Neurons and the synaptic basis of the fMRI signal associated with cognitive flexibility

Anja Stemme,a,* Gustavo Deco,b Astrid Busch,a and Werner X. Schneidera

aDepartment of Psychology, Ludwig-Maximilian University Munich, Leopoldstr. 13, D-80802 Munich, Germany
bICREA Research Professor, Computational Neuroscience, Universitat Pompeu Fabra, Passeig de Circumvallació 8, 08003 Barcelona, Spain

Keywords: Event-related fMRI; Cognitive set-shifting; Wisconsin Card Sorting Test; Attention; Working memory; PFC

Introduction

The Wisconsin Card Sorting Test (WCST) is well known to test cognitive flexibility in terms of set-shifting capabilities. Many fMRI studies with behaving monkeys as well as human subjects have shown transient neural activity in the Prefrontal Cortex (PFC), as indicated by an increase in the fMRI signal, following a rule change in the WCST or when using a WCST-like paradigm. We present a computational model, covering a limited number of PFC neurons and using precise biophysical descriptions, which is able to simulate WCS-like tests. Further, the detailed neuronal representation of the model allows us to calculate the resulting fMRI signal. Thus, we are able to analyze the adequacy of the model and its structure by comparing the calculated fMRI signal with the experimental data which in turn provides promising insights into the neural base of the increase in the fMRI signal.

© 2005 Elsevier Inc. All rights reserved.

Available online on ScienceDirect (www.sciencedirect.com).
promising framework for modeling the WCST in general, several items remain unsolved.

Braver et al. (1999) and O'Reilly et al. (2002) used a very similar approach to model subject behavior in a range of behavioral tests. Their model consists of simplified neural units receiving weighted excitatory, inhibitory and leaky input from other units. To cope with the rule switching and associated behavior in WCS-like tasks, an additional, dopamine-modulated, gating mechanism was introduced. The gating mechanism is responsible for the update of working memory representations and for shifting the attention to a new relevant feature dimension after a rule change and works as well by an error-driven modification of selected weights. Braver et al. (1999) suggested that a disturbed gating mechanism constitutes the shifting inability especially in schizophrenic patients. O’Reilly et al. (2002) replicated in their simulations the results of tests with frontally lesioned marmosets by damaging the neural units themselves. But, due to the lack of detailed biophysical representation, the same restrictions as stated above apply: abstract neural modeling only permits to replicate behavioral test data to a certain, limited extent and the integration of weight-modification processes does not permit to calculate a potential fMRI-signal and thereby validate the neurobiological plausibility of the model.

Therefore, in the present work, we use a detailed neuro-dynamical model describing main functional units involved in performing a test implementing a WCST-like paradigm. The functional units are established as pools of neurons and their identification is based on neurophysiological studies recording single cell data in behaving monkeys [see Miller (2000b) for a review, Asaad et al. (2000), White and Wise (1999), Wallis et al. (2001), Rainer and Miller (2002)]. The neurons are coupled via the three most common synaptic connection types found so far (NMDA—N-methyl-D-aspartate, AMPA—L-α-amino-3-hydroxy-5-methylisoxazol-4-propionic acid and GABA—gamma-aminobutyric acid) and are described using the so-called integrate-and-fire model which allows the production of realistic spiking dynamics. The synaptic connections are weighted thus generating the structure (i.e., the functional units) of the neural network model. In order to establish a set of weights eliciting the necessary model behavior in the simulations, rather than using a learning algorithm, we use the support of a mean field analysis (Brunel and Wang, 2001). During the test simulations, the previously determined weights will remain unchanged. Thus, the rule change does not require a weight-modification process.

Our model is used to simulate the experiments run by Nakahara et al. (2002) and Konishi et al. (1999) in a simplified manner. The calculation of the fMRI signal is thereby based on the synaptic activity which corresponds to the synaptic currents occurring during the simulations. Here, we refer to the work of Logothetis et al. (2001) who found that local field potentials (caused by synaptic currents) are more closely correlated to fMRI responses than the spiking outputs of a given area. Since we consider as well the summed spiking rate of the corresponding pools of the model as the base for the calculation of the fMRI signal, our work may provide further support for this hypothesis.

The entire model represents an extension of the framework of Brunel and Wang (2001) which has been applied already successfully to the description of selective working memory and attention [Almeida et al. (2004), Deco and Rolls (2005), Deco et al. (2004), Deco and Rolls (2003), Corchs and Deco (2002)].

Materials and methods

Description of the simulated WCST experiments

Experiment 1

Nakahara et al. (2002) used two trained monkeys and 10 human subjects to perform the tests as outlined in Fig. 1. A cue followed by three choice stimuli was presented on a screen. The subjects had to select one of the three choice stimuli which matched the cue stimulus in either the color or shape dimension by pressing a corresponding button. The relevant dimension was changed after six to eight successful responses without notice.

The measured fMRI-signal time-locked to the dimensional change is shown in Fig. 2. They found for the monkeys that the peak of the activation was located in the rostral bank at the ventral end of the inferior ramus of the arcuate sulcus. fMRI signals from these areas were transiently increased by 0.5 to 0.6% time-locked to the set-shifting event. For the human subjects, they observed that the most prominent shift-related activation of the PFC was located in the posterior part of the bilateral inferior frontal sulcus. In comparing monkeys and human subjects, they

---

Fig. 1. Experimental setup Nakahara et al. (2002) used in their WCST experiments: a cue stimulus was presented for 500 ms. Afterwards, three choice stimuli were presented on the screen. One of the choice stimuli matched to the cue stimulus in the color dimension, another in the shape dimension and the other in neither dimension. Subjects were required to match the cue stimulus to one of the choice stimuli based on the currently relevant dimension by pressing one of the three buttons. Visual feedback noticed subjects whether a response was correct or not. After six to eight successive correct answers, the relevant dimension was changed. ITI (intertrial interval), 2 s. (Reprinted from Nakahara et al. (2002), Fig. 1B.)
concluded that the main shift-related activity is located in the posterior part of the ventrolateral PFC across both species. They further hypothesized that these PFC regions represent homologous regions across these species with regard to cognitive set-shifting processes.

**Experiment 2**

Konishi et al. used seven healthy human subjects in a computerized version of the WCST. One target and four reference cards were presented on a screen. Subjects were required to choose one of the four reference cards which matched the target card in either the color, form or number dimension by pressing the corresponding button. After 10 consecutive correct answers, the relevant dimension for the answers was changed (see Fig. 3).

Furthermore, they distinguished two test conditions:

- Original condition: as described so far, subjects had to detect the new relevant condition themselves by a trial-and-error procedure.
- Instruction condition: subjects were explicitly informed about the new relevant dimension.

For both conditions, the fMRI signal was computed, time-locked to the dimensional set-shift. Konishi et al. found increased transient activity in the left and right inferior frontal sulci. Also, they found that the increase in the fMRI signal is significantly lower for the instruction condition compared to the original condition (see Fig. 4).

**Description of the neurodynamical model**

The model architecture we propose represents an extension of the framework first introduced by Brunel and Wang (2001). It consists of two different types of neurons, 80% excitatory pyramidal cells and 20% inhibitory interneurons, which are grouped into two appropriate types of pools: excitatory and inhibitory pools. The ratio of pyramidal cells versus interneurons is consistent with neurophysiological findings (Abeles, 1991). We chose to model $N_E = 1600$ excitatory neurons and $N_I = 400$ inhibitory neurons. The network of neurons is fully connected with different connection strengths. Each neuron is modeled as an ‘Integrate-and-Fire’ neuron taking into account three different synaptic connection types: two excitatory – AMPA and NMDA connections – and one inhibitory – GABA. The three different synaptic connection types are represented and computed using equivalent electrical circuits, consisting basically of a resistance parallel to a conductance with type-specific parameter values for conductivity and resistance. This way of neuron representation allows a very realistic simulation of the neurobiological processes as the following aspects are taken into account:

1. A neuron receives different types of input. If the summarized input is above a certain threshold, the neuron spikes and produces itself type-dependent post-synaptic excitatory or inhibitory input to other neurons.
2. After emitting a spike, the neuron is not able to spike again for a certain period of time (refractory period).
3. AMPA, NMDA and GABA inputs act in a different way. AMPA is considered as a ‘fast’ component, whereas NMDA is considered as a rather ‘slow’ input component with an additional magnesium-controlled voltage dependence. GABA, as the inhibitory component, acts opposite to AMPA and NMDA (different mathematical sign).

