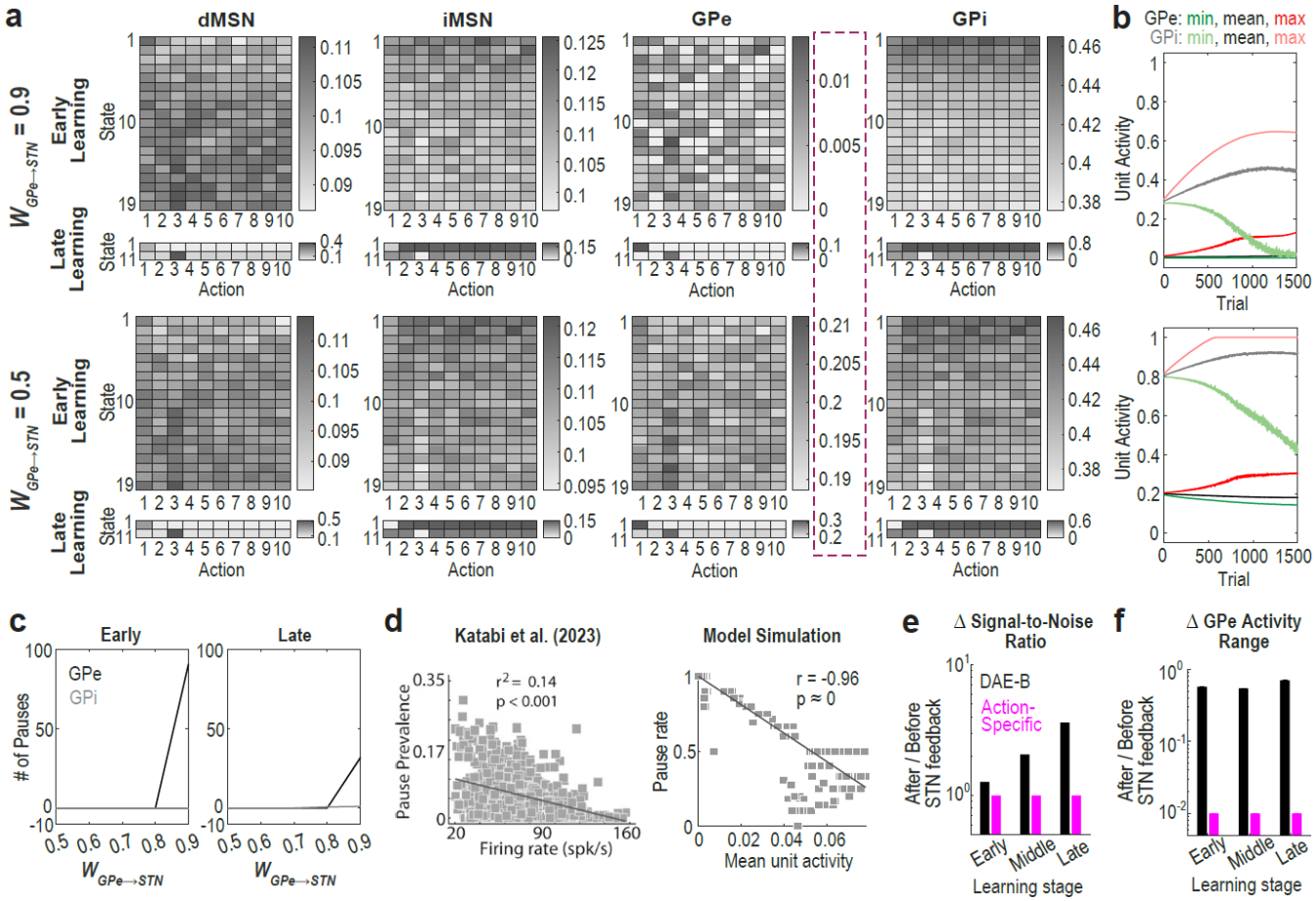
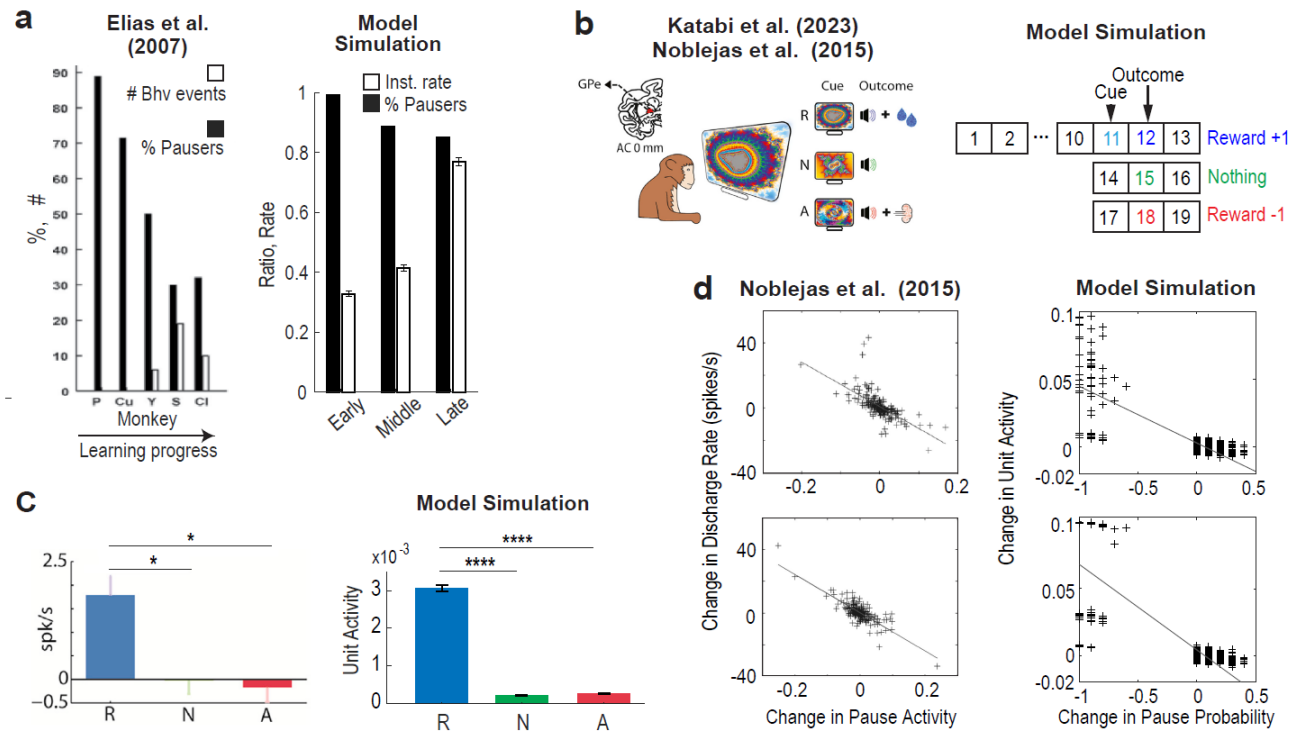


