A neural model of working memory

Authors: Sanjay G Manohar[1]*, Nahid Zokaei[2,4], Sean J Fallon[2], Tim Vogels[3], Masud Husain[1,2]

Affiliations:

[1]Nuffield Department of Clinical Neurosciences, University of Oxford, OX3 9DU.

[2]Department of Experimental Psychology, University of Oxford.

[3]Centre for Neural Circuits and Behaviour, University of Oxford

[4]Oxford Centre for Human Brain Activity, University of Oxford

*Correspondence to: Sanjay G Manohar, sanjay.manohar@ndcn.ox.ac.uk.

Summary

Working memory, the ability to keep recently encountered information available for immediate processing, has been proposed to rely on two mechanisms that appear difficult to reconcile: self-sustained neural firing, or the opposite—activity-silent synaptic traces. Here we show that both phenomena can co-exist within a unified system in which neurons hold information in both activity and synapses. Rapid plasticity in flexibly-coding neurons allows features to be bound together into objects, with an important emergent property being the focus of attention. One memory item is held by persistent activity in an attended or "focused" state, and is thus remembered better than other items. Other, previously attended items can remain in memory but in the background, encoded in activity-silent synaptic traces. This dual functional architecture provides a unified common mechanism accounting for a diverse range of perplexing attention and memory effects that have been hitherto difficult to explain in a single theoretical framework.

Acknowledgments: Research was funded by an MRC Clinician Scientist Fellowship to SGM, a Wellcome Trust Principal Fellowship to MH and NIHR Oxford Biomedical Research Centre. SGM conceived the model, ran simulations, and wrote the manuscript. TV advised on modelling, NZ, SJF and MH co-wrote the manuscript. We thank Prof. Mark Stokes and Dr. Nicholas Myers for comments on a draft, and Dr Eva Feredoes whose discussions inspired some of the concepts.

Introduction

Our capacity to hold and manipulate information over delays of a few seconds has long been thought to be subserved by the persistent firing of neurons during the delay (Funahashi, 2017; Fuster and Alexander, 1971). However, a number of recent studies have instead proposed “activity-silent” working memory, in which synaptic weights hold information during the delay, even in the absence of neuronal firing (Silvanto, 2017; Sreenivasan et al., 2014; Mongillo et al., 2008; Stokes, 2015). This dispute comes at a time when it is also becoming clear that working memory (WM) is not a homogeneous store. When we hold multiple items in WM, strong attentional effects are apparent. For example, people are faster and more accurate to recall the last item encoded, or the last item that was brought to mind (Chun et al., 2011; Oberauer, 2002; Souza and Oberauer, 2016; Zokaei et al., 2014a).

One item in memory, termed the ‘focus of attention’, appears to be in a privileged state. It is decodable using functional MRI and is susceptible to TMS, unlike the unfocused items which considered to be stored but in a non-privileged state (Lewis-Peacock et al., 2012; Sprague et al., 2016). In contrast, unfocused items can only be decoded after their latent representation is re-activated (Rose et al., 2016; Wolff et al., 2017). Thus both active and inactive representations may coexist in WM, and items can move between these two states (LaRocque et al., 2014; Postle, 2016; Zokaei et al., 2014b). Computational neural models of both active (Compte et al., 2000; Zenke et al., 2015) and silent (Mi et al., 2017; Mongillo et al., 2008) WM have been separately postulated, but neither type of model on their own accounted for shifts of attention within WM. Here we unify these divergent approaches in a new class of memory model.

Rapid synaptic plasticity at the millisecond scale has been used to explain how a pattern of inputs can be remembered (Fiebig and Lansner, 2017; Sandberg et al., 2003). In these synaptic models, simultaneously-activated neurons become more strongly connected, so that when a partial pattern is later presented, the original combination of active neurons can be re-activated, by associative recall. At longer time scales, synaptic plasticity may underlie associative episodic memory (Burgess and Hitch, 2005; Rizzuto and Kahana, 2001), or long term memory, which can provide the synaptic backdrop to support an active WM (Litwin-Kumar and Doiron, 2014, 2014; Zenke et al., 2015). Rapid plasticity in auto-associative networks can also account for serial recall of sequences of items (Fiebig and Lansner, 2017; Howard and Kahana, 2002) – including serial order effects such as primacy and recency (Farrell and Lewandowsky, 2002) – because new information may use up free space, or overwrite old information (Matthey et al., 2015; Sandberg et al., 2003). Because the physiological meaning of a neuron’s firing depends upon its input and output connections, plasticity in these models could lead to neurons whose activity represents different things on different trials – a property that we characterize here as *flexible coding*. However these models do not produce stable persistent-activity states in feature-selective neurons, which has long been considered a hallmark of WM (Funahashi, 2017).

