1

2

3

4

5　　　　A functional spiking-neuron model of activity-silent working memory in humans

6　　　　　　　based on calcium-mediated short-term synaptic plasticity

7

8

9　　　　Matthijs Pals[1], Terrence C. Stewart[2], Elkan G. Akyürek[3], & Jelmer P. Borst[1*]

10

11

12　　¹Bernoulli Institute, University of Groningen, Groningen, The Netherlands.

13　　²Centre for Theoretical Neuroscience, University of Waterloo, Waterloo, Ontario, Canada.

14　　³Department of Experimental Psychology, University of Groningen, Groningen, The Netherlands.

15

16

17

18　　*Corresponding author:

19　　E-mail: j.p.borst@rug.nl (JPB)

Abstract

In this paper, we present a functional spiking-neuron model of human working memory (WM). This model combines neural firing for encoding of information with activity-silent maintenance. While it used to be widely assumed that information in WM is maintained through persistent recurrent activity, recent studies have shown that information can be maintained without persistent firing; instead, information can be stored in activity-silent states. A candidate mechanism underlying this type of storage is short-term synaptic plasticity (STSP), by which the strength of connections between neurons rapidly changes to encode new information. To demonstrate that STSP can lead to functional behavior, we integrated STSP by means of calcium-mediated synaptic facilitation in a large-scale spiking-neuron model. The model simulated a recent study that measured behavior and EEG activity of participants in a delayed-response task. In this task, a visual grating had to be maintained in WM, and compared to a subsequent probe. It was demonstrated that WM contents could be decoded from the neural activity elicited by a task-irrelevant stimulus that was displayed during the activity-silent maintenance period. In support of our model, we show that it can perform this task, and that both its behavior as well as its neural representations correspond to the human data. We conclude that information in WM can be effectively maintained in activity-silent states by means of calcium-mediated STSP.

Author Summary

Mentally maintaining information for short periods of time in working memory is crucial for human adaptive behavior. It was recently shown that the human brain does not only store information through neural firing – as was widely believed – but also maintains information in activity-silent states. Here, we present a detailed neural model of how this could happen in our brain through short-term synaptic plasticity: rapidly adapting the connection strengths between neurons in response to incoming information. By reactivating the adapted network, the stored information can be read out later. We show that our model can perform a working memory task as accurate as human participants

46   can, while using similar mental representations. We conclude that our model is a plausible and

47   effective neural implementation of human working memory.

Introduction

49   The ability to temporarily hold information in working memory (WM) is a crucial part of
50   day-to-day life: it is what allows us to remember someone's name at a cocktail party, what ingredients
51   to buy at the supermarket for dinner, and which platform we need to go to when changing trains [1,2].
52   The maintenance of information in WM is often studied by means of a delayed-response task, in
53   which a briefly presented memory item is followed by a delay period [3,4]. The delay period ends
54   with the presentation of a probe that the participants need to compare to the memorized item. The
55   maintenance of information during the delay period of such tasks was long thought to be mediated by
56   continuously spiking neurons [5,6]. Although neural spiking is certainly important for WM, it was
57   recently shown that spiking activity during delay periods can be intermittent or even absent [7–11].
58   This suggests that information may be stored instead using activity-silent mechanisms, for instance
59   through transient connectivity patterns in the brain [2,12,13]. The spiking activity observed previously
60   might reflect the initial phase necessary to initialize new synaptic weights, active maintenance of the
61   focus of attention [14–17], or the read-out of information from working memory [13,18].

62   One of the candidate mechanisms for storing information in activity-silent states is short-term
63   synaptic plasticity [STSP; 19], which entails rapid changes in the strength of connections between
64   neurons to reflect new information being presented to the network [12]. Indeed, it was previously
65   shown that synapses in areas implicated in WM can be facilitated [20,21], potentially as a
66   consequence of residual calcium building up in presynaptic terminals [19,22]. In this way residual
67   calcium effectively leaves a 'synaptic trace' of what is currently stored in WM. An elegant
68   implementation of activity-silent storage by means of STSP was proposed by Mongillo and
69   colleagues [12], who developed a model that can maintain information through calcium-mediated
70   synaptic facilitation in recurrent networks of simulated spiking neurons. In response to a particular
71   input to the network, a subset of the neurons fires, with the result that their outgoing connections are
72   facilitated. Subsequently, stored information can be read out by applying a network-wide non-specific
73   input that will be mostly subthreshold for non-facilitated neurons but leads to firing of facilitated
74   neurons.
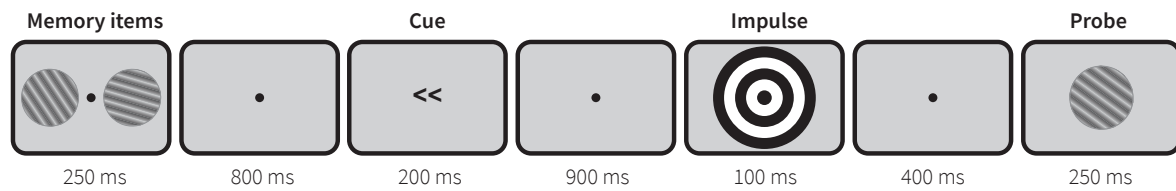
Fig 1. The retro-cue delayed-response task. After the presentation of the memory items, a cue indicates which grating needs to be maintained in WM for judgement of the probe. Decodable EEG activity is elicited by the task-neutral impulse, which is presented between the cue and the probe.

75   In the current study, we show that the mechanism proposed by Mongillo and colleagues [12]

76   not only results in efficient and robust storage, but also in effective, functional human behavior. We

77   integrated their calcium-mediated STSP mechanism in a large-scale spiking-neuron model that can

78   perform a delayed-response task. To evaluate this model, we used a previously reported

79   electroencephalography (EEG) dataset of a visual WM task, in which activity-silent memory states

80   were measured [3]. To this end, Wolff and colleagues developed an innovative method to probe

81   activity-silent brain states [3,23]. They showed that when the WM network is perturbed by a high-

82   contrast task-neutral stimulus during maintenance, ensuing neural activity reveals what is currently

83   held in an activity-silent state.

84   In their experiment [3], each trial started with the display of two randomly oriented gratings

85   (Fig 1). After an 800 ms fixation period, this was followed by a cue indicating which of the two

86   stimuli had to be maintained in memory. In order to examine the contents of WM during the

87   subsequent delay part of the trial, an impulse stimulus was presented 900 ms later. At the end of each

88   trial, participants had to indicate whether a probe stimulus was rotated clockwise or counter-clockwise

89   with respect to the cued memory item. To track the contents of WM, a decoding analysis was applied

90   to the EEG data [3]. It was shown that decoding accuracy quickly dropped to chance level after

91   presentation of the memory items, but returned when the probe was presented. This indicates that

92   between the presentation of the memory items and the probe, information is maintained in an activity-

93   silent (or at least quiescent) state. In addition, it was shown that it is possible to decode the orientation

94   of the cued memory item from the EEG data in response to the impulse stimulus. Thus, when the WM

95 network was perturbed by a task-neutral stimulus, the ensuing signal allowed for decoding of the

96 current contents of the activity-silent state. Interestingly, already after the presentation of the cue, only

97 the orientation of the cued memory item could be decoded, indicating that the uncued stimulus was

98 quickly forgotten, or actively cleared from memory.