Each neuron receives a certain background input from neurons outside the network modeled. Thereby, it is taken into account that neurons always show a certain level of activity, that is, a

![Fig. 2. fMRI signal change for the monkeys as measured and estimated by Nakahara et al. (2002) with the fitted models of hemodynamic response, time-locked to the dimensional change. Arrows point to activated regions in the inferior prefrontal cortex. (Reprinted from Nakahara et al. (2002), Fig. 2B.)](image)

![Fig. 3. Experimental setup Konishi et al. (1999) used in their WCST experiments. A target and four reference cards were presented on a screen. Subjects had to choose one of the four reference cards which matched the target card in either the color, form or number dimension by pressing the corresponding button. After 10 consecutive correct answers, the relevant dimension for the answers was changed. They distinguished two test conditions: original and instruction condition. In the instruction variant subjects were explicitly informed of the new valid rule in case of a dimensional change. (Reprinted from Konishi et al. (1999), Fig. 1A.)](image)
spiking rate of approximately 3 Hz for pyramidal cells and 9 Hz for interneurons, which is called the ‘spontaneous rate’. The external background input is modeled as an AMPA-mediated Poisson train of spikes arriving from $N_{ext} = 800$ neurons with a rate of 3 Hz. Thus, the total background noise each modeled neuron receives comes out to $v_{ext} = 800 \times 3 \text{ Hz} = 2.4 \text{ kHz}$.

The neuronal connections are weighted in order to establish the structure (i.e., various pools) within the network model. We assume that the connections were established by Hebbian learning, that is, the coupling will be strong if a pair of neurons shows correlated activity, and weak if two neurons are activated in an uncorrelated way. Consequently, neurons within a specific pool are coupled with the relatively strong weight $w = 2.1$, whereas interpool connections are comparatively weak.

To identify specific pools (i.e., functional units) within the architecture, we refer to single cell recordings in monkeys. These recordings exemplified that single neurons show rule-specific (Wallis et al., 2001; White and Wise, 1999) as well as object-specific (Rainer and Miller, 2002) activity in a range of behavioral tests. These results led us to the assumption that groups of neurons (i.e., the pools) code for specific stimulus features as well as for abstract rules in the tests we aim to simulate.

The design of the model (compare Fig. 5) reflects the main aspects of the WCS tests discussed above:

1. A certain set of visual stimuli requires a specific response of the subject based on a given rule. We will consider a test setup where a first stimulus consisting of two different feature dimensions (color and shape) is presented to the model followed by a second stimulus.
2. After the presentation of the second stimulus, a response is required: ‘Match’, if the presented pair of stimuli matches with respect to the currently relevant feature dimension (i.e., the relevant rule), ‘Non-Match’ otherwise.
3. The relevant rule might change with or without notice requiring a different answer for the same set of stimuli.

Hence, the model comprises two pools serving as ‘rule pools’, representing two different, possible active rules and four stimulus-specific pools, representing two times two different stimulus properties (i.e., two different shapes and two different colors). A stimulus with certain features is presented to the model (e.g., shape number one, color number two) by adding an extra Poisson input to the specific pools (i.e., to shape pool No. 1 and color pool No. 2). Thus, the external AMPA-mediated input to the neurons within this specific pool is increased to $v_{ext} + \lambda_{stimulus}$. To raise and hold competition, the rule pools receive continuously a low attentional biasing input, $\lambda_{bias}$. During the simulations, the spiking dynamics of the different pools indicate which answer the model would give to a presented series of two stimuli. This answer may be ‘Match’ or ‘Non-Match’, depending on the active rule and whether the stimulus series has the same feature (‘Match’) with respect to the according dimension or not (‘Non-Match’). At the end of the trial, we introduce an unspecific extra external input to the network representing the feedback the model would receive to the previously given answer. The feedback input is provided simultaneously to both of the rule pools, thus $v_{ext}$ is increased by $\lambda_{bias}$ and $\lambda_{feedback}$. In case of a correct answer, we refer to the feedback input as ‘positive feedback’ and ‘negative feedback’ in case of an incorrect answer. However, the feedback input itself is in both cases an external, unspecific AMPA-mediated input to both of the rule pools, differing just in the amount of the value.

We further refer to the two rule pools together as ‘rule module’ and to the two times two stimulus-specific pools as two ‘memory modules’. They are called memory modules because their pools serve as a kind of working memory for the stimulus dimensions.

Each pool selective for a specific function consists of a fraction of excitatory neurons, $f N_{E}$. We chose $f = 0.0625$, thus each selective pool incorporates 100 neurons. The remaining excitatory neurons are organized into a pool called ‘non-selective’ pool. This pool is used to introduce some noise into the network and helps generating the almost Poisson spike firing patterns of neurons in the simulation that are a property of many neurons observed in the cortex (Brunel and Wang, 2001). The inhibitory neurons are grouped to form one inhibitory pool which implements a global competition between all neurons in the network.

The synaptic coupling strength, that is, the weight between the inhibitory neurons and the excitatory neurons, is set to 1 as is the weight between the neurons of all pools and the non-selective pool neurons. The values of the remaining weights (see as well Table 1) were varied using a mean field analysis. The mean field analysis (see Brunel and Wang, 2001; Loh, 2003) provides the possibility to compute the neuronal behavior of the network under stationary conditions and therefore allows to gain rough insights into the consequences of changing a certain weight. The complete set of mathematical equations describing the neuronal model, the parameter values and the mean field analysis are presented in Appendices A and B. In Appendix C concrete values for the specific weights obtained by the mean field analysis are discussed (Appendix C.1) as well as

![Image](https://example.com/figure4.png)

Fig. 4. Time course of percentage increase of the fMRI signal in the left and right inferior frontal sulci of one subject as determined by Konishi et al. (1999) for the original and instruction condition. Konishi et al. found that the increase in the fMRI signal is significantly lower for the instruction condition compared to the original condition. (Reprinted from Konishi et al. (1999), Fig. 4B.)
Having found an appropriate set of weights we are able to run the neurodynamical simulations which enable us to study the behavior of the various pools over time and to compute the fMRI signal. There is an ongoing discussion whether the fMRI-signal reflects synaptic activity or the spiking rate in the measured brain region (see Logothetis et al., 2001; Heeger and Ress, 2002 for a review). In our approach, we used the synaptic activity, calculated as the absolute sum of the three different synaptic currents (AMPA, NMDA and GABA) occurring in the model neurons, as a base to compute the resulting fMRI signal using the hemodynamic response function as estimated by Glover (1999). This approach yields qualitative and quantitative results which are comparable to the experimentally determined fMRI signal change. A similar approach was used already successfully by Deco et al. (2004). Glover proposed that a good estimation of the hemodynamic response to a certain event is given by:

\[ h(t) = e^{t^2} - 3e^{2t^2} \]

\[ c_i = \frac{1}{\text{max}(t^2, e^{t^2})} \]

with \( n_1 = 6.0, t_1 = 0.9 \text{ s}, n_2 = 12.0, t_2 = 0.9 \text{ s} \) and \( a_2 = 0.2 \) as verified in a range of experiments (see Glover, 1999). Thus, the fMRI signal change can be calculated as:

\[ S_{\text{fMRI}}(t) = \int_0^\infty h(t-t')I_{\text{syn}}(t')dt' \]

with \( I_{\text{syn}} = \text{abs} (I_{\text{NMDA}}) + \text{abs} (I_{\text{AMPA}}) + \text{abs} (I_{\text{GABA}}) \). \( I_{\text{syn}} \) is considered for 100 ms intervals and normalized with the mean value of synaptic activity occurring during the simulations. However, it is possible to take the spiking rate as well as the base for calculating the resulting fMRI signal change. In this case, we calculate:

\[ S_{\text{fMRI}}(t) = \int_0^\infty h(t-t')SR(t')dt' \]

where \( SR \) equals the normalized, summed spiking rate of all pools.

**Results**

**Simulation results**

For the simulation of the spiking dynamics, we chose the following weight set (for the motivation to use this weight set and further parameter variations please refer to Appendix C):

- \( w_+ = 2.1 \)
- \( w_- = w_{\text{num}} = 0.1 \)
- \( w_{\text{rr}} = 1.2, w_{\text{non}} = 1.0 \)
- \( w_{\text{mrr}} = 1.0, w_{\text{mm}} = 1.1 \)

This weight configuration leads to the memory pools competing with each other (\( w_+ \) and \( w_{\text{num}} \)). Also, the rule pools are competitively active (\( w_- \)). There is no competition between the rule pools and the memory pools but each rule pool supports ‘its’ memory pools (\( w_{\text{rr}} \)) to a greater extend than the contrary ones (\( w_{\text{mrr}} \)). The memory pools in turn support the rule pools but this time in the opposite direction (\( w_{\text{mrr}} \) and \( w_{\text{mm}} \)). This is an important factor for the rule change as explained below. The weight \( w_+ \) assures the bistable properties of the pools (see well Appendix C.3).