In contrast, in sustained activity models, items are held WM by virtue of delay-period activity (Compte et al., 2000; Funahashi, 2015; Funahashi et al., 1989), which relies on positive feedback to allow stimulus-induced activity to persist or resonate, leading to an “attractor” state. (Chumbley et al., 2008; Wimmer et al., 2014; Zipser et al., 1993). Although such active maintenance may also depend upon rapid changes in synaptic weights (Hansel and Mato, 2013;

Pereira and Wang, 2015), these models do not generally allow memory recall from a silent inactive state.

The present work unites persistent activity attractors with silent synaptic storage. In our new class of memory model, both active and silent representations are essential to WM. We propose that persistent activation serves as the *focus of attention* that encodes recent activity patterns into synapses. Rapid plasticity in flexibly-coding neurons allows features to be bound together into objects, with an emergent property being that the last item is maintained actively. Recent, previously-attended items are preserved instead in synaptic traces. They are in a non-privileged state but, importantly, can be re-activated by partial information.

We propose that attention arises from the interaction between two distinct types of neural representation: fixed *feature* neurons, and *freely-conjunctive* neurons (**Fig.1A**). Feature neurons may be sensory, motor or conceptual. They have fixed receptive fields or tuning curves – as observed in posterior cortical areas. In contrast, the freely-conjunctive neurons can rapidly change their connection weights with the feature cells, and therefore their activity does *not* represent a fixed feature or item in memory. Instead, through rapid plasticity on each trial, a conjunctive cell will come to encode a conjunction of simultaneously active features, by forming a reciprocal associative mapping to feature-selective neurons.

Persistent activity arises by mutual excitation between feature and conjunction neurons. The conjunction neurons form a limited-capacity store that can hold many kinds of information in one place. Thus, our model bridges the gap between neuron-level descriptions and the psychological notion of a *general-purpose register*, sometimes termed a "memory slot" (Cowan, 2010; Luck and Vogel, 1997), a concept which has not as yet been characterized at the level of single prefrontal neurons. Such registers are difficult to explain unless individual neurons can encode different types of WM content at different times. Our model permits this by allowing rapid synaptic changes so conjunctive neurons can represent many kinds of information, depending on the recent context.

We suggest that two lines of evidence point to such conjunction neurons being located in prefrontal cortex (PFC): firstly, PFC is highly active in memory and manipulation (Eriksson et al., 2015; Postle et al., 2006), yet secondly, information is not always easy to decode (Christophel et al., 2012; Cogan et al., 2017; Kamiński et al., 2017). Although WM contents can undoubtedly be decoded from many PFC neurons, about 60% of prefrontal neurons appear to be nonselective, and even for those that are selective, they often show less than a 50% modulation of their firing rate by information in WM (Miller et al., 1996; Parthasarathy et al., 2017). This apparently-nonselective component of prefrontal activity could reflect transient and flexible coding by conjunctive units.

In this study, we first aim to provide a single common mechanism accounting for a diverse range of perplexing attention and memory effects. Second, we attempt to explain neurophysiological data where items in memory initially produce persistent activity, which then falls "silent" when attention shifts to new information (Konecky et al., 2017). Third, we aim to explain why many imaging studies conclude that attention and working memory are "distributed" processes involving both prefrontal and sensory brain areas (Christophel et al., 2017; Gayet et al., 2017, 2017; Xu, 2017). In our simulations, we chose to examine the extreme situation where conjunctive neurons are fully nonselective for features. This limiting scenario is of course implausible, since no single prefrontal neuron could receive input from every feature neuron. However we argue that it is a highly illustrative paradigmatic case. In reality prefrontal neurons will necessarily have some degree of selectivity, but here we focus only on characterizing the novel concept of how rapid plasticity can give rise to flexible coding, and therefore we model *purely* conjunctive neurons as distinct from feature-selective neurons.

