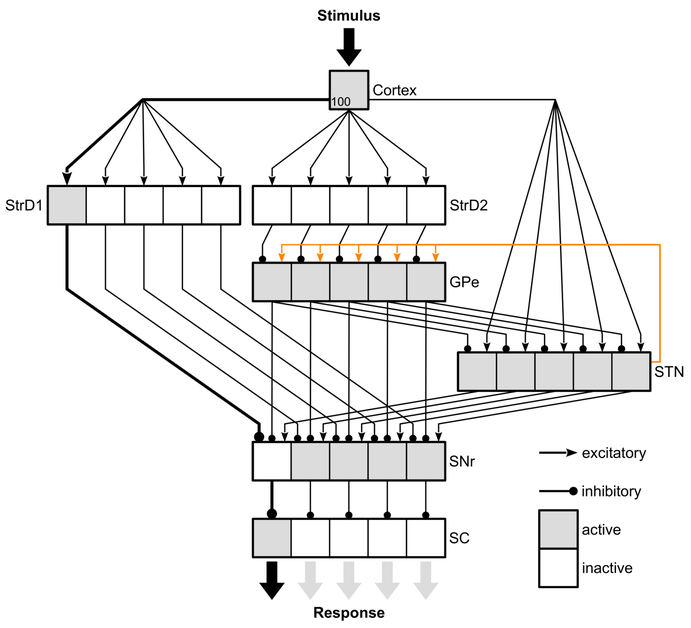
Synaptic plasticity in the STN – GPe loop biases exploration towards past rewarded responses

Oliver Maith[1], Javier Baladron[2], Wolfgang Einhäuser[3], Fred H. Hamker[1]

1: Department of Computer Science, Chemnitz University of Technology, Chemnitz, Germany
2: Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Santiago, Chile
3: Institute of Physics, Chemnitz University of Technology, Chemnitz, Germany

Despite much research and growing experimental data, an adequate understanding of decision-making at a mechanistic description level is still lacking. In our study, we investigate exploration behaviour by combining neuro-computational modeling and a behavioural experiment. We show that quite complex behaviour can be explained by a small sub-circuit - the STN-GPe loop - within the basal ganglia. Our experiment has been motivated by a prediction of our neuro-computational model of the basal ganglia. A particular novelty of this model is dopamine-modulated synaptic plasticity in the connectivity between the subthalamic nucleus (STN) and the external globus pallidus (GPe) of the basal ganglia. This adaptive sub-circuit enables the basal ganglia to stimulate alternative responses following negative reward prediction errors biased by past experiences. After a reversal, the indirect pathway not only inhibits the previously correct response but also retrieves the information stored in the STN-GPe loop and excites responses that were rewarded in the past. This extension of the basal ganglia function fits well with the recent observations, that address subcircuits in the basal ganglia and assigning various functions to it. The STN-GPe loop is often referred to malfunction, e.g. in Parkinson disease, but to our knowledge, however, little attention has been paid to its possible functional role, such as in exploration behaviour.
We tested the model prediction by means of a new version of a reversal learning task – a 5-choice reversal learning task with alternating position-reward contingencies – and analyzed if and how humans incorporate previous experience when exploring response options. With our new task, we extend the reversal learning research in which there has been no focus on exploration behavior so far. We found that humans preferentially explore previously rewarded response options which was in good quantitative agreement with our model's prediction. In particular, this preference evolves in a continuously progressive manner, suggesting an implicit learning process rather than explicit rule application. Our results point towards an interesting function of the STN-GPe loop.

The STN→GPe learning rule is based on homeostatic plasticity. It has an LTP component caused by below-average activity and an LTD component caused by above-average activity.



The model is composed of neural patches (squares), each with a population of 100 Izhikevich spiking neurons and connected to each other according to the major basal ganglia pathways.
Following each selection, SC-feedback toward the striatum increases the firing rate of the StrD1 and StrD2 sub-population encoding the saccade to the selected position. For simplicity, local projections in the striatum and the SC, as well as the feedback from SC to StrD1 and StrD2 and from cortex to SC are not shown.