For the simulations, a cue stimulus was presented for 500 ms, followed by the presentation of a sample stimulus for 500 ms, followed by a feedback period of another 500 ms. During

---

**Table 1**

Overview of synaptic coupling strength (weights) within the model architecture

<table>
<thead>
<tr>
<th>Pools</th>
<th>RC</th>
<th>RS</th>
<th>C1</th>
<th>C2</th>
<th>S1</th>
<th>S2</th>
<th>NonSp</th>
<th>Inhib</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC</td>
<td>( w_+ )</td>
<td>( w_- )</td>
<td>( w_{\text{rm}} )</td>
<td>( w_{\text{rrm}} )</td>
<td>( w_{\text{rr}} )</td>
<td>( w_{\text{mrr}} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>RS</td>
<td>( w_- )</td>
<td>( w_+ )</td>
<td>( w_{\text{rn}} )</td>
<td>( w_{\text{rnr}} )</td>
<td>( w_{\text{mrr}} )</td>
<td>( w_{\text{mnm}} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C1</td>
<td>( w_{\text{mm}} )</td>
<td>( w_{\text{mm}} )</td>
<td>( w_- )</td>
<td>( w_{\text{rm}} )</td>
<td>( w_{\text{mrm}} )</td>
<td>( w_{\text{mnm}} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C2</td>
<td>( w_{\text{mm}} )</td>
<td>( w_{\text{mm}} )</td>
<td>( w_- )</td>
<td>( w_{\text{rm}} )</td>
<td>( w_{\text{mrm}} )</td>
<td>( w_{\text{mnm}} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S1</td>
<td>( w_{\text{mm}} )</td>
<td>( w_{\text{mm}} )</td>
<td>( w_- )</td>
<td>( w_{\text{rm}} )</td>
<td>( w_{\text{mrm}} )</td>
<td>( w_{\text{mnm}} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S2</td>
<td>( w_{\text{mm}} )</td>
<td>( w_{\text{mm}} )</td>
<td>( w_- )</td>
<td>( w_{\text{rm}} )</td>
<td>( w_{\text{mrm}} )</td>
<td>( w_{\text{mnm}} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NonSp</td>
<td>( w_{\text{mm}} )</td>
<td>( w_{\text{mm}} )</td>
<td>( w_- )</td>
<td>( w_{\text{rm}} )</td>
<td>( w_{\text{mrm}} )</td>
<td>( w_{\text{mnm}} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Inhib</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

RC denotes the color rule pool; C1 denotes color pool number one. NonSp denotes the non-specific pool of neurons and Inhib the pool with the inhibitory neurons. The weight between two pools of one module is denoted \( w_- \), the weight between the two memory modules is named \( w_{\text{mm}} \). Furthermore we differentiated the various connections between the rule module and the memory modules. Each rule pool drives a certain memory module and this weight is called \( w_{\text{mm}} \), \( w_{\text{mm}} \) for the opposite direction. That is, the rule pool called RC for color rule provides a specific input for the pools coding for the color dimension of the stimulus and vice versa. Contrary, the weight between a rule pool and the other memory module is denoted \( w_{\text{mm}} \) and again \( w_{\text{mm}} \) for the opposite direction. As we have to make sure that the network only elicits spontaneous activity if there is no external input besides the background noise, we used the weights from the non-selective pool to the selective pools to compensate for the varied other weights. That is, these weights were calculated so that the sum of all weighted input connections to a given pool equals 1. Thus, the weight \( w_{\text{mm}} \) from the non-selective pool to a selective pool \( x \) turns out to be:

\[ w_{\text{mm}} = 1 - \sum_i f_i w_{\text{ir}} \]

where \( w_{\text{ir}} \) denotes the weight from the selective pool \( i \) to the pool \( x \) and \( f_i \) equals the respective fraction of neurons in pool \( i \).
the stimulus periods, the corresponding memory pools received the extra input $\lambda_{\text{stimulus}}$ of 150 Hz (s. above). The rule pools received a continuous unspecific input $\lambda_{\text{bias}}$ of 60 Hz. During the feedback period, the additional feedback input was provided unspecifically to the rule pools. The feedback input $\lambda_{\text{feedback}}$ was chosen to be 150 Hz, if the answer the model would give to the presented pair of stimuli was correct (positive feedback) and 300 Hz, if it was incorrect (negative feedback). The feedback input to the rule pools was provided after the feedback delay $T_{\text{feedback}}$.

Fig. 5. Neuronal model for the simulations. $I =$ Pool of inhibitory neurons. $N =$ pool of non-specific neurons. $RC =$ Pool for the color rule, $RS =$ pool for the shape rule. $C1 =$ color pool number one, $S1 =$ shape pool number one, $C2 =$ color pool number two, $S2 =$ shape pool number two. The network is fully connected. For the strengths of the unprinted connections to and from the inhibitory and non-specific pools please refer to the text and Table 1. External input is provided to the rule pools (bias input, $\lambda_{\text{bias}}$, and feedback input for correct and wrong answers, $\lambda_{\text{feedback}}$) and stimulus-specific input to the respective color and shape pools.

Fig. 6. Spiking dynamics of the rule pools, the color and shape pools during the simulations of the original WCST. The color and shape pools show activity according to the active rule and the presented stimuli. After a negative feedback, the activity of the rule pools inverts: The previously active pool becomes inactive and the previously inactive pool becomes active (see e.g., after 13 s).
Hz otherwise (negative feedback). The input was provided in addition to the 2.4 kHz external input, the neurons already receive as described above. In the present configuration, the model would always give a correct answer (according to the currently active rule), thus the negative feedback was provided only in case of a dimensional change of the valid rule.

Fig. 7. Extract of the spiking dynamics during WCST simulations showing how it is possible to differentiate model answers: decreasing and increasing spiking rates after the presentation of stimulus S1 in the left diagram would lead to a ‘non-match’ answer. In the right diagram, we see a constant or slightly increasing level of the spiking rate after the presentation of stimulus C2 which would turn out to a ‘match’ answer.

Fig. 8. Spiking dynamics of the rule pools, the color and shape pools during the simulations of the instruction WCST. Contrary to the simulation of the original WCST here only the new relevant rule pool receives the extra external feedback input. This circumstance leads to the straight increase of the spiking rate of the corresponding rule pool (see, e.g., after 13 s).
Fig. 6 shows the resulting spiking dynamics for a simulation consisting of 24 single trials where a rule change was assumed after eight trials: the top diagram shows the spiking train of the rule pools. First, the shape rule became active and accordingly the shape memory pools showed transient activity corresponding to the shape dimension of the presented stimuli (bottom diagram of Fig. 6). Due to correct answers, the positive feedback input was provided after every trial to the rule pools. This feedback led only to an increase of the spiking rate of the active rule pool although it was provided to both pools. Thus, the provision of the unspecific – positive feedback led to a kind of reward, a strengthening – of the currently active rule. After eight correct trials, we assumed a rule change, thus the negative feedback input was provided to the rule pools (top diagram of Fig. 6, e.g., after 13 s). This time, the application of the higher negative feedback led to the desired rule change, thus afterwards the color rule becomes active and, accordingly, the color memory pools started to react corresponding to the presented stimuli. The reason why only the negative feedback of 300 Hz and not the positive feedback of 150 Hz led to a rule change might be explained as follows: the higher input to the rule pools led to a strong competition between the two pools which resulted in an advantage for the previously inactive rule pool. This pool received an higher amount of input from the memory pools due the weight $w_{\text{nonr}}$ being higher than the weight $w_{\text{mr}}$. In other words, we can say that the memorization of the previously relevant stimulus dimension (i.e., represented by the corresponding active memory pools) led to the activation of the new relevant dimensional rule.

Another interesting point when considering the spiking dynamics of the memory pools in Fig. 6 is the spiking behavior in the first trial after the rule change. In these trials, the memory pool representing the now invalid stimulus dimension, is to a certain degree active as well. This circumstance might hold an explanation for set-shifting costs or errors occurring immediately after a rule change.