Supplementary Fig. 1: Reproduction of Fig. 2 using $W_{GPe\rightarrow GPi}$= 1.

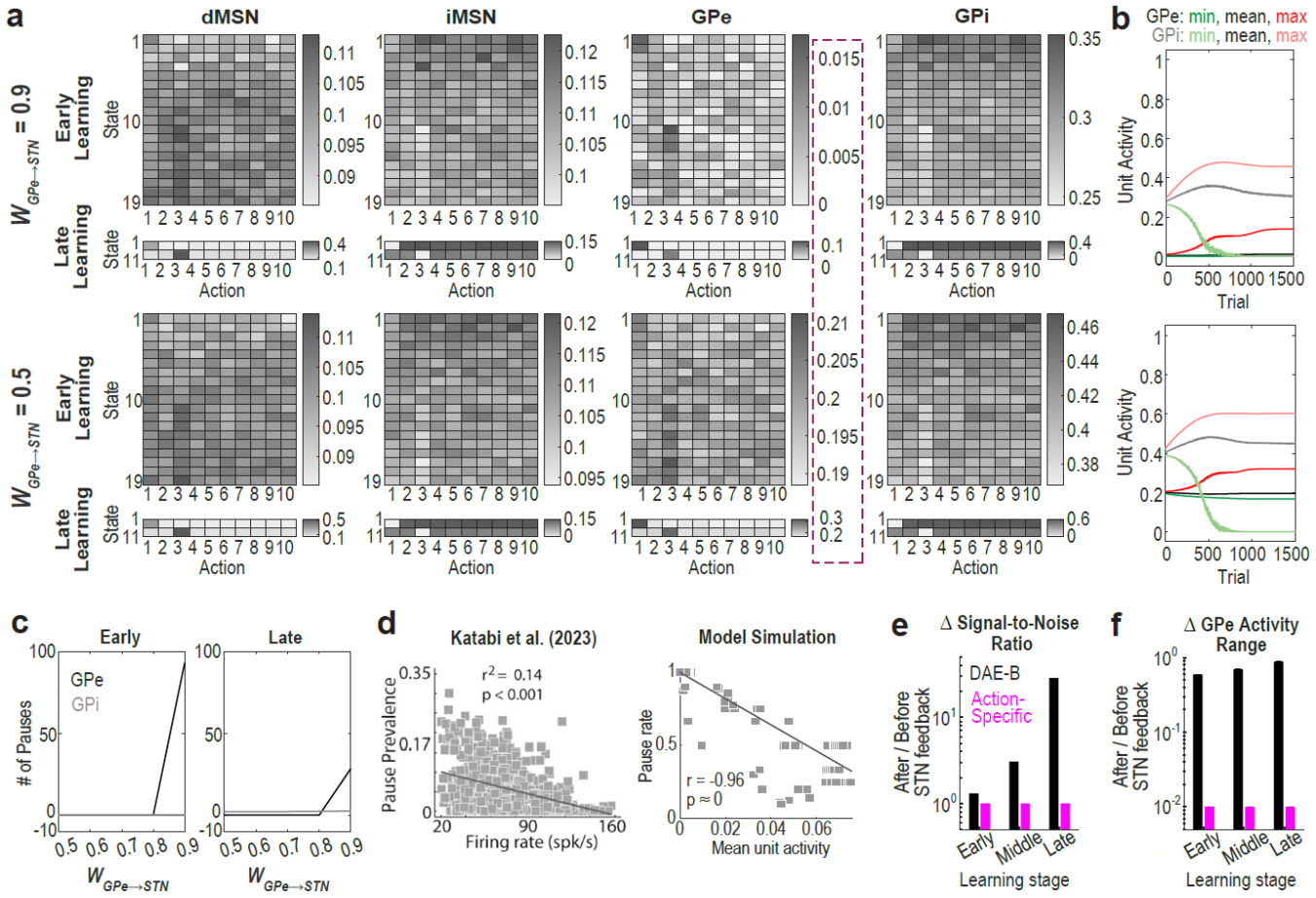


Supplementary Fig. 2: Reproduction of Fig. 3 using $W_{GPe \to GPi} = 1$.