99 To test our model, we let it perform the same experiment – including the application of the

100 impulse perturbation method – and compared both our model's performance as well as its mental

101 representations and underlying spiking behavior to the human data.

102

Results

104 **Model Architecture**

105 In order to implement a functional spiking-neuron model of WM we used Nengo, a

106 framework for building large-scale brain models that link single cell activity to demonstrative

107 cognitive abilities [24–26]. In this framework, information is represented by vectors of real numbers

108 that can be encoded and decoded from the collective spiking activity of a population of neurons.

109 Connections between neural populations allow for both communication and transformation of

110 information. Here, Nengo acts as a 'neural compiler': given a desired function, the connection weight

111 matrix between populations is calculated so that this function is approximated. Besides pre-

112 calculating connection weights, plasticity can be introduced by making use of biologically plausible

113 learning rules [27].

114 To account for short-term synaptic plasticity, we integrated the calcium kinetics mechanism

115 proposed by Mongillo and colleagues [12] in the model. Accordingly, synaptic efficiency between

116 two neurons is dependent on two parameters: the amount of available resources to the presynaptic

117 neuron (reflecting neurotransmitters) and the presynaptic calcium level. Each time a neuron fires, the

118 amount of available resources decreases, reducing synaptic efficiency. As resources are quickly

119 replenished (in the order of 200 ms), this results in short-term depression of firing rates. However, at

120 the same time calcium flows into the presynaptic terminals, *increasing* synaptic efficiency. Because

6
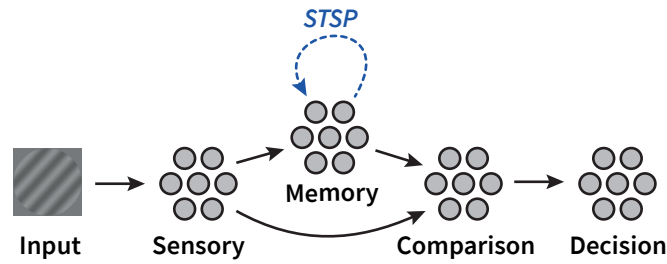
Fig 2. Model architecture. The model is divided in two modules (only one is pictured) representing the two visual hemispheres. Stimuli enter via a sensory population that transforms the input into a vector. This vector is then sent to a recurrently connected memory population exhibiting STSP. The comparison population integrates information from both the sensory and memory ensembles, the result of which is interpreted by the decision population.

121 calcium is much slower to return to its baseline levels than the resources, the synaptic connection is

122 facilitated in the long-term, for about 1.5 seconds.

123   This STSP mechanism was applied to the recurrent connections of two working memory

124 populations in our model. As described above, the aim is to simulate a dataset of a delayed-response

125 task in which the orientation of two memory items has to be compared to a probe ([3]; Fig 1). In this

126 task, significant EEG lateralization was observed at posterior electrodes after presentation of the cue.

127 We therefore hypothesized that distinct populations of neurons are responsible for processing visual

128 stimuli presented in the left and right visual field. Correspondingly, the model was divided into two

129 independent modules, each responsible for perceiving and representing one of the two incoming

130 stimuli (Fig. 2).

131   In order to demonstrate that our model is able to deal with real-world input, the stimuli from

132 [3] were presented to the model. The sensory populations use two-dimensional Gabor filters as

133 encoders [25,28,29]. As a result, the information present in the gratings – including their direction – is

134 encoded into 24-dimensional vectors that are passed on to the memory populations. That is, the

135 information encoded into the neurons is a compressed representation of the input image, using the top

136 24 singular values as per SVD (see Methods for more details). The memory populations contain

137 recurrent connections exhibiting STSP, in line with previous models of WM and anatomical areas

138   associated with WM [e.g., 21,30,31]. Consequently, the first stimulus during a trial will drive

139   facilitation of recurrent connections representing this stimulus. Neural activity resulting from

140   subsequent stimuli will be affected by this change in connectivity.

141   This enables implementing decision making as a match-filter process [8,13]. To decide on the

142   orientation change of the probe compared to the relevant memory item, both the sensory and memory

143   populations communicate the orientation of the gratings to a comparison population. When a probe is

144   presented, the orientation received from the sensory population is driven entirely by the incoming

145   stimulus. However, the orientation of the memory population is driven by a dynamic combination of

146   activity resulting from the incoming stimulus and activity from facilitated connections as a result of

147   the encoded memory item (i.e. hysteresis). In other words, the orientation represented by the memory

148   population reflects the orientation of the probe 'tuned' by the orientation of the memory item stored in

149   facilitated synapses, over time reverting to the new probe stimulus. To estimate the orientation

150   difference between the memory item and the probe, the outgoing connections from the comparison

151   layer subtract the two represented orientations. The resulting one-dimensional value reflects the

152   signed difference between the orientation of the memory item and that of the probe stimulus.

153

154   **Neural representations**

155   The model simulated the experiment reported in [3] and illustrated in Fig 1. In the original

156   paper, it was shown that decoding accuracy quickly dropped after presentation of the memory items

157   but returned again during presentation of the probe – as would be expected for an activity-silent

158   maintenance mechanism. Correspondingly, we examined the spiking activity and quality of

159   representations of our model during the task, in order to validate that any maintenance of information

160   in our model is realized in activity-silent states and not by persistent firing.

161   Fig 3 shows the spiking activity of the neurons in the memory populations of both modules

162   during one trial (A: cued module, B: uncued module), together with the mean amount of resources (x)

163   and calcium (u) in these populations. In both modules, there is spiking activity during and shortly