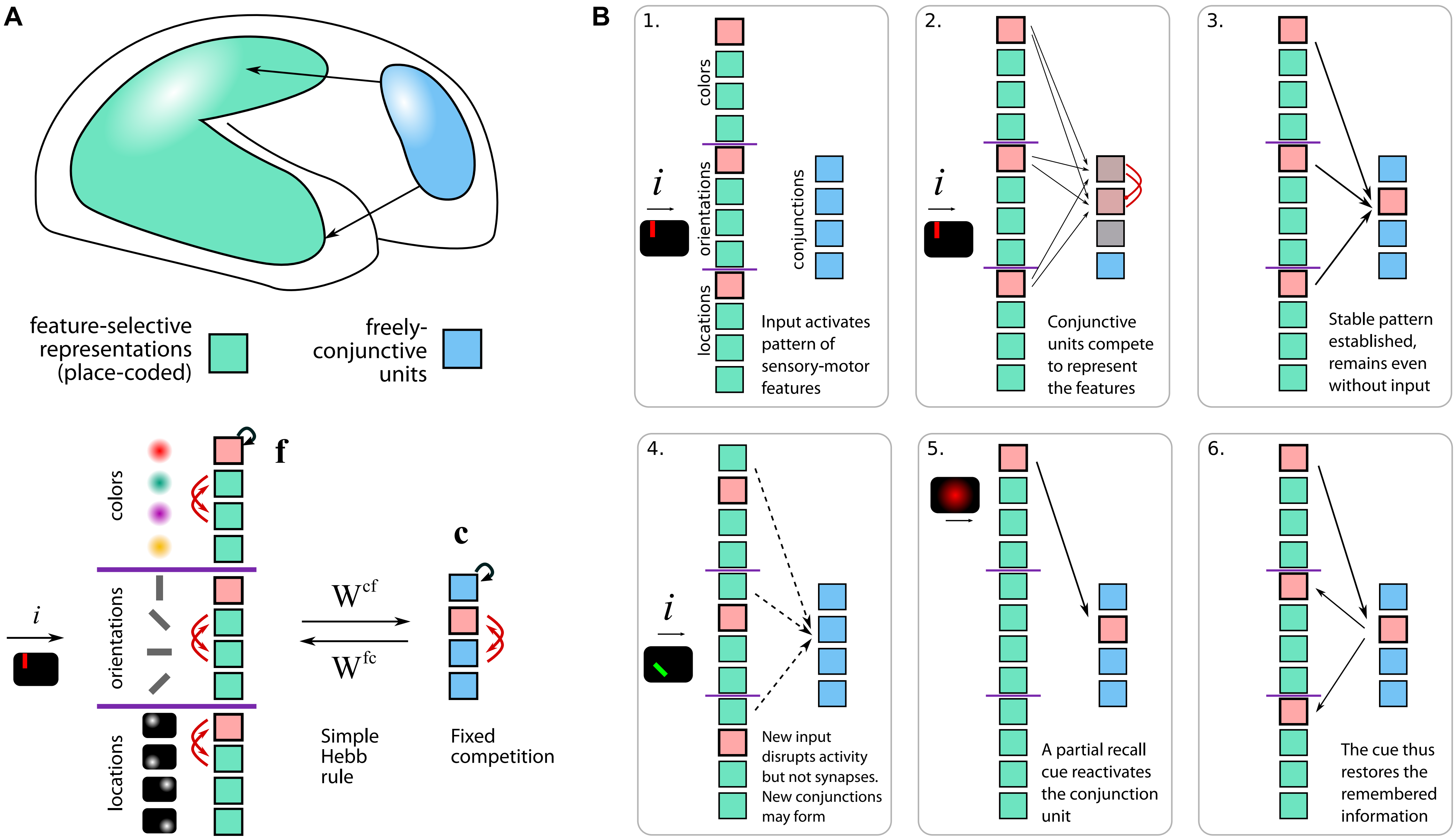

A
feature-selective representations (place-coded)
freely-conjunctive units
B
1.
colors
orientations
locations
conjunctions
i
Input activates pattern of sensory-motor features
2.
i
Conjunctive units compete to represent the features
3.
Stable pattern established, remains even without input
4.
i
New input disrupts activity but not synapses. New conjunctions may form
5.
A partial recall cue reactivates the conjunction unit
6.
The cue thus restores the remembered information
colors
orientations
locations
i
f
c
W^cf
W^fc
Simple Hebb rule
Fixed competition

Fig.1: Conjunctive neurons to support attention and working memory

A Two populations of neurons are distinguished based on their inputs. Posterior neurons (green) encode sensory-motor features, whereas prefrontal neurons (blue) are "conjunctive": i.e. they are able to rapidly increase or decrease their synaptic connectivity with patterns of feature neurons, using a Hebbian associative rule. We simulated 12 feature-selective neurons (**f**) and 4 freely-conjunctive neurons (**c**). An active combination of neurons (pink) causes strengthening of synapses in both directions, producing a stable attractor across brain areas. **c**=conjunctive cells, **f**=feature cells. W=synaptic weights, ***i***=sensory input.