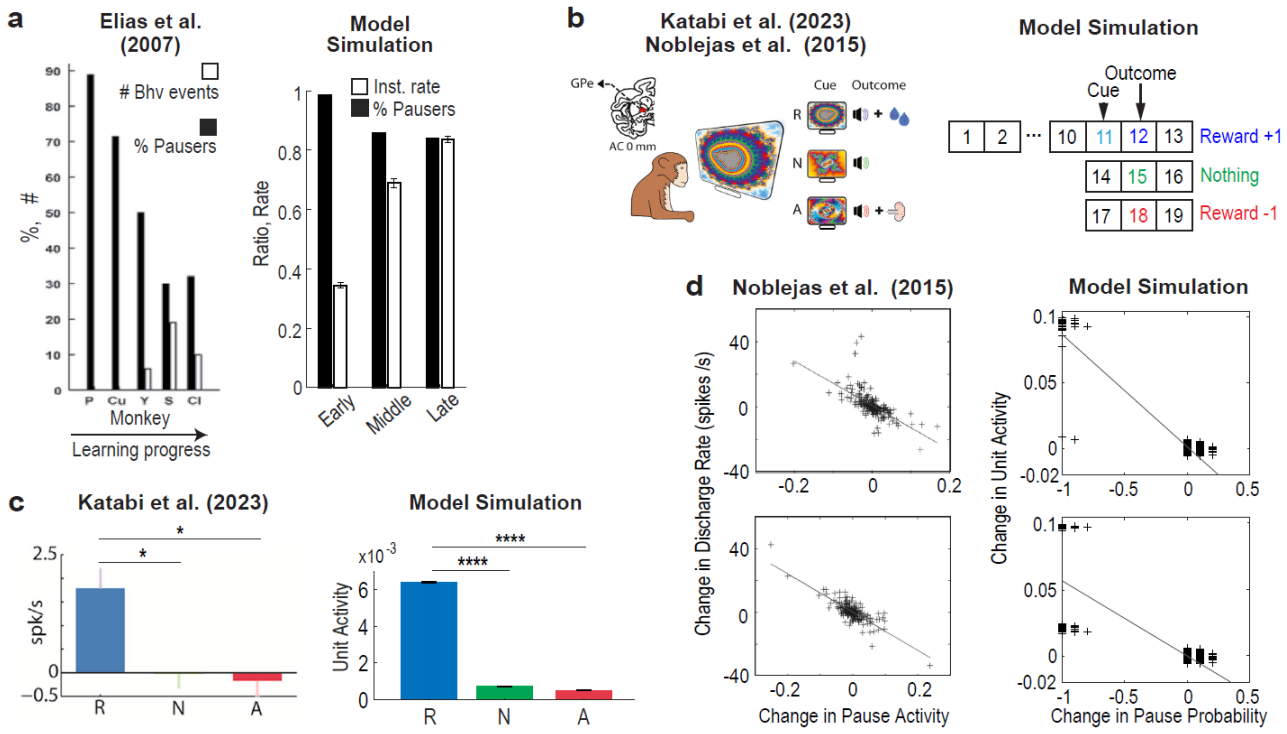
a: Page's trend tests revealed a significant increasing trend in instrumental response rates ($p = 9.72 \times 10^{-15}$, $z = 7.74$, $L = 1310$) and a significant decreasing trend in pauser rates ($p = 2.20 \times 10^{-16}$, $z = 11.31$, $L = 1361$; tested on reverse order).

c, right: ****$p = 2.56 \times 10^{-34}$ (Wilcoxon rank-sum test).

d, right: Pearson correlation analysis yielded: $r = -0.79$, $p = 6.06 \times 10^{-216}$ (top); $r = -0.83$, $p = 6.05 \times 10^{-254}$ (bottom). Pearson correlation analysis restricted to data points with y-values < 0.05 yielded: $r = -0.69$, $p = 6.85 \times 10^{-139}$ (top); $r = -0.79$, $p = 1.13 \times 10^{-200}$ (bottom).



Supplementary Fig. 3: Reproduction of Fig. 2 using $W_{GPe\rightarrow GPi}$= 3.