164   after presentation of the memory items, the impulse stimulus, and the probe, but not in between. The
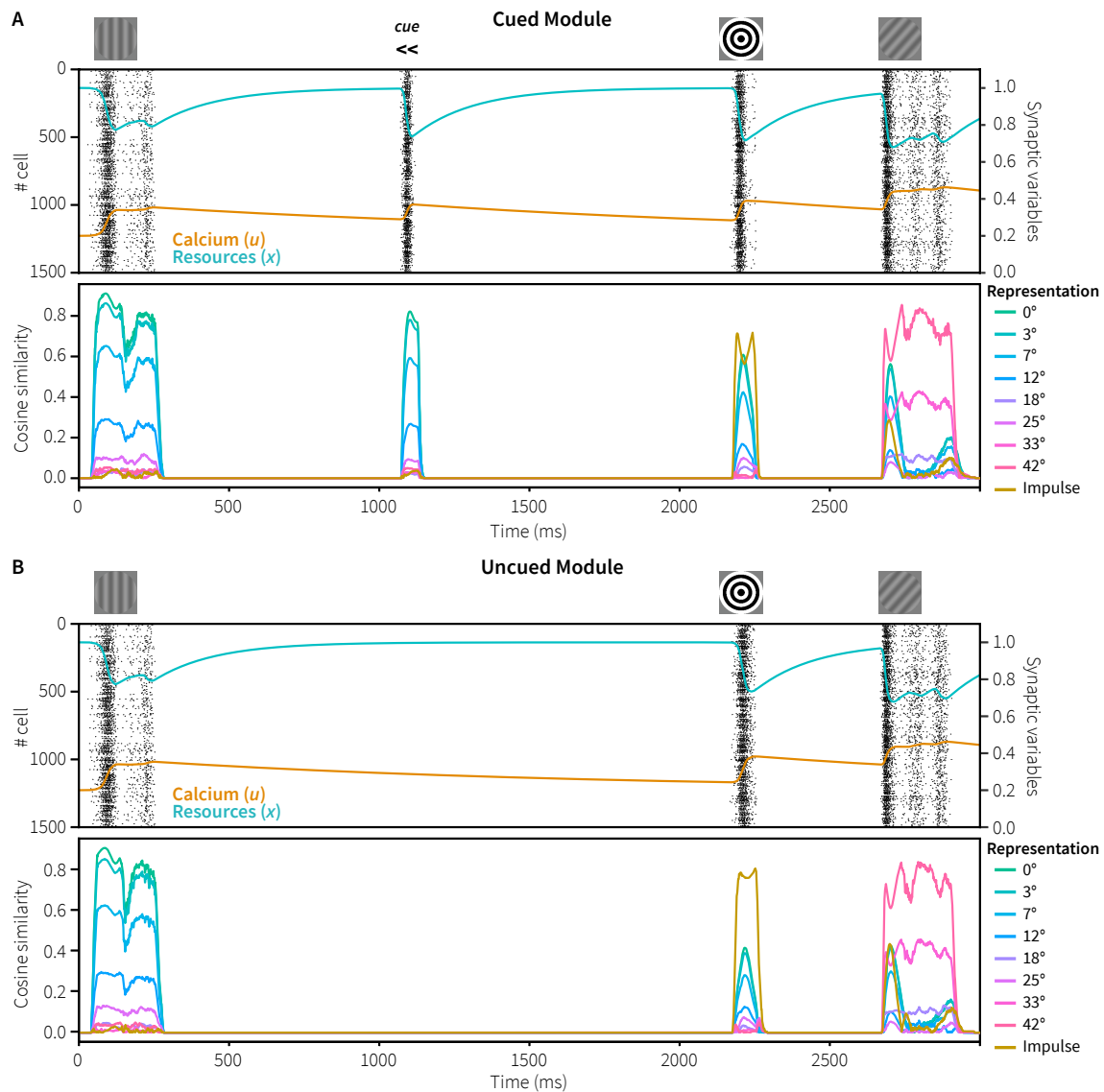
Fig 3. Spiking behavior and representations. Top: spiking activity of the memory populations of the cued (left) and uncued (right) modules, including resource and calcium levels during a trial. Bottom: cosine similarity between the vector represented by the memory populations and ideal vectors, averaged over 100 trials with 0° memory items and 42° probes with constant within-trial phase.

165   spiking activity causes the amount of available resources and the calcium level to decrease and

166   increase, respectively. The resulting short-term depression can directly be observed as the amount of

167   spiking declines after the onset of a stimulus, although it periodically reactivates.

168   In the original experiment, a retro-cue that indicated which of the two previously presented

169   items needed to be memorized was presented 800 ms later, which was followed by significant

9

170   lateralization at posterior electrodes. To mimic this, the memory population of the cued module is

171   briefly reactivated by means of a non-specific population wide input [cf. 12]. This not only re-

172   activates the memory item, but also helps to maintain the stimulus for a longer time period, as

173   reactivation of facilitated synapses will lead to re-facilitation of those connections.

174   Next, we analyzed the vectors represented by the memory populations of both the cued and

175   uncued module. Fig 3 (bottom of each panel) shows the absolute cosine similarity between the vector

176   represented by the memory populations and the ideal vectors of potential representations. To clearly

177   illustrate the difference between the two modules, the mean cosine similarity was calculated over 100

178   trials in which both modules were presented with the same memory item and probe, with a rotation of

179   0° and 42°, respectively. Note that in the simulation of the real experiment, the cued and uncued

180   modules are never presented with the same memory item.

181   During presentation of the initial memory item of 0°, the vectors represented by both modules

182   are very similar to the ideal 0° vector. In addition, the cosine similarity is inversely correlated with the

183   angular difference between the represented vector of 0° and potential representations, indicating that

184   similar stimuli are represented by similar vectors and firing patterns. As was the case in the original

185   experiment, during the delay periods we could not decode what is being represented by the neural

186   populations as there is no spiking activity – indicating activity-silent memory. However, in response

187   to the non-specific reactivation of the cued model at 1050 ms, there was spiking activity that clearly

188   represents the originally encoded vector. It therefore appears that neural connections representing the

189   memory item were indeed facilitated, and that mainly those connections and neurons get activated in

190   response to the non-specific reactivation elicited by the cue.

191   One of the main results of the original study was that the EEG activity in response to the

192   impulse stimulus contained the orientation of the cued memory item, and not of the uncued item [3].

193   This was taken to show that a stimulus is only maintained in an activity-silent state if it is still needed

194   for the task. If not, it is quickly forgotten or actively cleared from the network. To see if our model

195   has both the same storage and forgetting capabilities, we examined the vectors represented by the

196   memory populations of the cued and uncued module during presentation of the impulse (Fig 4; cf. Fig
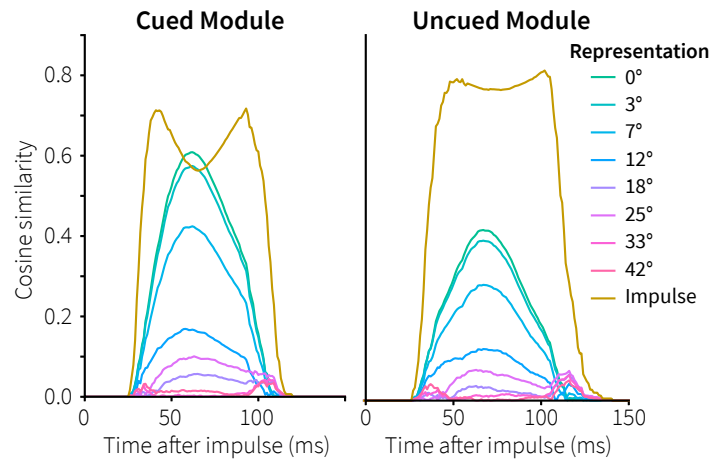
Fig 4. Cued and uncued memory representations in response to the impulse stimulus.