An important point – especially in view of the envisaged latter construction of an explicit ‘answer module’ – is the discrimination of the potential model answers. Fig. 7 shows extracts of the spiking trains of the shape pools (left) and color pools (right) in order to illustrate how the model answers can be differentiated. The left diagram shows the spiking trains when first the stimulus S2 was presented and afterwards stimulus S1. The right diagram shows the spiking train in a trial where first the stimulus C2 was presented and afterwards the same stimulus C2. The time course of these spiking trains allows to differentiate the model answers: during the sample stimulus period, we see decreasing activity of shape pool S2 and increasing activity of shape pool S1. Thus, the model answer would turn out to ‘Non-Match’ for the left diagram. For the case where the same stimulus feature was presented in the cue and sample period (right diagram), we see a constant level of activity, hence, in this case, the model answer would be ‘Match’. These spike trains show on the one hand that it is possible to use a simple algorithm to determine the model answers during the simulations.

![Graphs showing spiking dynamics](image)
Hence, it is not necessary to analyze the spike trains off-line and calculate the answers after the simulation. The answers can be calculated on line during the simulations and thus the feedback can be provided accordingly. On the other hand, they provide the base for the construction of an explicit answer module whose primary task will be the detection of a decreasing spiking rate.

For the simulations of an instruction variant of the WCST comparable to the experiments run by Konishi et al. (1999), we modified the feedback input to the rule pools. In case of a rule change, we provided now an ‘instruction feedback’ of as well 300 Hz only to the new relevant rule pool. This way of feedback provision corresponds to subjects being explicitly informed about the new valid rule in the experiments. The resulting spiking dynamics are shown in Fig. 8: The new valid rule (see e.g., after 13 s) becomes activated immediately and the memory pools now follow the new active rule instantly. These circumstances might be translated into lower set-shifting costs and error rates for subjects in this variant of the WCST.

fMRI calculations

For the calculation of the resulting fMRI signal during the simulations, we compared the time course of the signal change when using (a) the synaptic activity and (b) the summed spiking rates of all pools as the base for the fMRI calculations. For this purpose, we simulated 110 trials (cue-sample-feedback periods) and changed the rule after every 10 correct answers in order to gain a sufficient amount of data and an adequate interval between two rule changes. Following the method used by Konishi et al. (1999), we sampled the calculated fMRI data every 2 s time locked to the provision of the negative feedback (first indication of the rule change) which we called as well ‘time zero’. Thus, we used the calculated fMRI data at time points -5, -3, -1, 1, 3, 5, 7, 9 and 11 s relative to time zero.

Fig. 9 shows the individual time course of the fMRI signal for nine single rule changes and the resulting mean fMRI signal change graph when using the synaptic activity as the base for the calculation of the fMRI signal for the original WCST and the instruction variant. The mean fMRI signal shows a peak transient increase of about 0.6% roughly 6 s after the rule change event for the original WCST variant (top diagrams in Fig. 9). For the instruction variant, the peak value is about 0.35%. Nakahara et al. (2002) and Konishi et al. (1999) obtained comparable results in their experiments (see Figs. 2 and 4), differing only to a certain, small degree for the period of -5 to -1 s relative to time zero. For this period of time, our calculations resulted in a decreasing fMRI signal in contrast to the results of Konishi et al. (1999) as shown in Fig. 4. However, when averaging the fMRI data for all seven subjects, Konishi et al. also obtained a decreasing fMRI signal change (Konishi et al. (1999), Figs. 4C and D). Therefore, we considered the discrepancy between the fMRI calculations and the experimentally determined fMRI data for the time period of -5 to -1 s to be of only minor importance. Thus, in summary, the simulation results fit qualitatively and even quantitatively very well with the experimentally determined fMRI data. It is noteworthy that the qualitative correspondency not only holds for the mean fMRI signal change time course across all rule changes but as well for

Fig. 10. Same as Fig. 9 with the normalized summed spiking rate used as the base for the calculation of the fMRI signal.
the single events. Due to a rather limited variance, every single fMRI signal change time course (left diagrams in Fig. 9) is qualitatively comparable to the experimentally determined fMRI time course (Figs. 2 and 4).

This is not the case when considering the spiking rates occurring during the simulations as the base for the calculation of the fMRI signal (Fig. 10). The mean fMRI signal change time course now shows an even larger decrease for the time period of −5 to −1 s relative to time zero and a just rather small increase afterwards. Besides this qualitative and quantitative discrepancy between the mean fMRI signal change time course (Fig. 10, right) and the experimentally determined fMRI data, the single time courses spread to a very large degree (Fig. 10, left diagrams).

These results show on the one hand that we developed a reasonable neuronal model for the simulation of WCST-like tasks and the resulting fMRI signal and support on the other hand the hypothesis that fMRI data reflect synaptic activity rather than spiking rates.

Discussion

In the present work, we introduced a neuronal model based on detailed biophysical descriptions of single neurons which is capable to simulate WCS-like tasks and to account for the respective fMRI data. The identification of neuronal pools in the model was based on single cell recordings in behaving monkeys. The weights between the pools were adapted using a mean field analysis. The rule switching was achieved by supplying unspecific feedback input to the rule pools and thus did not require a weight modification process as was necessary by the models discussed in the Introduction. An hemodynamic response function was used to estimate the fMRI signal change during the simulations of the trials. The results obtained from the calculations of the fMRI signal fitted the experimentally determined data very well and supported the conclusion that our model describes the neuronal dynamics involved in WCS-like tasks quite well. Furthermore, the hypothesis that synaptic activity rather than spiking rate is reflected in experimentally determined fMRI data was encouraged.

The proposed model also offers an explanation for subjects’ inability to change to a new valid rule: improper memorization of stimulus dimensions might at least theoretically account for a temporal difficulty to switch to a new rule. If the memory pools for some reasons do not show a sufficient amount of activity, the switch to a new rule obviously would not be possible. This circumstance could lead to more than one erroneous trial to complete a rule change. Besides, especially when considering patients with permanent rule switching difficulties, the model suggests that the feedback input to the rule pools might be too low to provide the required competition which effectively ended in a rule change in our simulations.

It is remarkable that the neurodynamical model presented in this work did not require a “global pulse of excitation”, initiated by a motor response, to be delivered to the network in order to reset the PFC as criticized by Cohen et al. (2002, p. 224). We simply deliver an unspecific feedback input to the rule pools which effectively leads to the rule change and thus enables the memorization of the new relevant feature dimension. Besides, the model would allow as well “the successive encoding of items during a serial recall task”: within a trial, the working memory representation (i.e., the active memory pool) of a previously presented stimulus is ‘erased’ as soon as a new competitive stimulus is presented. Therefore, in this work, we do not face the limitation that a (motor) response is necessary to enable the update of working memory representations (Cohen et al., 2002, p. 224).

However, the model faces some other limitations:

First, currently only two different rules are supported. But when considering cognitive flexibility in terms of set-shifting, which is our main focus for the near future, it is not important to implement three different rules. Cognitive flexibility is required already when dealing with two different rules. Therefore, we consider the implementation of three rules and the modeling of the corresponding trial-and-error component for further study.

A second, more important issue is the development of an explicit ‘answer module’ and the consideration of error rates. It would be highly desirable that the model emits error rates corresponding to subject error rates. Here, the current model configuration offers already some promising perspectives: a lower value for the parameter \( \lambda_{\text{bias}} \) would lead to arbitrary rule changes. In this case, the model would give an incorrect answer (due to the irrelevant stimulus dimension because of the arbitrary rule change) and the negative feedback provided afterwards would lead the model to return to the correct dimension again. Initially, it is apparently necessary to analyze subject errors, that is, the unmotivated errors, which are not related to a rule change. At present, the model offers the following two possibilities why subjects might make mistakes: they might respond according to the wrong stimulus dimension or they did not memorize a presented stimulus properly. Both factors are to be integrated into the model by varying the parameters \( \lambda_{\text{bias}}, \lambda_{\text{feedback}} \) and \( \lambda_{\text{stimulus}} \) (see as well Appendix C).