The StrD2 population inhibits GPe and disinhibits the STN neurons of the same sub-channel

$$\Delta_{LTP} = \left(0.9 - \frac{A_{pre}}{\bar{A}_{pre}}\right)^{+} + \epsilon \cdot \left(0.9 - \frac{A_{post}}{\bar{A}_{post}}\right)^{+}$$

$$\Delta_{LTD} = \left(\frac{A_{pre}}{\bar{A}_{pre}} - 1.1\right)^{+} + \epsilon \cdot \left(\frac{A_{post}}{\bar{A}_{post}} - 1.1\right)^{+}$$

$$\tau_w \frac{dw}{dt} = (DA(t))^{+} \cdot \left(\left(1 - \frac{w}{w_{max}}\right) \cdot \Delta_{LTP} - \left(\frac{w}{w_{max}}\right) \cdot \Delta_{LTD}\right)$$

A_{pre} and A_{post} are spike traces

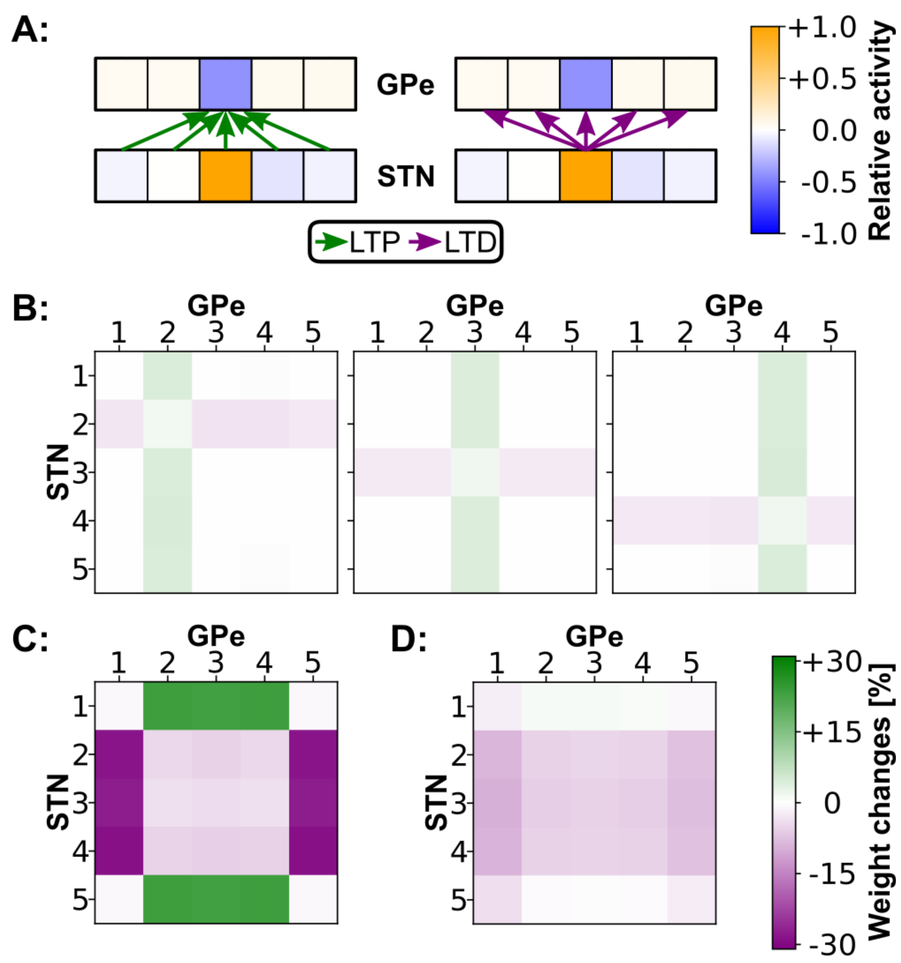
During positive-reward prediction errors, long-term potentiation (LTP) is caused by below-average neuronal activity (pre- or post-synaptic, $\epsilon = 10$), and long-term depression (LTD) is caused by above-average activity.

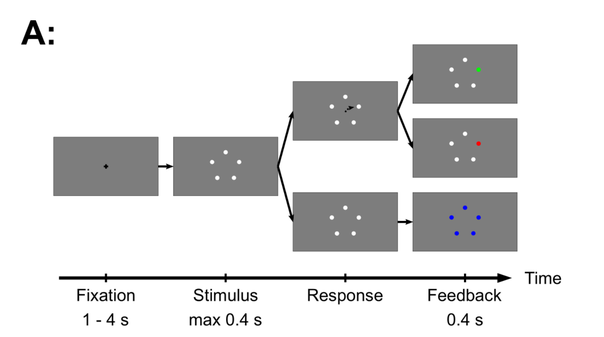
A: Relative activity in the STN-GPe sub-populations after a rewarded response, here selecting "position 3" (averaged over 15 consecutive rewarded trials of one block). Consecutive rewarded trials cause LTP in the synapses from all STN neurons to those GPe neurons of the selected sub-channel and LTD in the synapses of the neurons from the selected STN sub-channel to all GPe neurons.