Cosine similarity between the representations in the memory populations and ideal vectors in response. The memory item presented before the impulse had a rotation of 0°.

197   3, 2150-2300 ms). In both modules, the memory populations start representing the impulse stimulus.

198   When the facilitated recurrent connections of the cued and uncued items become activated, both

199   modules also represent the original 0° memory item. However, only for the cued module does the

200   represented vector become (very briefly) more similar to the ideal memory item than to the impulse

201   vector, offering a potential explanation of why only the cued, and not the uncued memory item, could

202   be decoded after the impulse [3].

203   To summarize: in both the cued and the uncued modules, STSP encodes the initial stimulus.

204   In the cued model, facilitated connections are re-facilitated at the moment of cue, counteracting the

205   gradual calcium decay that goes on in both modules (Fig 3). As a result, once the impulse arrives,

206   only the cued model has sufficiently facilitated connections specific to the memory item to generate a

207   response larger than the impulse representation (Fig 4). Note that the uncued memory population was

208   not actively cleared, but that the calcium levels of the facilitated synapses simply decayed away as it

209   was not reactivated at the moment of the cue.

210
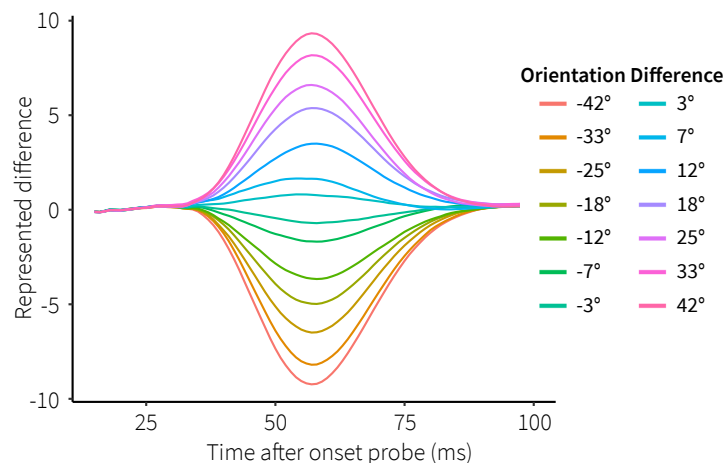
Fig 5. Represented difference in the decision population. Colors indicate experimental orientation differences, averaged over all trials.

Behavior

212   In order to see if our model not only matches neural activity, but also gives rise to functional

213   behavior similar to human participants, we evaluated its performance. First, to see if the information

214   maintained in the facilitated synapses can be used to produce a relevant response, we inspected the

215   value represented by the decision population in the cued module. This population receives the angular

216   difference between the memory item and the probe from the comparison population, and thus

217   represents a measure of difference between the orientations decoded from the sensory and memory

218   populations. Fig 5 shows the represented value for the possible orientation differences between

219   memory items and probes, averaged over all simulated trials. First, it takes a moment for the probe

220   information to reach this population. Second, the facilitated synapses become activated, reactivating

221   the memory item, thereby leading to different representations in the sensory and memory populations,

222   and thus to a difference in the decision population. Finally, the probe starts overriding the memory

223   representation, reducing the difference until both populations represent the probe and the difference

224   has disappeared. Overall, both the sign and magnitude of the orientation difference are clearly

225   represented in the decision signal.

226   To translate this decision signal into a response, we integrated the decision activation after the

227   presentation of the probe. Integrating the evidence corresponding to two distinct decisions has been

12
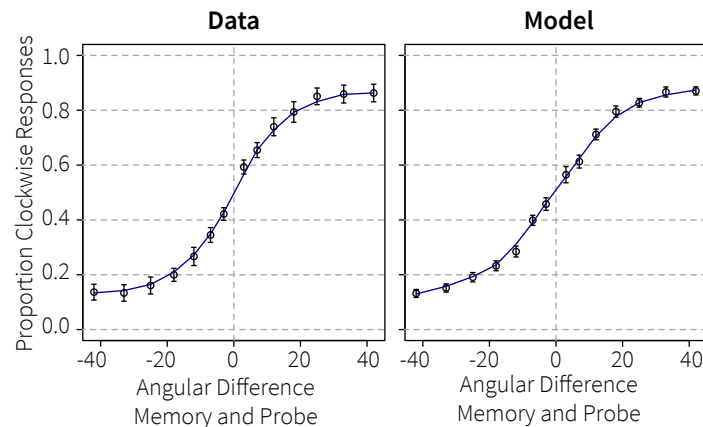
Fig 6. Performance. Proportion of clockwise responses dependent on the angular difference between the cued memory item and the probe.

228   widely used before in accumulator models of perceptual decision making [e.g., 32]. We did not model

229   motor processes, but simply interpreted a positive result as a clockwise response and a negative result

230   as a counter-clockwise response. Fig 6 shows that the model's proportion of clockwise responses

231   across orientation differences follows a similar S-shape as the human responses.

232

Discussion

234   We developed a functional spiking-neuron model to explain recent theories of activity-silent

235   human working memory. Whereas incoming information is encoded through spiking, maintenance of

236   information was realized by short-term synaptic plasticity based on calcium kinetics [12]. This

237   mechanism can maintain information effectively for short periods of time without requiring neural

238   spiking. In support of the model, we simulated a recent EEG study that applied an innovative impulse

239   perturbation method [3] to reveal the content of activity-silent WM. Both the model's choice

240   behavior, as well as its mental representations corresponded well to the human data. This

241   demonstrates that calcium-mediated STSP not just results in robust maintenance of arbitrary stimuli,

242   as shown earlier [12], but can also simulate effective human behavior based on real-world stimuli.

243   With regard to localization, the model was used to simulate data from Wolff and colleagues

244   [3], who reported posterior EEG effects. However, WM is often attributed to prefrontal areas [e.g.,

245   14,46]. Activity-silent maintenance has likewise been found in both posterior [3,8,47] and frontal

13

246   [18,44] regions. It appears that especially sensory working memory should be attributed to the

247   relevant sensory systems [47,48], instead of to a centralized system. While the exact function of the

248   different regions implicated in WM might differ, the neural substrate and mechanisms might be

249   similar, and could potentially all be explained by the proposed STSP mechanism.

250   A number of design choices warrant discussion. First, the employed neurons do not have a

251   baseline firing rate, as is evident by the lack of any spiking activity during the delay-period of a trial

252   (Fig 3). In order to clearly demonstrate activity-silent maintenance of information, we defined the

253   tuning curves of the sensory and memory neurons so that they only fire when presented with input.