A third issue is the relevance of the irrelevant stimulus dimension. Wallis et al. (2001) and Chen et al. (2001) found that irrelevant rule and stimulus neurons show a certain degree of activity, while in our simulations so far, the irrelevant memory and rule pools were almost completely silent. Of course, it has to be taken into consideration that the experimental setups of Wallis et al. (2001) and Chen et al. (2001) differed from the setups we modeled as they did not require an actual switch to another rule and hence the respective cognitive flexibility within an experiment. Thus, it is first of all at least questionable whether the rule neurons Wallis et al. found are comparable to the rule neurons we introduced in the presented model. To reliably determine whether the irrelevant stimulus dimension affects subject responses (and thus whether the irrelevant stimulus pools should show as well a certain level of activity), we consider the evaluation of subject response time distributions as a more promising approach: if the irrelevant stimulus dimension has no effect, the response times would not differ when both stimulus dimensions require the same answer (e.g. match in color, match in form) compared to the stimulus dimensions requiring different answers (e.g., match in color, non-match in form). The present model configuration does not show such effects. Due to the parameter \( w_{\text{stim}} = 0.1 \), the stimulus dimensions are strongly competitive and thus response times emitted by the answer module under development would not vary, respectively. However, higher values for \( w_{\text{stim}} \) will allow parallel activity and thus, respectively varying response times.
Finally, the feedback signals, currently modeled as external AMPA input to the rule pools, are quite clearly to be translated into responses of dopamine neurons. Here, we consider primarily the work of Schultz (1998) to be of major importance and the main issue how to correlate dopamine releases to the neuronal pools of our model in order to elicit the desired test behavior. This translation into dopamine responses is of course as well an important point concerning the prediction of patient behavior.

To gain the necessary experimental data for the extended modeling of error rates and response times, we are working at present on behavioral experiments using a ‘Delayed-Matching-to-Sample’ (DMS) WCS-like test setup comprising two different rules. Subjects have to state whether a presented series of two stimuli match or do not match with respect to the feature dimension ‘same object’ or ‘same location’ with and without explicit instruction of the new valid rule. The goal is to model subject behavior in these tests where we expect as well that response time distributions enable us to draw reliable conclusions concerning the activity of neurons reflecting irrelevant stimulus dimensions. Further on, we aim to simulate and predict patients’ behavior (e.g., patients with focal brain lesions and schizophrenics).

Acknowledgments

This work was supported by the German Science Foundation (DFG), Project Number FOR 480/1-1. Special thanks are directed to Rita Almeida and Marco Loh (University Pompeu Fabra, Department of Computational Neuroscience, Barcelona) and to Joe Krummenacher (Ludwig-Maximilian University Munich, Department of Psychology) for valuable discussion and advise. Gustavo Deco acknowledges support through the German Ministry for Research (BMBF grant 01HBC01A).

Appendix A. Biased competition network of integrate-and-fire neurons

The calculation of the spiking dynamics in the presented model is based on the framework of Brunel and Wang (2001) where each neuron is described by an integrate-and-fire model. Thus the sub-threshold membrane potential $V(t)$ is given by:

$$C_m \frac{dV(t)}{dt} = -g_m(V(t) - V_L) - I_{\text{syn}}(t)$$  \hspace{1cm} (A.1)

where $C_m$ is the membrane capacitance, which turns out to be 0.5 nF for excitatory neurons and 0.2 nF for inhibitory neurons; $g_m$ is the membrane leak conductance: 25 nS for excitatory neurons and 20 nS for inhibitory neurons; $V_L$ is the resting potential of $-70 \text{ mV}$ and $I_{\text{syn}}$ is the total synaptic current flow into the cell. When the membrane potential $V(t)$ reaches the value $V_{thr} = -0.50 \text{ mV}$ a spike is generated and afterwards the membrane potential is set to the value $V_{reset} = -0.55 \text{ mV}$.

The synaptic current is given by the sum of glutamatergic, AMPA ($I_{\text{AMPA,rec}}$) and NMDA ($I_{\text{NMDA,rec}}$) mediated, recurrent excitatory currents, one AMPA ($I_{\text{AMPA,ext}}$) mediated external excitatory current and one inhibitory GABAergic current ($I_{\text{GABA}}$):

$$I_{\text{syn}}(t) = I_{\text{AMPA,ext}}(t) + I_{\text{AMPA,rec}}(t) + I_{\text{NMDA,rec}}(t) + I_{\text{GABA}}(t).$$ \hspace{1cm} (A.2)

where

$$I_{\text{AMPA,ext}}(t) = g_{\text{AMPA,ext}}(V(t) - V_E) \sum_{j=1}^{N_E} s_j^{\text{AMPA,ext}}(t)$$  \hspace{1cm} (A.3)

$$I_{\text{AMPA,rec}}(t) = g_{\text{AMPA,rec}}(V(t) - V_E) \sum_{j=1}^{N_E} w_j s_j^{\text{AMPA,rec}}(t)$$  \hspace{1cm} (A.4)

$$I_{\text{NMDA,rec}}(t) = g_{\text{NMDA}}(V(t) - V_E) \sum_{j=1}^{N_E} w_j s_j^{\text{NMDA}}(t)$$  \hspace{1cm} (A.5)

$$I_{\text{GABA}}(t) = g_{\text{GABA}}(V(t) - V_i) \sum_{j=1}^{N_I} s_j^{\text{GABA}}(t)$$  \hspace{1cm} (A.6)

with $V_E = 0 \text{ mV}$ and $V_i = -70 \text{ mV}$, $w_j$ are the synaptic weights as specified in Table 1. For the conductances, we used the following values:

$g_{\text{AMPA,ext}} = 2.08 \text{ nS}$, $g_{\text{AMPA,rec}} = 0.052 \text{ nS}$, $g_{\text{NMDA}} = 0.1635 \text{ nS}$ and $g_{\text{GABA}} = 0.625 \text{ nS}$ for excitatory neurons and $g_{\text{AMPA,ext}} = 1.62 \text{ nS}$, $g_{\text{AMPA,rec}} = 0.0405 \text{ nS}$, $g_{\text{NMDA}} = 0.129 \text{ nS}$ and $g_{\text{GABA}} = 0.4865 \text{ nS}$ for inhibitory neurons. These values are obtained from Brunel and Wang (2001) in consideration of the number of neurons used in our network ($N_E = 1600$ excitatory neurons and $N_I = 400$ inhibitory neurons). Brunel and Wang estimated the conductances so that in an unstructured network the excitatory neurons have a spontaneous spiking rate of 3 Hz and the inhibitory neurons a spontaneous rate of 9 Hz as found experimentally. The NMDA synaptic current has an extra component, taking into account that NMDA currents are mediated, recurrent excitatory currents, one AMPA ($I_{\text{AMPA,ext}}$) mediated external excitatory current and one inhibitory GABAergic current ($I_{\text{GABA}}$):
\[
\frac{dV_{\text{GABA}}(t)}{dt} = - \frac{\Delta_{\text{GABA}}(t)}{\tau_{\text{GABA}}} + \sum_k \delta(t - t_k^j), \tag{A.11}
\]

where \(\tau_{\text{NMDA,decay}} = 100\) ms is the decay time for NMDA synapses, \(\tau_{\text{AMPA}} = 2\) ms for AMPA synapses and \(\tau_{\text{GABA}} = 10\) ms for GABA synapses; \(\tau_{\text{NMDA,rise}} = 2\) ms is the rise time for NMDA synapses (the rise times for AMPA and GABA are neglected because they are smaller than 1 ms) and \(\alpha = 0.5\) ms\(^{-1}\). The sums over \(k\) represent a sum over spikes formulated as \(\delta(t)\) emitted by presynaptic neuron \(j\) at time \(t_k^j\).

**Appendix B. Mean field parameter approximation**

For the mean field approximation we used the implementation of Loh (2003) which was derived in Brunel and Wang (2001) using the assumption that the network of integrate-and-fire neurons is in a stationary state. Thus, the calculation of the potential of a neuron can be simplified to:

\[
\tau_x \frac{dV(t)}{dt} = - V(t) + \mu_x + \sigma_x \sqrt{\tau_x} \eta(t) \tag{B.1}
\]

where \(V(t)\) is the membrane potential, \(x\) labels the populations, \(\tau_x\) is the effective membrane time constant, \(\mu_x\) is the mean value the membrane potential would have in the absence of spiking and fluctuations, \(\sigma_x\) measures the magnitude of the fluctuations and \(\eta\) is a Gaussian process with absolute exponentially decaying correlation function with time constant \(\tau_{\text{AMPA}}\). The quantities \(\mu_x\) and \(\sigma_x^2\) are given by:

\[
\mu_x = \frac{(T_{\text{ext}}\nu_{\text{ext}} + T_{\text{AMPA}}n_x + \rho_1 N_x V_E + \rho_2 N_x (V') + T_I w_{\text{ex}} v_I V_I + V_{\text{reset}})}{B_x} \tag{B.2}
\]

\[
\sigma_x^2 = \frac{g^2_{\text{AMPA,ext}} (\langle V \rangle - V_E)^2 N_x V_{\text{ext}} + \tau_{\text{AMPA}}^2}{g_m \tau_m^2} \tag{B.3}
\]

where \(w_{\text{ex}}\) are the weights from the inhibitory neurons to the pool \(x\), \(V_{\text{ext}} = 3\) Hz, \(v_I\) is the spiking rate of the inhibitory pool, \(\tau_m = C_m/g_m\) with the values for the excitatory or inhibitory neurons depending of the pool considered.

