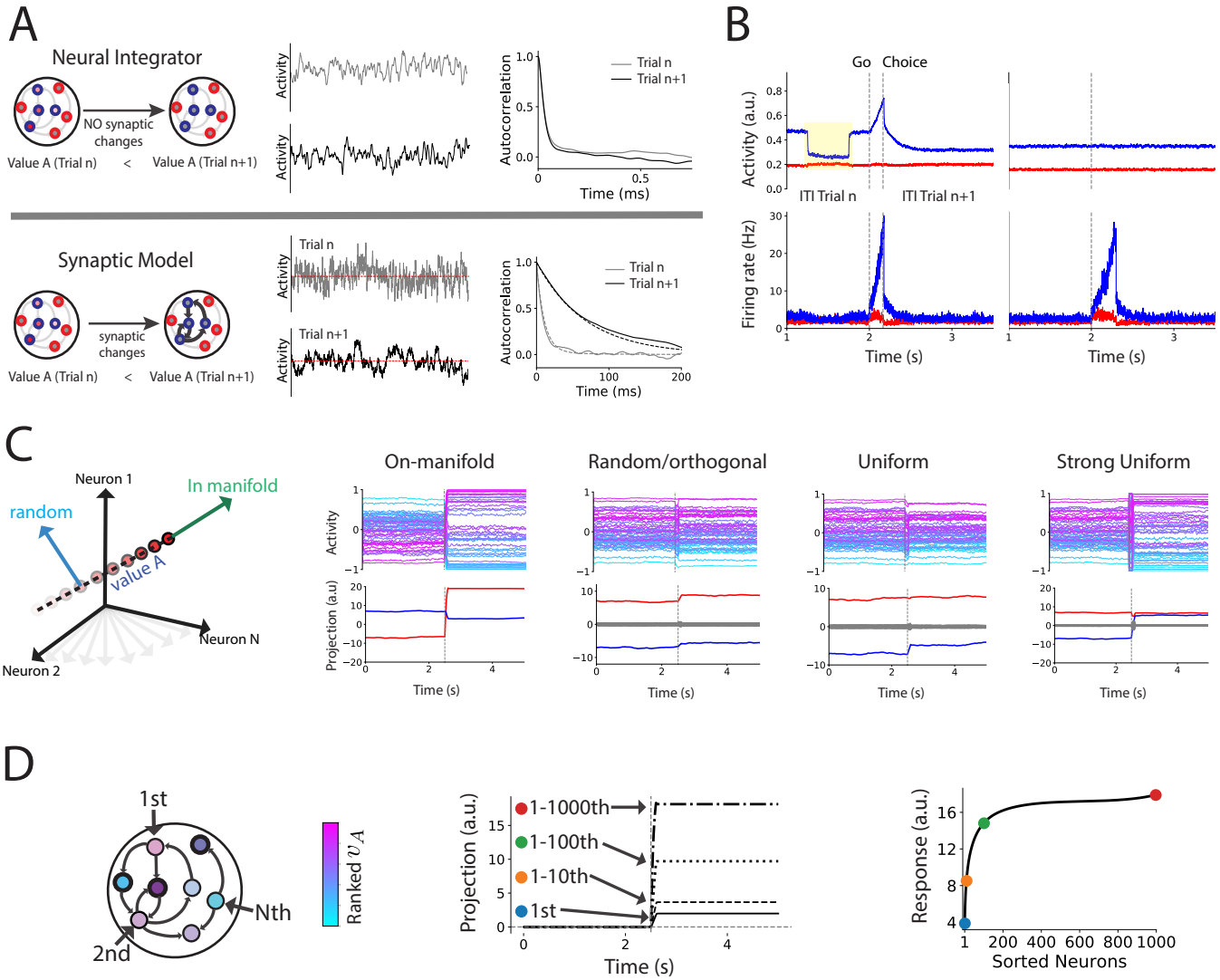


**Fig. 2.** Response to perturbations: Neural integrator vs. synaptic model. (A) First column: schematics for the synaptic dynamics across trials. Second column: neural activity for two consecutive trials with different action value encodings. Third column: autocorrelation function calculated from the ITI activity. Dashed traces: analytical formula (see details in SI). Top row: neural integrator. Bottom row: synaptic model. (B) Left column: perturbation during the ITI (yellow shaded region) in the valuation network. Right column: synaptic model with no feedback. (C) First column: schematic of on-manifold and random/orthogonal perturbations. Examples of on-manifold, random/orthogonal, and uniform perturbations (at the vertical dashed lines). (D) Engineered perturbation experiment for the neural integrator. Neurons are ranked according to the entries of the action value encoding axis $\vec{v}_A$ (first column). Change in action value encoding when the 1st, 1-10th, 1-100th, and all ranked neurons are stimulated with the same magnitude (second column). Third column: number of stimulated neurons vs. the magnitude of the perturbation response (see SI).