254   However, background firing could simply be added to the model without affecting the functioning of

255   our model, as has been done in the past ([e.g., 12,33]; see also below). Second, the number of neurons

256   per population and the number of dimensions used to represent the stimuli were set to reflect human

257   behavior. In general, adding more neurons will improve the representation of vectors and the

258   approximation of the functions computed over those. Increasing the number of dimensions expands

259   what can represented [25,34,35]. Thus, changing the number of neurons and dimensions will change

260   the quality of the representations and will influence the number of errors made during the task. Here,

261   we estimated parameters to roughly match human performance; we do not have a principled reason

262   either for using 1000 or 1500 neurons per population or 24 dimensions. Finally, information was

263   represented using Nengo's default vector representation, which provides an intuitive method to link

264   neural spiking to representation and function [25]. However, representing information differently

265   should not affect the basic functioning of the model as all connections and the STSP mechanism are

266   implemented at the neural level.

267

Representations in WM

269   As discussed above, in the current model information is maintained without *any* intermittent

270   firing (Fig 3). This directly contradicts the original analysis of the dataset [3], where the represented

271   stimulus could be decoded for some time after its offset. In addition to full activity-silent

272   maintenance, Mongillo and colleagues [12] observed a bi-stable regime in their model: with added
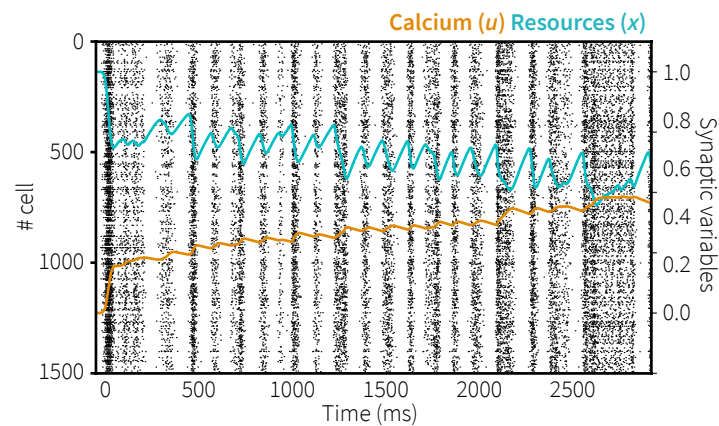
Fig 7. Bi-stable maintenance. Background noise puts the model in a bi-stable regime where facilitated connections reactivate spontaneously on the time scale of τ_D.

273   background noise, neurons with facilitated connections reactivated spontaneously. Consequently, due

274   to the dynamics of u and x, the reactivated neurons will be briefly depressed before being facilitated

275   again, leading again to reactivation. In this regime, the time between subsequent reactivations is on

276   the scale of τ_D – the time constant of the available resources – as it is controlled by the recovery from

277   the synaptic depression. A brief exploratory analysis shows that such a bi-stable regime can also be

278   added to our model, as is illustrated in Fig 7. This provides the model with an additional method of

279   maintaining information, possibly over a longer period of time. It also indicates a potential

280   explanation for the delay-activity observed in the original analysis [3]: non-specific background or

281   recurrent input after presentation of a stimulus might temporarily have pushed the network into this

282   regime.

283   A functional role for delay activity in WM might be tracking the focus of attention [15–

284   17,36]. Wolff and colleagues [3] performed a second experiment, similar to the one discussed in this

285   paper. Again, two memory items were presented at the start of the trial. However, in this version both

286   were eventually probed, and participants were aware of which of the two items would be probed first.

287   Before the first probe, the prioritized item could be much better and longer decoded than the un-

288   prioritized item, a pattern that reversed after the first probe. This could be simulated with our model

289   by giving the prioritized item a persistent firing state as in Figure 7, while the un-prioritized item

290   might be completely maintained in an activity-silent state as in Figure 3. This also provides a

291   functional explanation for psychological theories that state that a single focal WM item can be used

292   without any time cost [16,37–39], while other items in working memory incur a cost estimated at 200

293   ms [15,40–43] – the latter potentially being due to the costs of reactivating the information from a

294   non-active state.

295   In our model, we have assumed that the coding of the information itself is static, that is, the

296   same facilitated connections are used repeatedly. However, there has recently been increasing

297   evidence for a dynamic coding framework, which states that information maintained in a WM

298   network traverses a highly dynamic path through neural activation [44]. It is not yet clear how this

299   relates to our model, although a possible clue might come from a model by Singh and Eliasmith [45].

300   Neural populations in their model represent two dimensions, where one dimension represents time

301   and the other a stimulus. Their model elegantly captures a wide variety of observed neural responses

302   during a WM task; the inclusion of time as a dimension represented by the neurons in the network

303   naturally leads to a dynamic firing pattern over time.

304

Related Models

306   Recently, Myers and colleagues [8] described a related non-spiking neural population model

307   with similar functionality as the current model, although they did not match human data directly.

308   Their model consisted of a three-layer architecture: a stimulus layer, a template layer, and a decision

309   layer not unlike the sensory, memory, and comparison population in our model. A critical difference

310   between the two models is that their decision layer only receives input from the template layer, while

311   in our case it receives input from both the sensory and memory populations. The template layer in

312   Myers' model acts like a match-filter: it is able to maintain a stimulus orientation, and when presented

313   with a subsequent probe orientation convey the signed difference between the two to the decision

314   layer. The memory population in our model can likewise be viewed as a match filter. After onset of

315   the probe, the represented orientation shifts to the orientation of the probe from the direction of the

316   orientation of the memory item. This shift in itself indicates a degree of difference between the two

317 orientations, including the sign of this difference. One could potentially measure this with a neural

318 population that computes a time derivative with respect to the orientation [49]. However, exploratory

319 analysis indicated this to be less robust than our current method.

320   Another closely related model was proposed by Barak and colleagues [18]. Their model

321 consisted of a sensory and memory population. After presentation of a stimulus, connections from the

322 sensory population to the memory population will be facilitated. Subsequently, during the delay

323 period, an increasing uniform current is applied to the network which activates the neurons in the

324 memory population that have facilitated incoming connections. During presentation of a subsequent

325 probe, mutual inhibition between the sensory and memory population will guide decision making.

326 This model explains observed ramping up of activity during anticipation of a probe. However, it is not

327 clear whether the gradually increasing external current is essential to extract the information

328 maintained in the facilitated connections in the memory representation. It can be expected that in the

329 brain bottom-up stimulus driven activity might also be able to activate the information stored in

330 connections, for instance when the timing of the probe is unknown.

331

Conclusion

333   To conclude, our model shows that maintenance of information in WM by means of calcium-

334 mediated STSP can lead to functional behavior. It is broadly consistent with current theories

335 regarding activity-silent storage in human WM and is able to show a variety of effects observed

336 during a visual delayed-response task. Furthermore, it provides a solid basis for exploring a model

337 that incorporates psychological theories on the focus of attention [15–17,37] by combining activity-

338 silent maintenance with storage through persistent firing.