The remaining quantities are calculated as follows:

\[
S_x = 1 + T_{\text{ext}}\nu_{\text{ext}} + T_{\text{AMPA}} n_x + (\rho_1 + \rho_2) N_x + T_I w_{\text{ex}} v_I
\]

\[
\tau_x = \frac{C_m}{g_m S_x}
\]

\[
n_x = \sum_{j=1}^{p} f_j w_{j,x} v_j
\]

\[
N_x = \sum_{j=1}^{p} f_j w_{j,x} \psi(v_j)
\]

\[
\psi(v) = \frac{v_{\text{NMDA}}}{1 + v_{\text{NMDA}}} \times \left(1 + \frac{1}{1 + v_{\text{NMDA}}} \sum_{n=1}^{\infty} \left(\frac{1}{n!}\left(\frac{- \mathbf{\tau}_{\text{NMDA,rise}}}{1 + \mathbf{\tau}_{\text{NMDA}}} + \mathbf{\tau}_{\text{NMDA,decay}}\right)^n T_n(v)\right)\right)
\]

\[
T_n(v) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \frac{\mathbf{\tau}_{\text{NMDA,rise}}^{n+k} (1 + v_{\text{NMDA}})}{\mathbf{\tau}_{\text{NMDA,rise}}(1 + \mathbf{\tau}_{\text{NMDA}}) + \mathbf{\tau}_{\text{NMDA,decay}}}
\]

\[
\mathbf{T}_{\text{ext}} = \frac{g_{\text{AMPA,ext}} N_{\text{ext}} \nu_{\text{AMPA}}}{g_m}
\]

\[
\mathbf{T}_{\text{AMPA}} = \frac{g_{\text{AMPA,rec}} N_{\text{E,AMPA}}}{g_m}
\]

\[
\rho_1 = \frac{g_{\text{NMDA}} N_{\text{E}}}{g_m}
\]

\[
\rho_2 = \beta \frac{g_{\text{NMDA}} N_{\text{E}} (V_E - V_{\text{reset}})(J - 1)}{g_m J^2}
\]

\[
J = 1 + \gamma \exp(-\beta(V_j))
\]

\[
T_I = \frac{g_{\text{GABA}} N_{\text{I,AMPA}}}{g_m}
\]

\[
\langle V_x \rangle = \mu_x - (V_{\text{reset}} - V_{\text{reset}}) v_x \tau_x,
\]

where \(p\) is the number of excitatory pools, \(f_x\) is the fraction of neurons in the excitatory pool \(x\), \(w_{j,x}\) the weight of the connections from pool \(x\) to pool \(j\), \(\nu_x\) is the spiking rate of the inhibitory pool, \(\tau_r = C_W/g_m\) with the values for the excitatory or inhibitory neurons depending of the pool considered.

The spiking rate of a pool as a function of the defined quantities is then given by:

\[
\nu_x = \phi(\mu_x, \sigma_x), \tag{B.5}
\]

where

\[
\phi(\mu_x, \sigma_x) = \left(\tau_p + \tau_x\int_{\frac{\mu_x}{\mu_x}}^{\infty} du \sqrt{\pi} \exp(u^2)[1 + \text{erf}(u)]\right)^{-1}
\]

\[
\alpha(\mu_x, \sigma_x) = \frac{(V_{\text{reset}} - \mu_x)}{\sigma_x} \left(1 + 0.5 \frac{\tau_{\text{AMPA}}}{\tau_x}\right)
\]

\[
+ 1.03 \sqrt{\frac{\tau_{\text{AMPA}}}{\tau_x} - 0.5 \frac{\tau_{\text{AMPA}}}{\tau_2}}
\]

\[
\beta(\mu_x, \sigma_x) = \frac{(V_{\text{reset}} - \nu_x)}{\sigma_x}
\]

with \(\text{erf}(u)\) being the error function and \(\tau_p\) the refractory period which is considered to be 2 ms for excitatory neurons and 1 ms.
for inhibitory neurons. To solve the equations defined by Eq. (B.5) for all $x$, we integrate numerically Eq. (B.4) and the differential equation below, which has fixed point solutions corresponding to Eq. (B.5):

$$\tau \frac{dv_x}{dt} = -v_x + \phi(\mu_x, \sigma_x).$$  \hspace{1cm} (B.6)

For all simulation steps described in Appendix C.1, the mean-field Eqs. (B.4) and (B.6) were integrated using the Euler method with step size 0.2 and 5000 iterations.

Appendix C. Parameter evaluations

In the following sections, we describe the evaluation of selected model parameters. For evaluation purposes we kept a certain set of parameters fix while varying another one. In this way, it is possible to examine the model behavior at boundaries. It should be noted however that in varying the complete set of parameters at once (i.e., the weights and the external input) other combinations might expose which are as well capable to simulate the experiments as necessary. However, the principal relationship of the parameters and the principal ability to simulate the experiments are first of all necessary. Thus, to look for a complete set of different model parameters would be very time consuming while not providing any further insights.

In Appendix C.1, we describe how we set up the mean field analysis in order to find a fitting set of weights. In Appendix C.2, we evaluate variations of the external input the model receives; in Appendix C.3, we describe consequences of varying the parameter $w_{-}$.  

C.1. Evaluation of weight combinations using the mean field analysis

The mean field analysis allows to study the behavior of the network under stationary conditions, that is, for a given set of weights and external input frequencies the resulting average spiking rates of the various pools are calculated. As we need to estimate the behavior of the network for a series of input conditions, reflecting the presentation of a stimulus sequence and the feedback to the response of the network, we have to use a series of mean field parameter scans as well. Therefore, we defined the following exemplary set of test conditions:

1. a cue stimulus is presented to the network consisting of color number one and shape number one (C1, S1);
2. next, a second stimulus is presented to the network, called sample stimulus, consisting of color number one and shape number two (C1, S2);
3. we assume a correct answer of the network so we add an unspecific positive feedback input to the rule pools;
4. the next cue is presented to the network, consisting of color number two and shape number one (C2, S1);
5. the next sample stimulus is presented, this time consisting of color number two and shape number two (C2, S2);
6. now, we assume a rule change so we add an unspecific negative feedback input to the rule pools RC and RS;
7. another cue stimulus is presented to the network: color number two and shape number one (C2, S1).

For steps (2) to (7), we configured the initial spiking rates of the pools for the mean field analysis as calculated in the previous step. We chose the bias input to the rule pools to be 60 Hz (external AMPA input), the positive feedback input to be 150 Hz, the negative feedback input to be 300 Hz and the stimulus input to be 150 Hz. Considering that all neurons already receive an external input of 2.4 kHz totally from the 800 neurons outside the network modeled, the provision of the 300 Hz negative feedback input means, for example, that these neurons now receive 2.4 + 0.3 kHz of external input. Thus, the total input increase is rather low.

The weights of all pools to and from the inhibitory pool were set to 1 as were the weights from all pools to the non-selective pool of neurons (see as well Table 1). The weight $w_{+}$ was set to 2.1 and the remaining weights were varied in the range from 0.1 to 1.5. Consistent with the WCS task the network has to carry out, we expected the pools to show the following characteristics:

- After step (1) the rule pools should show a clear different level of spiking activity. The difference should be at least 10 Hz and the pool showing a higher spiking rate is considered to be active. Consistent with the active rule and the presented cue stimulus stimulus one of the memory pools should show a clear level of activity. That is, if the rule pool RC is active we want the memory pool C1 to be active as well.
- After step (2), the same rule pool should be active and one of the memory pools, corresponding to the active rule and the sample stimulus.
- After step (3), we expect a similar activity pattern as after step (2)
- After step (4), we expect the same rule pool to be active as in steps (1), (2) and (3) and additionally, the corresponding memory pool, C2 or S1 according to the active rule.
- After step (5), we expect the same behavior as after step (2)
- We do not expect a special activity pattern after step (6). Due to the stationary conditions in the mean-field analysis both rule pools will show a high level of activity after step (6).
- After step (7), we expect the same behavior as after step (1), however this time we want the contrary rule to be active.