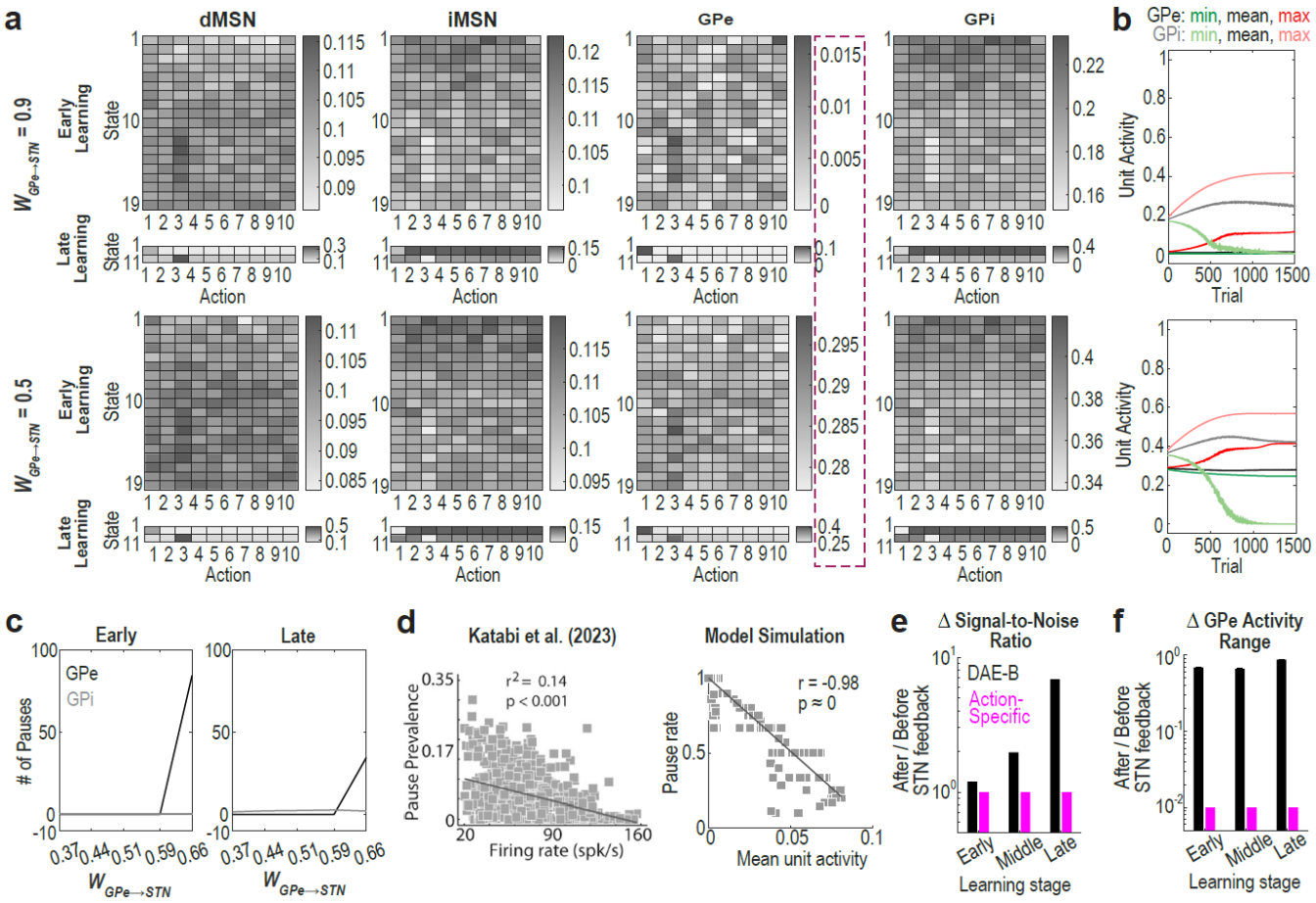


Supplementary Fig. 4: Reproduction of Fig. 3 using $W_{GPe \to GPi}$ = 3.

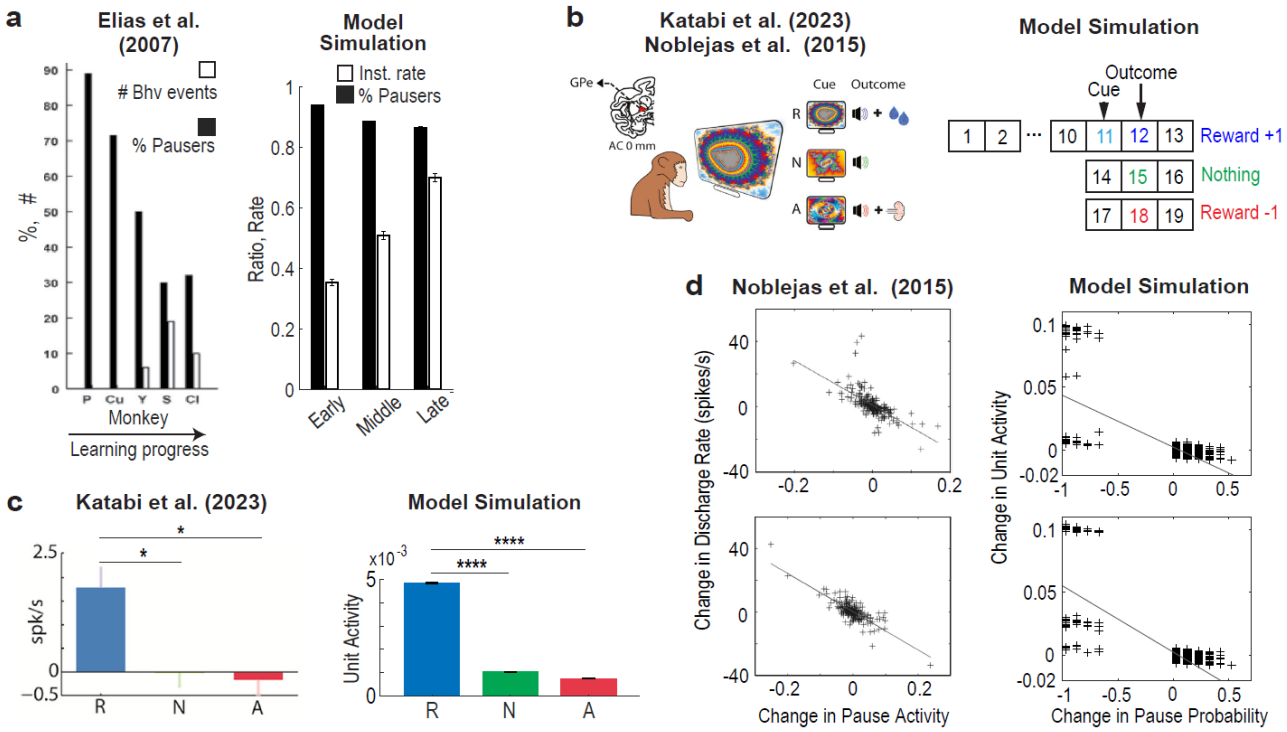
a: Page's trend tests revealed a significant increasing trend in instrumental response rates ($p = 2.20 \times 10^{-16}$, $z = 9.55$, $L = 1336$) and a significant decreasing trend in pauser rates ($p = 2.20 \times 10^{-16}$, $z = 10.04$, $L = 1343$; tested on reverse order).

c, right: ****$p = 2.56 \times 10^{-34}$ (Wilcoxon rank-sum test).

d, right: Pearson correlation analysis yielded: $r = -0.94$, $p \approx 0$ (top); $r = -0.81$, $p = 9.10 \times 10^{-230}$ (bottom). Pearson correlation analysis restricted to data points with y-values < 0.05 yielded: $r = -0.15$, $p = 4.64 \times 10^{-6}$ (top); $r = -0.84$, $p = 3.38 \times 10^{-257}$ (bottom).