339

Methods

Model

342   **Nengo.** The model was implemented using Nengo, a Python library for simulating large-scale

343 neural models with a clear link between spiking activity and representation [24–26]. Nengo makes use

17

344   of a theoretical framework called the Neural Engineering Framework [NEF; 50]. Information is

345   represented as a vector of real numbers that can be encoded and decoded from the collective spiking

346   activity of populations of neurons. Encoding is mediated by giving each neuron a non-linear tuning

347   curve that characterizes their general response to the incoming signal. Decoding is a linear process:

348   the activity of each neuron in a population is weighted by a constant and summed over time in order

349   to decode the represented vector. Connections between populations allow for the communication and

350   transformation of the information. Here the NEF calculates the connection weight matrix between

351   populations to approximate a desired function. In addition, connection weights can be learned and

352   adapted through several biologically plausible learning rules, both supervised and unsupervised [27].

353   **Short-term synaptic plasticity.** Short-term synaptic plasticity was implemented in Nengo

354   following the calcium kinetics mechanism of Mongillo and colleagues ([12]; available at

355   https://github.com/Matthijspals/STSP). Because spiking leaky integrate-and-fire (LIF) neurons are

356   computationally efficient while retaining a degree of biological plausibility, we added this mechanism

357   to the existing Nengo implementation of LIF neurons. Synaptic efficiency is based on two parameters:

358   the amount of available resources to the presynaptic neuron (x, normalised between 0 and 1) and the

359   fraction of resources used each time a neuron fires (u), reflecting the residual presynaptic calcium

360   level. When a neuron fires, its resources x are decreased by ux, mimicking neurotransmitter depletion.

361   At the same time, its calcium level u is increased, mimicking calcium influx into the presynaptic

362   terminal. Both u and x relax back to baseline with time constants τ_D (0.2s) and τ_F (1.5s), respectively.

363   This results in a system where after a neuron fires its outgoing connections will be depressed on the

364   time scale of τ_D and facilitated on the timescale of τ_F as illustrated in Fig 8.

365   For all LIF neurons to which we apply STSP, every time step u and x are calculated according

366   to equation 1.1 and 1.2, respectively:

367
$$\frac{dx}{dt} = \frac{1-x}{\tau_D} - ux\delta(t - t_{sp})$$
(1.1)

368
$$\frac{du}{dt} = \frac{U-u}{\tau_F} + U(1-u)\delta(t - t_{sp})$$
(1.2)

18
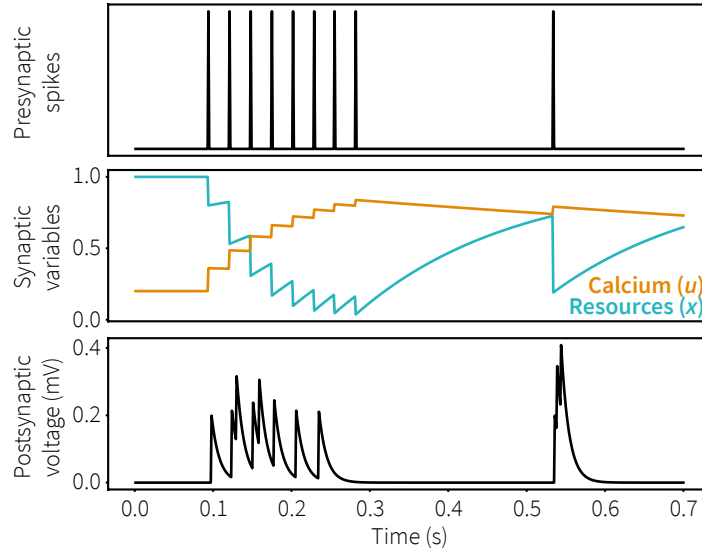
Fig 8. STSP mechanism. Top: spikes in presynaptic neuron. Middle: calcium (u) and resources (x) of presynaptic neuron, u increases and x decreases when the presynaptic neuron spikes. Bottom: resulting postsynaptic voltage; note the synaptic depression at the end of the first spike train and synaptic facilitation at the later spike. Nengo reproduction of Fig 1A in [12].

369   where x represents the available resources, u the residual calcium level, τ_D the depressing time

370   constant, δ the Dirac delta function, t the simulation time and t_{sp} the time of a presynaptic spike. In

371   equation 1.2, τ_F represents the facilitating time constant and U the calcium baseline level. Outgoing

372   connection weights of neurons implementing STSP are determined by both their initial connection

373   weight and their current synaptic efficiency. Initial connections weights are calculated by the NEF,

374   while synaptic efficiency is set to the product of the current value of u and x of the presynaptic

375   neuron, normalised by their baseline value:

376
$$w_{ij} = \frac{xu}{U} w_{ij}^0 \qquad (1.3)$$

377   where w_{ij} represents the connection weight between neuron i and j and w_{ij}^0 the initial connection

378   weight.

379   **Architecture.** The overall architecture of the model is shown in Fig 2 (the model is available

380   for download at https://github.com/Matthijspals/STSP). The sensory and decision populations consist

381   of 1000 LIF neurons, the memory and comparison populations of 1500 LIF neurons. Biologically

19

382   relevant parameters were left to default, which are consistent with neocortical pyramidal cells [24].

383   Parameters U, τ_D and τ_F were set the same as in [12]. $\tau_F \gg \tau_D$ and τ_F on the order of 1s are consistent

384   with patch-clamp recordings of facilitated excitatory connections in the ferret prefrontal cortex [21].

385   To describe the relationship between neural representations and real-world stimuli it can be

386   assumed that the brain makes use of a statistical model, not unlikely a parametrized model, where a

387   small number of parameters capture the overall shape of the data [25]. To find such a model we need

388   a set of basis functions that will be good at describing both the incoming images and the encoders of

389   the neurons receiving these images. These basis functions can be found by applying singular value

390   decomposition (SVD) to a matrix containing both the images and the encoders. The images consisted

391   of the stimuli in the experiment, while the encoders were two-dimensional Gabor filters, defined by a

392   sinusoidal plane wave multiplied by a Gaussian function. Gabor filters have previously been shown to

393   accurately describe the response profile of simple cells in the cat [29] and macaque [28] striate cortex

394   and seem to underlie early stages of visual processing. Thus, the SVD mediates a biologically

395   plausible method that results in stimuli being represented by 24-dimensional vectors.