B: Average weight changes between the neurons from the STN to the GPe (summed over 10 consecutive rewarded trials for each response). Left: response 2 was selected. Center: response 3 was selected: Right: response 4 was selected. The resulting weight changes lead to a decline of connections to unrewarded sub-channels and an increase to rewarded sub-channels (2, 3, 4).

When an STN sub-population becomes active during exploration, it excites the GPe cells of frequently rewarded positions more strongly, biasing selection towards those positions. C: Never-rewarded experiment.

D: Rarely-rewarded experiment (where responses 2, 3 and 4 were frequently rewarded, and responses 1 and 5 rarely rewarded)





A: Single trial of the task. After a central fixation, five identical, small white circles at unique positions arranged on an imaginary circle are presented. Participants are required to saccade to one of the five positions within 400 ms. If a saccade response occurs, a positive-reward feedback (selected circle → green) or a negative-reward feedback (selected circle → red) is given. Without a response, a negative-reward feedback is given after 400 ms (all circles → blue).

$$DA \leftarrow R, \text{ reward delivery}$$

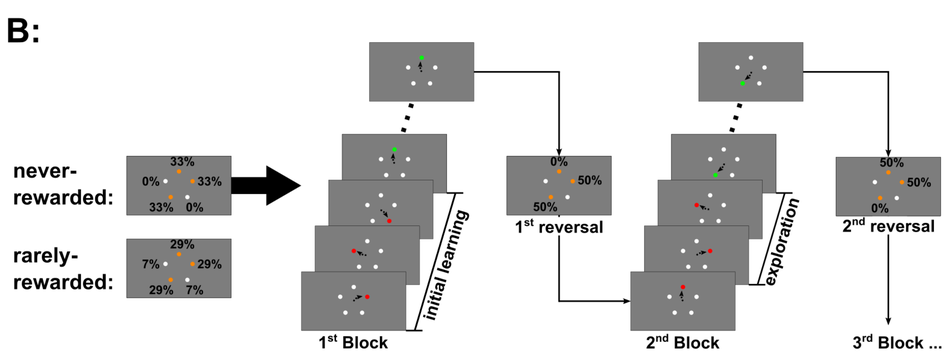
$$\tau_{DA} \frac{\mathrm{d}DA}{\mathrm{d}t} = -DA$$

$$R = \begin{cases} r - P_i, & \text{rewarded response i} \\ -r, & \text{unrewarded response} \\ -r/2, & \text{missing response} \end{cases}$$

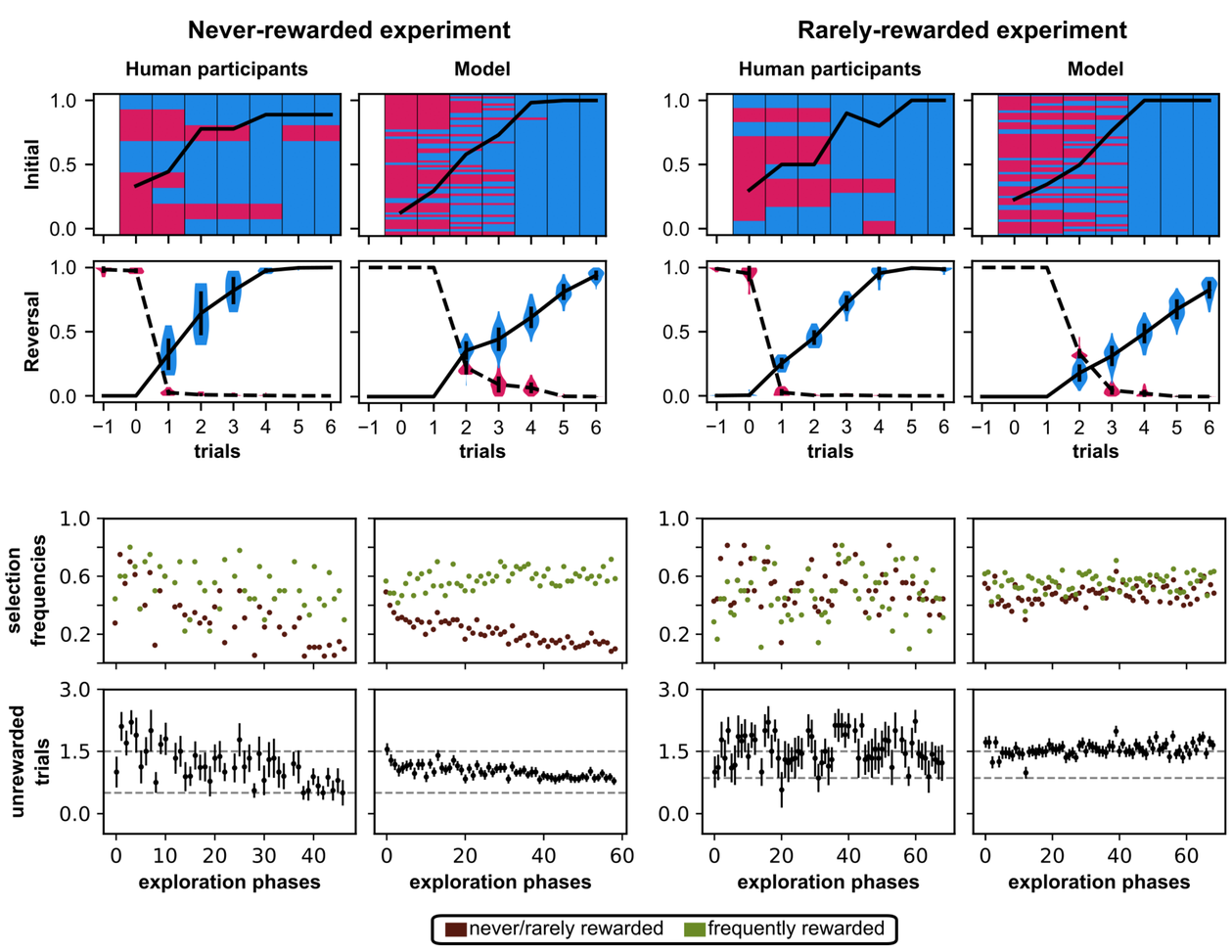
$$P_i = P_i + \frac{r}{\tau_P} \cdot \begin{cases} 1, & \text{rewarded response i} \\ -1, & \text{unrewarded response i} \end{cases}$$