Supplementary Fig. 5: Reproduction of Fig. 2 using $W_{STN\to GPe}$= 1.5 and $W_{GPe\to STN}$= 0.66.

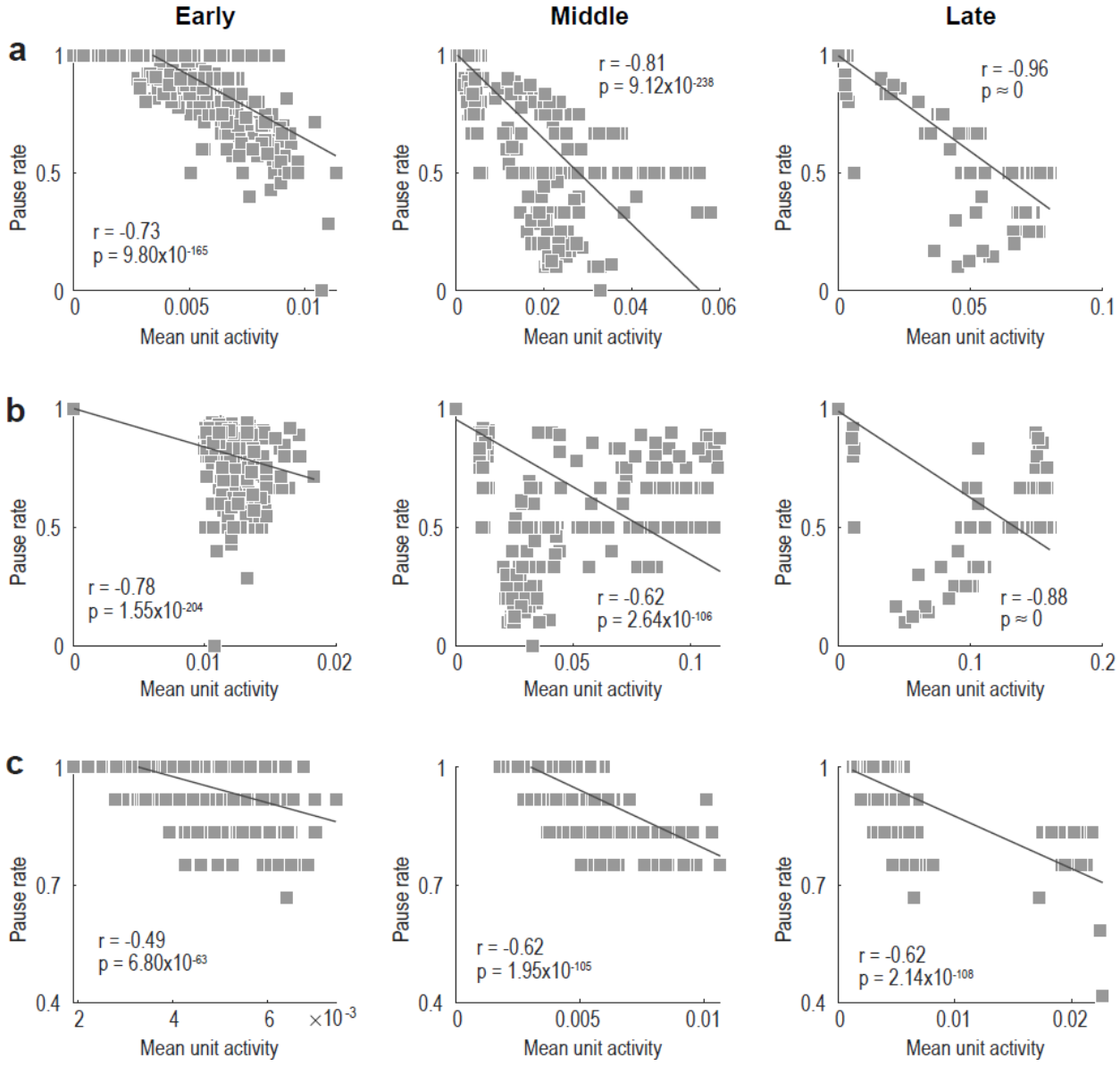


Supplementary Fig. 6: Reproduction of Fig. 3 using $W_{STN \to GPe}$= 1.5 and $W_{GPe \to STN}$= 0.66.

a: Page's trend tests revealed a significant increasing trend in instrumental response rates ($p = 1.48 \times 10^{-13}$, $z = 7.39$, $L = 1305$) and a significant decreasing trend in pauser rates ($p = 2.47 \times 10^{-10}$, $z = 6.33$, $L = 1290$; tested on reverse order).

c, right: ****$p = 2.56 \times 10^{-34}$ (Wilcoxon rank-sum test).

d, right: Pearson correlation analysis yielded: $r = -0.67$, $p = 2.70 \times 10^{-130}$ (top); $r = -0.77$, $p = 6.13 \times 10^{-193}$ (bottom). Pearson correlation analysis restricted to data points with y-values < 0.05 yielded: $r = -0.46$, $p = 6.95 \times 10^{-51}$ (top); $r = -0.77$, $p = 1.38 \times 10^{-188}$ (bottom).