396   Next, we specified for each possible vector – representing a particular grating – what the

397   corresponding decoded orientation should be. This information was used to define a function that

398   takes a 24-dimensional vector as input and returns the corresponding orientation. The NEF yields the

399   connection matrix at the neural level that approximates this function for the connections from the

400   sensory and memory populations to the comparison populations. Stimulus orientations were not

401   directly decoded as the angle θ, but rather by the sine and cosine of θ. Decoding sine and cosine of θ

402   is robust, as the ratio between the two determines the stimulus orientation independent of the

403   amplitude, which is not the case when decoding θ directly. Furthermore, the symmetry of the sine

404   functions provides a natural solution for the symmetric nature of the stimuli, as in the experiment a

405   stimulus with an orientation of -90° contains exactly the same pixels as a stimulus with an orientation

406   of 90° and therefore results in the same neural activity.

407

408 **Experimental Simulation**

409   **Stimuli.** Input to the model consisted of images of 128 by 128 pixels. Stimuli were generated

410 using Psychopy, an open-source Python application [51]. Stimuli consisted of a circle on a grey

411 background (RGB = 128, 128, 128). Memory items and probe stimuli were sine-wave gratings with a

412 diameter of 128 pixels and spatial frequency of 0.034 cycles per pixel. The phase was randomized

413 within and across trial. For each trial, the orientation of the memory items was randomly selected

414 from a uniform distribution of orientations. The angular differences between the memory item and the

415 corresponding probe stimulus were uniformly distributed across seven angle differences (3°, 7°, 12°,

416 18°, 25°, 33°, 42°), both clockwise and counter-clockwise. The impulse stimulus consisted of a

417 'bull's-eye' stimulus of the same size and spatial frequency as the memory items. It was presented at

418 twice the contrast compared to the grating stimuli, to each module.

419   **Procedure.** The model completed the retro-cue delayed-response task from [3]. Each trial

420 started with the presentation of two memory items to the sensory population of the corresponding

421 modules for 250 ms. In the original experiment, a retro-cue that indicated which of the two previously

422 presented items needed to be memorized was presented 800 ms later, which was followed by

423 significant lateralization at posterior electrodes. To mimic this, the memory population of the cued

424 module is briefly reactivated by means of a non-specific population wide input [cf. 12]. After another

425 fixation period, the impulse stimulus was presented to both sensory populations for 100 ms, 1100 ms

426 after the onset of the cue. After another delay of 400 ms, the probe was presented to the sensory

427 populations for 250 ms. To simulate different participants in the experiment, every 1344 trials the

428 random seed was reset and new random Gabor filters were generated to use as encoders for the

429 sensory populations. In total the model performed 30 sets of 1344 trials, reflecting 30 participants in

430 the original experiment.

431

432 **References**

433 1.   Baddeley AD, Hitch GJ. Working Memory. In: Bower GA, editor. The psychology of learning

434   and motivation. New York: Academic Press; 1974. pp. 47–89. doi:10.4324/9781912282418

435   2.   Barak O, Tsodyks M. Working models of working memory. Curr Opin Neurobiol. 2014;25:

436   20–24. doi:10.1016/j.conb.2013.10.008

437   3.   Wolff MJ, Jochim J, Akyürek EG, Stokes MG. Dynamic hidden states underlying working-

438   memory-guided behavior. Nat Neurosci. 2017;20: 864–871. doi:10.1038/nn.4546

439   4.   Rombouts JO, Bohte SM, Roelfsema PR. How attention can create synaptic tags for the

440   learning of working memories in sequential tasks. PLoS Comput Biol. 2015;11: e1004060.

441   doi:10.1371/journal.pcbi.1004060

442   5.   Fuster JMJ, Alexander GE. Neuron Activity Related to Short-Term Memory. Science (80-).

443   1971;173: 652–654.

444   6.   Goldman-Rakic PS. Cellular basis of working memory. Neuron. 1995;14: 477–485.

445   doi:10.1016/0896-6273(95)90304-6

446   7.   Lundqvist M, Herman P, Miller EK. Working Memory: Delay Activity, Yes! Persistent

447   Activity? Maybe Not. J Neurosci. 2018;38: 7013–7019. doi:10.1523/JNEUROSCI.2485-

448   17.2018

449   8.   Myers NE, Rohenkohl G, Wyart V, Woolrich MW, Nobre AC, Stokes MG, et al. Testing

450   sensory evidence against mnemonic templates. Elife. 2015;4: 1–25. doi:10.7554/eLife.09000

451   9.   Sreenivasan KK, Curtis CE, D'Esposito M. Revisiting the role of persistent neural activity

452   during working memory. Trends Cogn Sci. 2014;18: 82–89. doi:10.1016/j.tics.2013.12.001

453   10. Watanabe K, Funahashi S. Neural mechanisms of dual-task interference and cognitive capacity

454   limitation in the prefrontal cortex. Nat Neurosci. 2014;17: 601–611. doi:10.1038/nn.3667

455   11. Lundqvist M, Rose J, Herman P, Brincat SL, Buschman TJ, Miller EK. Gamma and Beta

456   Bursts Underlie Working Memory. Neuron. 2016;90: 152–164.

457   doi:10.1016/j.neuron.2016.02.028

458   12. Mongillo G, Barak O, Tsodyks M. Synaptic Theory of Working Memory. Science (80-).

459   2008;319: 1543–1546. doi:10.1126/science.1150769

460   13. Stokes MG. 'Activity-silent' working memory in prefrontal cortex: a dynamic coding

461   framework. Trends Cogn Sci. 2015;19: 394–405. doi:10.1016/j.tics.2015.05.004

462   14.   Borst JP, Anderson JR. Using Model-Based functional MRI to locate Working Memory

463   Updates and Declarative Memory Retrievals in the Fronto-Parietal Network. Proc Natl Acad

464   Sci USA. 2013;110: 1628–1633. doi:10.1073/pnas.1221572110

465   15.   Borst JP, Taatgen NA, Van Rijn H. The Problem State: A Cognitive Bottleneck in

466   Multitasking. J Exp Psychol Learn Mem Cogn. 2010;36: 363–382. doi:10.1037/a0018106

467   16.   Oberauer K. Access to information in working memory: Exploring the focus of attention. J

468   Exp Psychol Learn Mem Cogn. 2002;28: 411–421.

469   17.   Olivers CNL, Peters J, Houtkamp R, Roelfsema PR. Different states in visual working

470   memory: when it guides attention and when it does not. Trends Cogn Sci. 2011;15: 327–334.

471   doi:10.1016/j.tics.2011.05.004

472   18.   Barak O, Tsodyks M, Romo R. Neuronal population coding of parametric working memory. J

473   Neurosci. 2010;30: 9424–9430. doi:10.1523/JNEUROSCI.1875-10.2010

474   19.   Zucker RS, Regehr WG. Short-Term Synaptic Plasticity. Annu Rev Physiol. 2002;64: 355–

475   405. doi:10.1146/annurev.physiol.64.092501.114547

476   20.   Tsodyks M, Pawelzik K, Markram H. Neural Networks with Dynamic Synapses. Neural

477   Comput. 1998; 821–835.

478   21.   Wang Y, Markram H, Goodman PH, Berger TK, Ma J, Goldman-Rakic PS. Heterogeneity in

479   the pyramidal network of the medial prefrontal cortex. Nat Neurosci. 2006;9: 534–542.