$$-r \leq P_i \leq 0.8r$$

DA is set to the reward signal R and then decays exponentially leaving only a small window of time for dopamine-modulated plasticity.
R depends on the response of the model and is calculated from reward constant r=0.25 and a response-specific reward prediction signal P_i . P_i is updated at the occurrence of the corresponding response i.



B: The first block is initial learning; all following blocks start with a reversal thus contain an exploration. Unbeknownst to the participants, three positions are selected at the beginning of each experiment (highlighted here in orange) that are frequently rewarded during the different blocks of the task. The remaining two positions are never rewarded in the never-rewarded experiment and rarely rewarded in the rarely-rewarded experiment. Participants become aware of the reversal, when they receive negative-reward feedback for their previously rewarded choice. With the selection of an alternative position the exploration phase starts. When the newly rewarded position is discovered, the exploration phase ends.

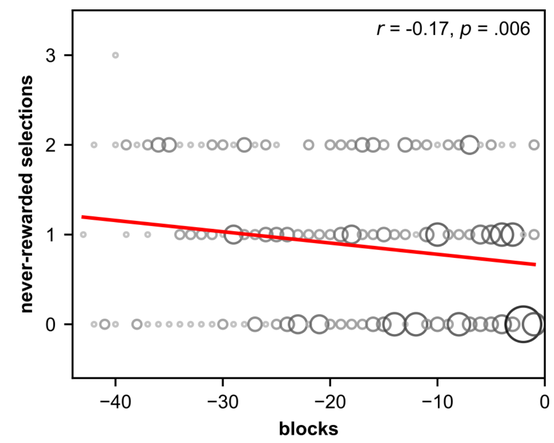


Human and model performance during the first trials of the initial learning phase (beginning of the first block) and reversal learning phases.

The selections of each individual participant/simulation are displayed in the background (red = unrewarded, blue = rewarded). Each row corresponds to a different participant/simulation.

Selection frequencies of unrewarded exploration trials

The dashed lines show the expected average numbers of unrewarded exploration trials for an ideal observer showing an exploration bias toward frequently rewarded positions (never-rewarded experiment: 0.5, rarely-rewarded experiment: 0.86) or not (1.5).



Exploration bias develops continuously: When we align the data to the block in which a participant has chosen for the last time a position that is never rewarded, our data indicate a continuous progress toward the bias. Red line: linear regression (Spearman correlation, r(258) = −0.17, p= 0.006, 95% CI -0.29 to −0.05).

Conclusion:
STN-GPe connections allow the STN-GPe loop to store information of rewarded responses and bias exploration towards these responses. Our model predicts a new function of the STN-GPe circuit that can learn to selectively stimulate alternative promising choices.

This research was funded by the German Research Foundation (DFG, 416228727) - SFB 1410 Hybrid Societies