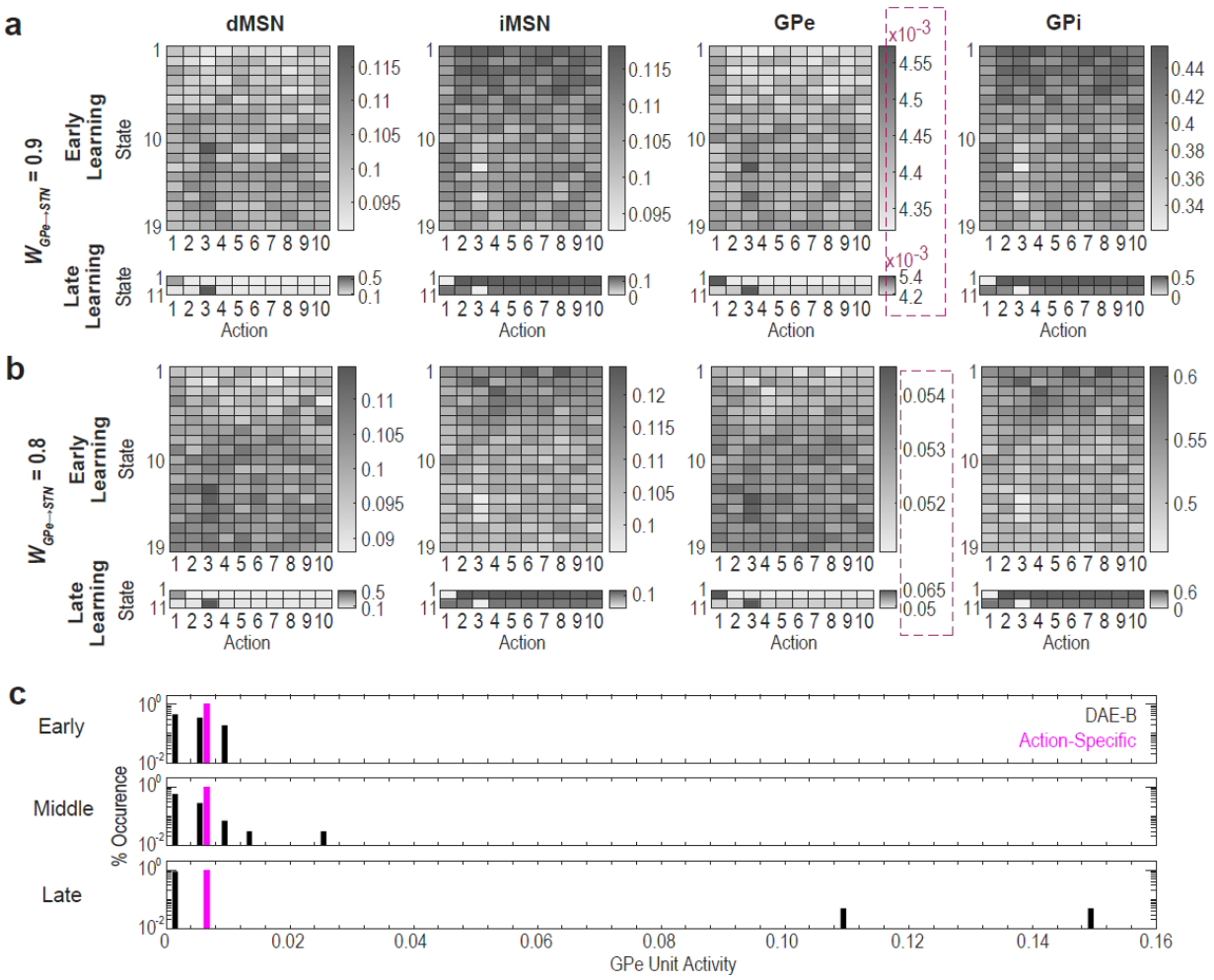


Supplementary Fig. 7: Reproduction of Fig. 2d (left) using corrected unit activity (b) and an alternative task design (c) based on Katabi et al. (2023)[24].

a: Replication of Fig. 2d (left) at early, middle and late stages, corresponding to trial 100, 600, and 1500, respectively.

c: Replication of **a** using corrected unit activity, in which mean unit activity was computed excluding pausing periods. As in Fig. 2d (right) and **a**, the model was trained on the task shown in Fig. 1c.

d: Replication of Fig. 2d (left) at early, middle and late stages, but with the model trained on the task illustrated in Fig. 3b.

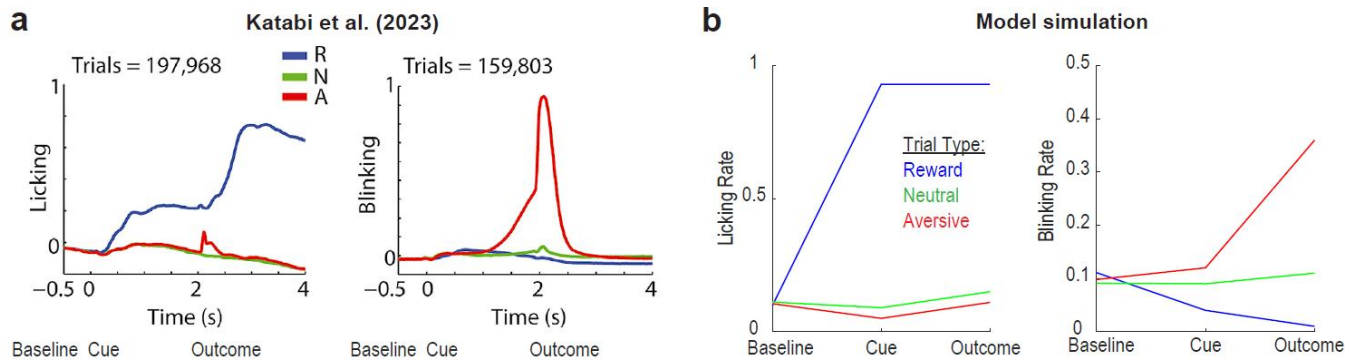


Supplementary Fig. 8: GPe pauses occur in an all-or-none fashion in a model with action-specific connectivity between the GPe and STN.

a: All GPe units display pausing activity (defined as activity < 0.01) when $W_{GPe\rightarrow STN}$ = 0.9.