The mean field analysis was configured to complete all seven calculation steps only if the individual criteria were met. However, simulations of the spiking dynamics using example weight combinations (which were deemed successful by the mean field calculations) revealed that only a very limited range of weight settings is able to elicit the desired behavior out of the network. The reason might be that for the above described 7-step-configuration, the timing dynamics play an important role. Contrary to the mean field calculations, in the spiking simulations stimuli are presented only for a limited period of time. Thus, when simulating the spiking dynamics the spiking rates of the pools do not need to converge to the values as estimated by the mean field analysis under stationary conditions. Especially, as we are simulating not only a single stimulus presentation but a series of different input sequences to various pools.

We detected the following constraints:

- $w_{-}$ and $w_{\text{mm}}$ have to be quite low as otherwise the memory pools neither react according to the presented stimuli nor to the active rule.
If $w_{\text{mm}}$ is greater than $w_{\text{m}}$, the wrong memory pools tend to get activated.

The same happens if $w_{\text{mon}}$ is greater than $w_{\text{rm}}$.

If $w_{\text{rm}}$ is too high, the memory pools do not act according to the presented stimuli but only according to the active rule.

$w_{\text{rm}}$ has to be lower than $w_{\text{nonr}}$ in order to drive the rule change.

Therefore we chose $w_{\text{m}}$ and $w_{\text{mm}}$ to be 0.1 and further examined the influence of the weights $w_{\text{rm}}$, $w_{\text{mr}}$, $w_{\text{mon}}$ and $w_{\text{nonr}}$. We set $w_{\text{mon}}$ to be

$$w_{\text{mon}} = w_{\text{rm}} - 0.1/0.2$$

and $w_{\text{nonr}}$ to be

$$w_{\text{nonr}} = w_{\text{mr}} + 0.1/0.2$$

The results are shown in Fig. 11. Only a very limited number of weight combinations completed all seven steps described above successfully. Diagram 1 (top left in Fig. 11) shows the result for $w_{\text{mon}} = w_{\text{rm}} - 0.1$ and $w_{\text{nonr}} = w_{\text{mr}} + 0.1$. Fifteen weight combinations successfully completed step 7, that is, the stimulus pools showed for all configured steps the desired firing patterns and the rule change in step seven succeeded. Further three weight settings completed step five of the above described mean field configuration which means that though the stimulus pools showed the desired firing pattern, the rule change did not work out. Diagrams 2 and 4 (top and bottom right in Fig. 11) show that these weight settings only let to the successful completion of step two, that is, the firing patterns of the pools were not satisfying after the provision of the positive feedback in step three. Thus, the difference of 0.2 between the weights $w_{\text{mr}}$ and $w_{\text{nonr}}$ (as settled in these diagrams) does not seem to fit. Finally, diagram 3 (bottom left) shows three weight combinations which successfully com-

![Fig. 11. Result of mean field analysis for four different weight configurations. We analyzed seven setup situations (steps) using the mean field calculations. Each step describes a configuration of input and start-up firing rates of the neuronal pools reflecting the progress in the WCST. Please refer to Appendix C.1 for a detailed description of the single steps. Diagram 1 (top left) shows that 15 weight combinations successfully completed step seven; further three weight settings completed step five. Diagrams 2 and 4 (top and bottom right) show that these weight settings only let to the successful completion of step two and finally diagram 3 (bottom left) shows three weight combinations which successfully completed step seven and further combinations which only completed step five or rather two.](image-url)
pleted step (7) and further combinations which only completed step (5) or rather step (2).

Again, we cross-checked the mean field results in single simulations of the spiking dynamics to verify that the model elicits the desired test behavior.

The following sets were selected (see Fig. 11):

1. Diagram 1: \( w_{rm} = 1.0, w_{mr} = 1.0 \)
2. Diagram 2: \( w_{rm} = 1.0, w_{mr} = 1.2 \)
3. Diagram 3: \( w_{rm} = 1.1, w_{mr} = 1.0 \)
4. Diagram 4: \( w_{rm} = 1.2, w_{mr} = 1.1 \)
5. Diagram 5: \( w_{rm} = 1.3, w_{mr} = 1.0 \)
6. Diagram 6: \( w_{rm} = 1.3, w_{mr} = 1.1 \)
7. Diagram 7: \( w_{rm} = 1.2, w_{mr} = 1.2 \)
8. Diagram 8: \( w_{rm} = 1.4, w_{mr} = 1.0 \)
9. Diagram 9: \( w_{rm} = 1.2, w_{mr} = 1.0 \)
10. Diagram 10: \( w_{rm} = 1.3, w_{mr} = 1.1 \)
11. Diagram 11: \( w_{rm} = 1.4, w_{mr} = 1.1 \)

For the simulations, the cue stimulus was presented for 500 ms, followed by the presentation of the sample stimulus for 500 ms, followed by a feedback period for as well 500 ms. We did not model the concrete answer of the network to a given pair of stimuli but assumed that it will respond always correct except in cases of a rule change.

It is necessary to analyze the spiking dynamics carefully: The model has to respond clearly to the presented stimuli and according to the active rule in order to provide the possibility to differentiate potential answers. Also, the active rule should change after a negative feedback input to the rule pools. We simulated all possible cue-sample-stimulus combinations and assumed a rule change after ten correct trials. Every simulation comprised 33 single trials.

Additionally, we performed simulations to verify the rule change behavior of the network. For this purpose, we chose to simulate 20 trials and assumed a rule change after only two correct trials in order to gain a greater amount of rule change situations. The simulation results of the above named weight sets can be summarized as follows:

- Simulations 1, 2, 3, 6 and 11: the stimulus pools did not show always distinct and satisfactory activity patterns according to the presented stimuli.
- Simulation 10: the according stimulus pools showed the desired behavior but we detected a spontaneous (i.e., without negative feedback input) rule change.
- Simulations 4, 5 and 7: delivered a spontaneous rule change and some non-satisfactory stimulus responses.
- Simulation 8: besides some non-satisfactory stimulus responses the network showed incorrect stimulus pools activities, that is, the wrong stimulus dimension showed a considerable activity level.
- Simulation 9: showed promising behavior with respect to all pools and was therefore chosen for further simulations and examination of the fMRI signal.

C.2. Evaluation of \( \lambda_{bias}, \lambda_{feedback} \) and \( \lambda_{stimulus} \)

The simulations described in Simulation results were conducted using the following parameter combination:

- \( w_{+} = 2.1 \)
- \( w_{-} = w_{mm} = 0.1 \)
- \( w_{rm} = 1.2, w_{mon} = 1.0 \)
- \( w_{rm} = 1.0, w_{mon} = 1.1 \)
- \( \lambda_{positiveFeedback} = 150 \) Hz
- \( \lambda_{negativeFeedback} = 300 \) Hz
- \( \lambda_{bias} = 60 \) Hz
- \( \lambda_{stimulus} = 150 \) Hz

This combination led to the desired behavior of the neurodynamical model in the simulations. The variation of selected parameters allows to determine boundaries for these parameters and further evaluate the model behavior.

First of all, if the network does not receive any input (\( \lambda_{bias}, \lambda_{stimulus} \) and \( \lambda_{feedback} \) set to zero) all pools show stable spontaneous activity of 3 Hz for the selective and nonselective pools and about 9 Hz for the inhibitory pool as designated by the framework of Brunel and Wang (2001). For the further evaluation of the external input \( \lambda_{bias}, \lambda_{feedback} \) and \( \lambda_{stimulus} \) we conducted again a mean field analysis first and used the seven-step configuration as described in Appendix C.1:

- \( \lambda_{stimulus} \) was varied starting with a value of 10 Hz up to 500 Hz in steps of 10. The remaining parameters were left unchanged. Due to this analysis, a minimum value of 60 Hz should already lead to the desired simulation behavior of the model. If \( \lambda_{stimulus} \) is set to a value below 60 Hz the rule pools would not show a sufficient and different level of activity and thus step (1) of the described configuration (see Appendix C.1) would not succeed. A maximum value for \( \lambda_{stimulus} \) up to the tested rate of 500 Hz was not detectable. Thus, for the factor \( \lambda_{stimulus} \) values from 60 Hz to at least 500 Hz should lead to the desired simulation behavior.
- \( \lambda_{feedback} \). The positive feedback was varied starting with a value of 0 Hz up to 200 Hz as well in steps of 10 Hz with the remaining parameters left unchanged. The analysis revealed a necessary minimum value for the positive feedback of 50 Hz. If set to values below, the model would not be able to keep the active rule after a correct response period and thus step (3) would not complete successfully. The maximum value for the positive feedback turned out to be 180 Hz. Higher values for the positive feedback would lead to undesired rule changes and thus step (3) would again not complete successfully. For the negative feedback, we analyzed the range starting at 210 Hz up to 500 Hz. The minimum value turned out to be 180 Hz while a maximum was not detectable up to a rate of 500 Hz. If the negative feedback is set to values below 250 Hz, the rule change would not succeed (step (7)).
- \( \lambda_{bias} \) was altered starting with 0 Hz up to 200 Hz as well in steps of 10 Hz. The analysis revealed a minimum value of 50 Hz while no maximum exposed. If the bias is set to values below 50 Hz the rule pools will not show a distinct and clear level of activity (step (1) fails).