480   doi:10.1038/nn1670

481   22.   Jackman SL, Regehr WG. The Mechanisms and Functions of Synaptic Facilitation. Neuron.

482   2017;94: 447–464. doi:10.1016/j.neuron.2017.02.047

483   23.   Wolff MJ, Ding J, Myers NE, Stokes MG. Revealing hidden states in visual working memory

484   using electroencephalography. Front Syst Neurosci. 2015;9: 1–12.

485   doi:10.3389/fnsys.2015.00123

486   24.   Bekolay T, Bergstra J, Hunsberger E, DeWolf T, Stewart TC, Rasmussen D, et al. Nengo: a

487   Python tool for building large-scale functional brain models. Front Neuroinform. 2014;7: 48.

488   doi:10.3389/fninf.2013.00048

489   25.   Eliasmith C. How to Build a Brain: A Neural Architecture for Biological Cognition. New

490   York, NY: Oxford University Press; 2013.

491   26.   Eliasmith C, Stewart TC, Choo X, Bekolay T, DeWolf T, Tang C, et al. A large-scale model of

492   the functioning brain. Science (80-). 2012;338: 1202–1205. doi:10.1126/science.1225266

493   27.   Bekolay T, Kolbeck C, Eliasmith C. Simultaneous unsupervised and supervised learning of

494   cognitive functions in biologically plausible spiking neural networks. Cognitive Science

495   Society; 2013. pp. 169–174.

496   28.   Ringach DL. Spatial Structure and Symmetry of Simple-Cell Receptive Fields in Macaque

497   Primary Visual Cortex. J Neurophysiol. 2002;88: 455–463. doi:10.1152/jn.2002.88.1.455

498   29.   Jones JP, Palmer LA. An Evaluation of the Two-Dimensional Gabor Filter Model of Simple

499   Receptive Fields in Cat Striate Cortex. J Neurophysiol. 1987;58: 1233–1258.

500   30.   Camperi M, Wang X-J. A model of visuospatial short-term memory in prefrontal cortex:

501   recurrent network and cellular bistability. J Comput Neurosci. 1998;5: 383–405.

502   31.   Miller P, Brody CD, Romo R, Wang XJ. A Recurrent Network Model of Somatosensory

503   Parametric Working Memory in the Prefrontal Cortex. Cereb Cortex. 2003;13: 1208–1218.

504   doi:10.1093/cercor/bhg101

505   32.   Palmeri TJ, Schall JD, Logan GD. Neurocognitive Modeling of Perceptual Decision Making.

506   In: Busemeyer JR, Wang Z, Townsend JT, Eidels A, editors. The Oxford Handbook of

507   Computational and Mathematical Psychology. New York, NY: Oxford University Press; 2015.

508   pp. 320–340.

509   33.   Stewart TC, Bekolay T, Eliasmith C. Learning to Select Actions with Spiking Neurons in the

510   Basal Ganglia. Front Neurosci. 2012;6. doi:10.3389/fnins.2012.00002

511   34.   Rasmussen D, Eliasmith C. A spiking neural model applied to the study of human

512   performance and cognitive decline on Raven's Advanced Progressive Matrices. Intelligence.

513   2014;42: 53–82. doi:10.1016/j.intell.2013.10.003

514   35.   Kajić I, Gosmann J, Stewart TC, Wennekers T, Eliasmith C. A Spiking Neuron Model of

515   Word Associations for the Remote Associates Test. Front Psychol. 2017;8: 48.

516   36.   Lewis-Peacock JA, Drysdale AT, Oberauer K, Postle BR. Neural evidence for a distinction

517   between short-term memory and the focus of attention. J Cogn Neurosci. 2012;24: 61–79.

518   doi:10.1162/jocn_a_00140

519   37.   Oberauer K. Design for a working memory. In: Ross BH, editor. Psychology of Learning and

520   Motivation. Academic Press; 2009. pp. 45–100.

521   38.   Cowan N. Attention and memory: An integrated framework. New York: Oxford University

522   Press; 1995.

523   39.   McElree B. Working memory and focal attention. J Exp Psychol Learn Mem Cogn. 2001;27:

524   817–835.

525   40.   Anderson JR. How Can the Human Mind Occur in the Physical Universe? New York: Oxford

526   University Press; 2007.

527   41.   Nijboer M, Borst JP, Van Rijn H, Taatgen NA. Contrasting Single and Multi-Component

528   Working-Memory Systems in Dual Tasking. Cogn Psychol. 2016;86: 1–26.

529   doi:10.1016/j.cogpsych.2016.01.003

530   42.   Borst JP, Taatgen NA, Stocco A, Van Rijn H. The Neural Correlates of Problem States:

531   Testing fMRI Predictions of a Computational Model of Multitasking. PLoS One. 2010;5:

532   e12966. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0012966

533   43.   Borst JP, Taatgen NA, Van Rijn H. What Makes Interruptions Disruptive? A Process-Model

534   Account of the Effects of the Problem State Bottleneck on Task Interruption and Resumption.

535   Proc CHI. Seoul, Korea: ACM Press; 2015.

536   44.   Stokes MG, Kusunoki M, Sigala N, Nili H, Gaffan D, Duncan J. Dynamic Coding for

537   Cognitive Control in Prefrontal Cortex. Neuron. 2013;78: 364–375.

538   doi:10.1016/j.neuron.2013.01.039

539   45.   Singh R, Eliasmith C. Higher-dimensional neurons explain the tuning and dynamics of

540   working memory cells. J Neurosci. 2006;26: 3667–3678. doi:10.1523/JNEUROSCI.4864-

541   05.2006

542   46.   Fuster J. The prefrontal cortex. 5th ed. New York, NY: Elsevier; 2015.

543   47.   Sreenivasan KK, Vytlacil J, D'Esposito M. Distributed and dynamic storage of working

544   memory stimulus information in extrastriate cortex. J Cogn Neurosci. 2014;26: 1141–1153.

545   doi:10.1162/jocn_a_00556

546   48.   Pasternak T, Greenlee MW. Working memory in primate sensory systems. Nat Rev Neurosci.

547   2005;6: 97–107. doi:10.1038/nrn1603

548   49.   Tripp BP, Eliasmith C. Population models of temporal differentiation. Neural Comput.

549   2010;22: 621–659. doi:10.1162/neco.2009.02-09-970

550   50.   Eliasmith C, Anderson CH. Neural Engineering: Computation, Representation, and Dynamics

551   in Neurobiological Systems. Cambridge, MA: The MIT Press; 2002.

552   51.   Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, et al. PsychoPy2:

553   Experiments in behavior made easy. Behav Res Methods. 2019;51: 195–203.

554   doi:10.3758/s13428-018-01193-y

555