b: No pauses are observed when $W_{GPe\rightarrow STN}$ = 0.8**.**

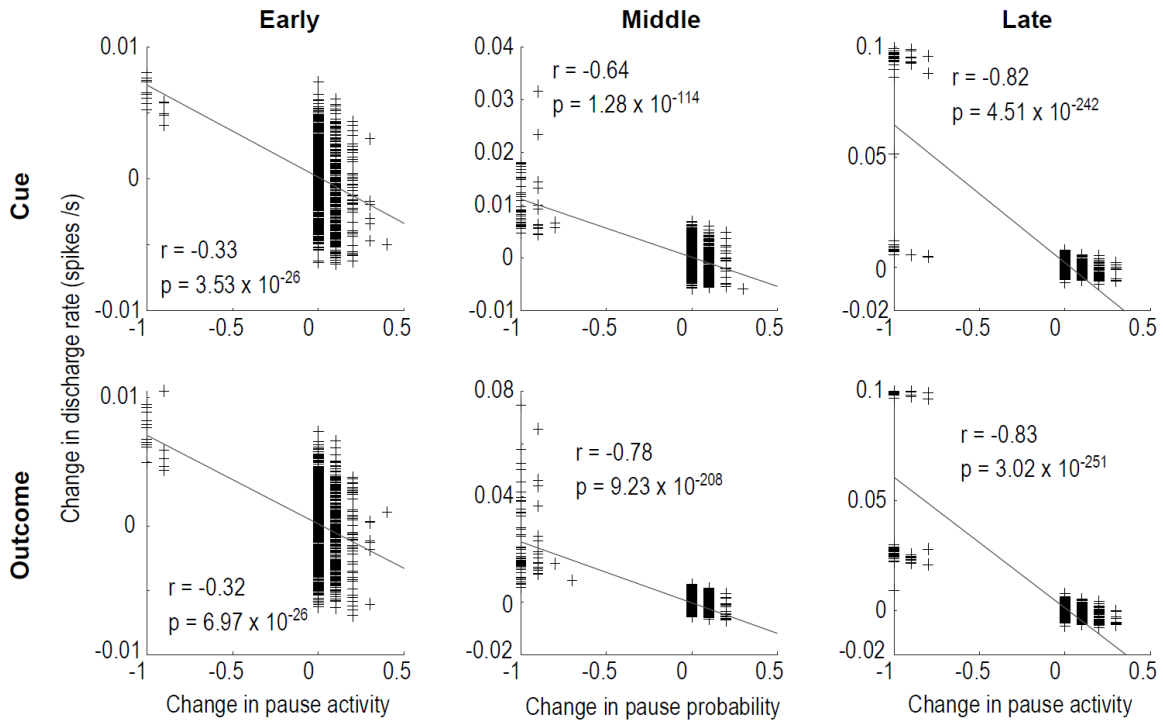
c: Histogram of GPe unit activity at early, middle and late learning stages.



Supplementary Fig. 9: Animal and model behavior during the task from Katabi et al. (2023)[24] and Noblejas et al. (2015)[18].

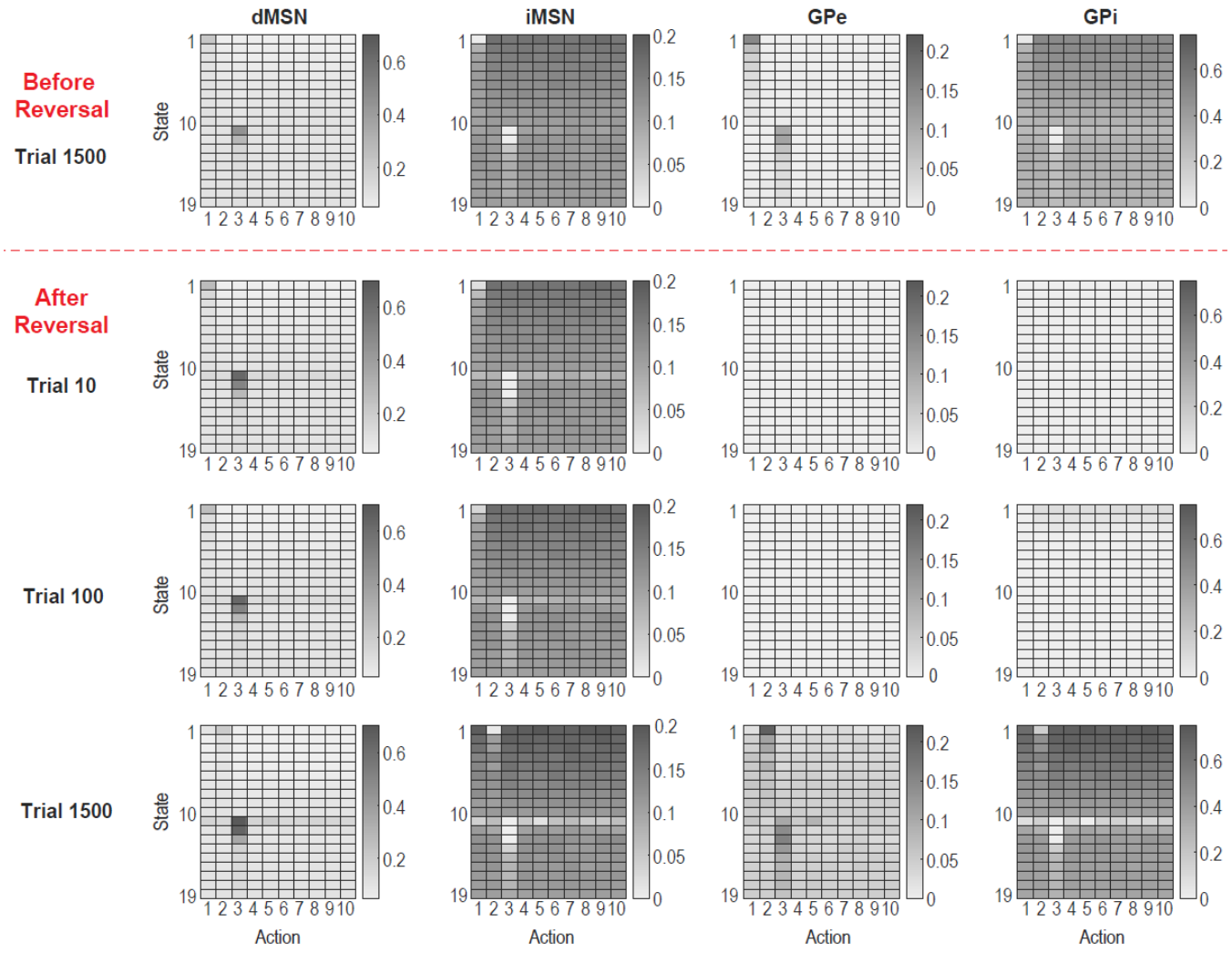
a: Licking (left) and blinking (right) during reward (blue), neutral (green) and aversive (red) trials in Katabi et al. (2023)[24] and Noblejas et al. (2015)[18]. The cue was presented for 2 s at time 0 s, and the outcome was delivered at 2 s. Figure adapted from[24].