\[ \text{Diagram 1:} \quad w_{+} = 2.1, w_{-} = w_{mm} = 0.1, w_{rm} = 1.2, w_{mon} = 1.0, w_{rm} = 1.0, w_{mon} = 1.1, \lambda_{stimulus} = 150 \text{ Hz} \]

\[ \text{Diagram 2:} \quad w_{+} = 2.1, w_{-} = w_{mm} = 0.1, w_{rm} = 1.2, w_{mon} = 1.0, w_{rm} = 1.0, w_{mon} = 1.1, \lambda_{stimulus} = 150 \text{ Hz} \]

\[ \text{Diagram 3:} \quad w_{+} = 2.1, w_{-} = w_{mm} = 0.1, w_{rm} = 1.2, w_{mon} = 1.0, w_{rm} = 1.0, w_{mon} = 1.1, \lambda_{stimulus} = 150 \text{ Hz} \]

\[ \text{Diagram 4:} \quad w_{+} = 2.1, w_{-} = w_{mm} = 0.1, w_{rm} = 1.2, w_{mon} = 1.0, w_{rm} = 1.0, w_{mon} = 1.1, \lambda_{stimulus} = 150 \text{ Hz} \]

\[ \text{Diagram 5:} \quad w_{+} = 2.1, w_{-} = w_{mm} = 0.1, w_{rm} = 1.2, w_{mon} = 1.0, w_{rm} = 1.0, w_{mon} = 1.1, \lambda_{stimulus} = 150 \text{ Hz} \]

\[ \text{Diagram 6:} \quad w_{+} = 2.1, w_{-} = w_{mm} = 0.1, w_{rm} = 1.2, w_{mon} = 1.0, w_{rm} = 1.0, w_{mon} = 1.1, \lambda_{stimulus} = 150 \text{ Hz} \]

\[ \text{Diagram 7:} \quad w_{+} = 2.1, w_{-} = w_{mm} = 0.1, w_{rm} = 1.2, w_{mon} = 1.0, w_{rm} = 1.0, w_{mon} = 1.1, \lambda_{stimulus} = 150 \text{ Hz} \]

\[ \text{Diagram 8:} \quad w_{+} = 2.1, w_{-} = w_{mm} = 0.1, w_{rm} = 1.2, w_{mon} = 1.0, w_{rm} = 1.0, w_{mon} = 1.1, \lambda_{stimulus} = 150 \text{ Hz} \]

\[ \text{Diagram 9:} \quad w_{+} = 2.1, w_{-} = w_{mm} = 0.1, w_{rm} = 1.2, w_{mon} = 1.0, w_{rm} = 1.0, w_{mon} = 1.1, \lambda_{stimulus} = 150 \text{ Hz} \]

\[ \text{Diagram 10:} \quad w_{+} = 2.1, w_{-} = w_{mm} = 0.1, w_{rm} = 1.2, w_{mon} = 1.0, w_{rm} = 1.0, w_{mon} = 1.1, \lambda_{stimulus} = 150 \text{ Hz} \]

\[ \text{Diagram 11:} \quad w_{+} = 2.1, w_{-} = w_{mm} = 0.1, w_{rm} = 1.2, w_{mon} = 1.0, w_{rm} = 1.0, w_{mon} = 1.1, \lambda_{stimulus} = 150 \text{ Hz} \]
The mean field analysis revealed so far a very interesting parameter range and thus a further support especially for the assumed behavioral equivalent of the provided external inputs: The positive feedback serves indeed as a ‘strengthen’[er], a reward of the active rule presumably comparable to the explicit answer ‘correct’ to a subject in an experiment. The factor \( \lambda_{\text{bias}} \) represents in a way the attention paid to the overall task. If set to an insufficient amount, the model – presumably again similar to a subject not paying attention – would not activate a rule to a sufficient level.

Simulations of the spiking dynamics using boundary parameter values showed further constraints and relationships concerning the external input to the model. \( \lambda_{\text{stimulus}} \) has to be at least 140 Hz. Otherwise, the memory pools would not react reliably appropriate to the sample stimulus but keep the activity due to the first cue stimulus input presented during the trials to the network.\(^2\)

For \( \lambda_{\text{bias}} \), a value of 20 to 30 Hz should already fit the simulation requirements as the positive feedback always strengthens the active rule which should lead to the necessary level of rule activity.\(^3\) For the negative feedback, a value of about 280 Hz turned out to be necessary in order to see a rule change reliably happen. The strengthening of the active rule due to the provision of the positive feedback (i.e., an increase in the spiking rate) is observable starting at a value of about 100 Hz. The variation of \( \lambda_{\text{bias}} \) in conjunction with the positive feedback seem to be a promising approach to integrate errors in the model behavior. If the bias is too low (0 to 20 Hz) and the rule pools are furthermore not sufficiently strengthened by the positive feedback, this leads to arbitrary rule changes (i.e., the contrary memory pools are activated without the negative feedback being provided) and thus the model would produce erroneous answers (see as well Discussion above).

However, for the calculation of the fMRI signal a robust and accurate model behavior is necessary as for these calculations only error-free trials are taken into consideration. Thus, in summary the evaluation showed so far that the chosen set of values for \( \lambda_{\text{bias}}, \lambda_{\text{feedback}} \) and \( \lambda_{\text{stimulus}} \) used in the simulations are in a good range to produce a robust model behavior.

**C.3. Evaluation of \( w_+ \)**

The parameter \( w_+ \) is on the one hand responsible for the identification of the various pools within the network; neurons within a pool are connected with each other by this strong weight \( w_+ \) which is almost twice as high compared to the weights between any two different pools within the network. On the other hand, \( w_+ \) is a factor tuning the bistability property of the model. ‘Bistability’ thereby means that if a pool receives a sufficient level of external input it enters into a state of transient activity and keeps this state even if the input is removed. If the input is not sufficient, the firing rate of the pool will return to the spontaneous level of activity after removing the input (thus bistability: stable spontaneous or stable transient activity). Considering the experiment simulations (see Fig. 6), we see that the memory pools always keep the activity according to the last presented stimulus until a new stimulus is presented. Thus, a new stimulus ‘erases’ the memorization of the last stimulus in the non-match case or strengthens it in the match-case. Thus, the transient activity of the memory pools is removed only by a competing stimulus input. The rule pools show always a clear level of transient (valid rule) or spontaneous (invalid rule) activity except for the feedback periods; in these periods the spiking rate increases (positive feedback) or the activity levels change (negative feedback). Lower values for the parameter \( w_+ \) decrease model stability, that is, the pools might not be able to keep stable transient activity whereas higher values for \( w_+ \) increase model stability and hence the model is more robust against external input variations. Concretely:

- \( w_+ = 1.9 \). The rule pools are not able to keep a sufficient amount of activity, that is, the rule pools do not reach a state of stable transient activity (average spiking rate for an active rule pool about 15 Hz compared to about 26 Hz in the original parameter erosion). Thus, the active rule changes at arbitrary times and accordingly the active memory pools. The memory pools themselves show often parallel activity (comparable to the first trial after a rule change in the original version (Fig. 6) but besides are able to keep the activity according to the presented stimulus.

- \( w_+ = 2.0 \). Here, we only see occasionally arbitrary rule changes but besides stable activity of rule and memory pools as deemed necessary for the simulations. The average spiking rate for an active rule pool turns out to be now about 20 Hz).

- \( w_+ = 2.2 \). Leads to a rule change failure rate of about 30%. The memory pools however still respond to the presented stimuli which means that \( \lambda_{\text{stimulus}} \) is still strong enough to erase previously memorized stimuli.

- \( w_+ = 2.3 \). Besides rule change failures we now see as well that the memory pools do not respond always to the presented stimuli. Thus, now \( \lambda_{\text{stimulus}} \) is too weak to erase a previously memorized stimulus.

**References**