b: Model behavior during the late stage, chosen to reflect the fact that the animals in Katabi et al. (2023)[24] and Noblejas et al. (2015)[18] were trained for months.



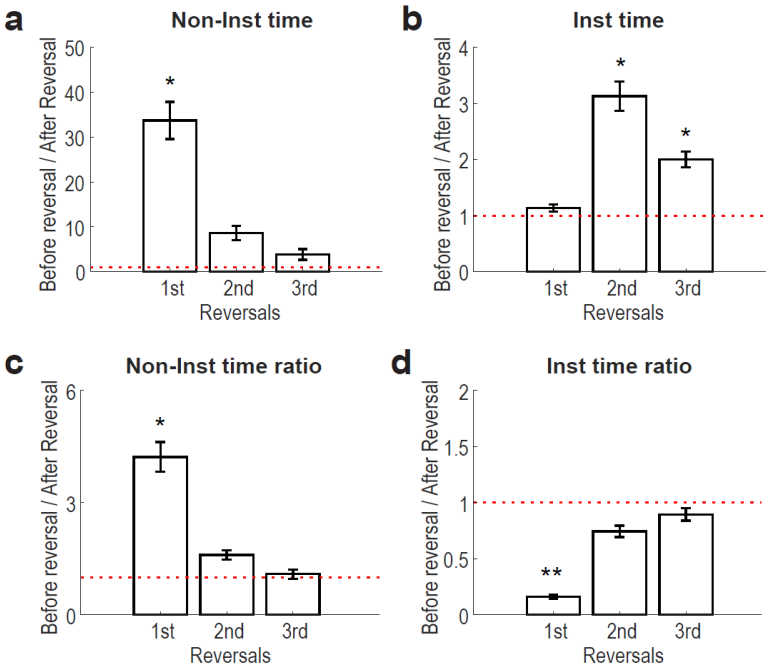
Supplementary Fig. 10: Changes in unit activity and pause likelihood during cue (top) and outcome (bottom) presentation.

Early, Middle, and Late learning stages correspond to trial 100, 600, and 1500, respectively. r and p in each plot indicate Pearson correlation coefficient and corresponding p-value. Each plot shows results from 100 simulations.



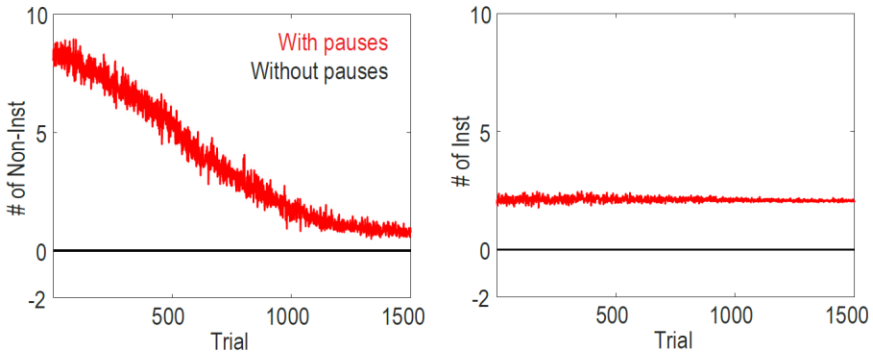
Supplementary Fig. 11: Model behavior before and after reversal when $Proto^{GPe \rightarrow STN}$ activity was artificially increased from the point of reversal onward.

Shortly after the reversal, almost no GPe units show excitation, abolishing relative activity differences among GPe units associated with distinct actions. $W_{GPe \rightarrow STN}$ = 0.9.



Supplementary Fig. 12: Animal behavior before and after each reversal.
In each plot, data from the day before reversal and the day of reversal were compared (N = 5).
a & b: Time spent on Non-Inst and instrumental behaviors.
c & d: Proportion of time spend on Non-Inst and instrumental behaviors relative to total session duration.
* $p < 0.05$ and ** $p < 0.0001$ (one-sample t-test).

After the first reversal, time spent on instrumental behaviors did not increased significantly, whereas time spent on Non-Inst increased substantially. A different pattern was observed following the second and third reversals: time spent on Non-Inst behaviors and the Non-Inst time ratio did not increased significantly, while time spent on instrumental behaviors increased significantly. These results suggest that the animals began to understand the task structure around the time of the second reversal and adopted a different exploration strategy—one more focused on instrumental behaviors.



Supplementary Fig. 13: Both Non-Inst (left) and Instrumental (right) behaviors consistently coincide with pauses throughout learning.

$W_{GPe \rightarrow STN}$ = 0.9. Results are averaged over 100 simulations.