B Sequence of proposed neuronal events during attention, encoding and retrieval in working memory. **1.** Sensory input activates features. In this case a vertical red bar located at the top left of the display activates separate feature neurons tuned to orientation, color and location. **2.** Features excite conjunctive neurons, which compete. **3.** The winning conjunction drives sustained activity. **4.** New input to the system (in this case an oblique purple bar at bottom left) disrupts current firing activity, but synaptic weightings remain. **5.** Probe feature (in this case red colour) re-activates the original conjunctive unit that encoded the red vertical bar. **6.** Conjunctive unit re-activates original features, completing recall.

Results

1. Operation of the Model

When a stimulus is perceived (**Fig.1B; Movie S1**), conjunctive neurons compete through lateral inhibition to become active in response to the combination of active features. In the example shown in **Fig.1** the conjuncton units learn rapidly to encode combinations of color, orientation and location (**Fig.1B.2**). During encoding into WM, the winning conjunctive unit sustains the activity of all co-active feature neurons through mutual excitation. This strengthens synapses in both directions through rapid Hebbian plasticity, further stabilizing the active pattern. Once a conjunctive unit succeeds in reciprocally activating a set of feature units, *attention is focused* on the activated features, binding the features of a compound stimulus into a perceptual object.

The reciprocal feature-to-conjunctive synapses keep the novel combination of features persistently active, even when the stimulus is no longer present (**Fig.1B.3**).

When a new stimulus arrives, a new pattern of sensory input destabilizes internal activity, thus triggering a shift of attention towards the newly activated features. A new conjunction may win, to become the new focus of attention. Crucially, however, synapses between the previous object's constituent features and one particular conjunctive unit remain strengthened even after those neurons fall silent (**Fig.1B.4**). Thus, presenting any one feature of a previously attended object (e.g. color, as shown in **Fig.1**) will act as a memory probe, re-activating the corresponding conjunction neuron (**Fig.1B.5**), and therefore also the other features that were associated with it (**Fig.1B.6**). The object's features are therefore recalled by auto-associative pattern completion, which brings them back into an attended, foreground state. Separate objects must always be encoded sequentially, which we suggest is plausible in light of the empirically observed attentional bottleneck in feature binding (Reynolds and Desimone, 1999).

To demonstrate the power of the model, we simulated a common visuospatial WM task (**Fig.2A**) in which participants remember the orientations of a set of colored bars (e.g. Gorgoraptis et al., 2011; Pertzov et al., 2016). Neurons were modelled as firing-rate units obeying a Hebbian plasticity rule (see **Methods**). Memory items were composed of combinations of features, and up to four unique items were presented sequentially to the feature units. After a delay, we probed one of the items by activating its color-feature alone, and recording whether its orientation was subsequently re-activated. Remarkably, just four color, orientation, location and conjunctive neurons each are needed to explain a wide range of behavioral and neurophysiological data, which no models have yet captured (**Table S1**).

Crucially both the activation and learning equations were implemented continuously over a block of trials, with blank input in between trials, so that encoding, recall and interference from the previous trial all arose naturally from the way stimuli were presented. We tuned the model to perform at levels comparable to humans at this task (see **Methods**). For clarity, here we elected to keep the model's operation almost identical for all the simulations, even though the experimental data we match come from a variety of tasks and measures.

2. Capacity limits and serial order in WM

A key feature of WM is its limited capacity. The more items held in memory, the less accurately they are remembered (Luck and Vogel, 1997; Bays and Husain, 2008). Simulated recall accuracy (**Fig.2B**) matched the set-size effect from classical visuospatial WM experiments (**Fig.2C**). This is because each additional stimulus competes for conjunctive neurons, and may corrupt or overwrite synaptic traces of previously-seen objects. Whether a previous item is overwritten is determined by the how well the currently-active features match the existing synaptic weights, which are themselves continuously subject to Hebbian rules. Therefore in our model, capacity is limited by interference between items in memory, in line with convergent evidence from multiple WM domains (Almeida et al., 2015; Farrell et al., 2016; Oberauer and Lewandowsky, 2014). Note that accuracy is lower than human data because the model chooses between four options rather than two, but varying the model parameters can make it arbitrarily more accurate (**Fig.S8,S9**). Importantly the model predicts the counterintuitive finding that storing extra features within a single object either occurs automatically(Allen et al., 2006) or else may incur no extra cost (Luck and Vogel, 1997). In fact our model predicts that in some situations a benefit can be observed for adding a new irrelevant but distinguishing feature to each object (**Fig.S7**).

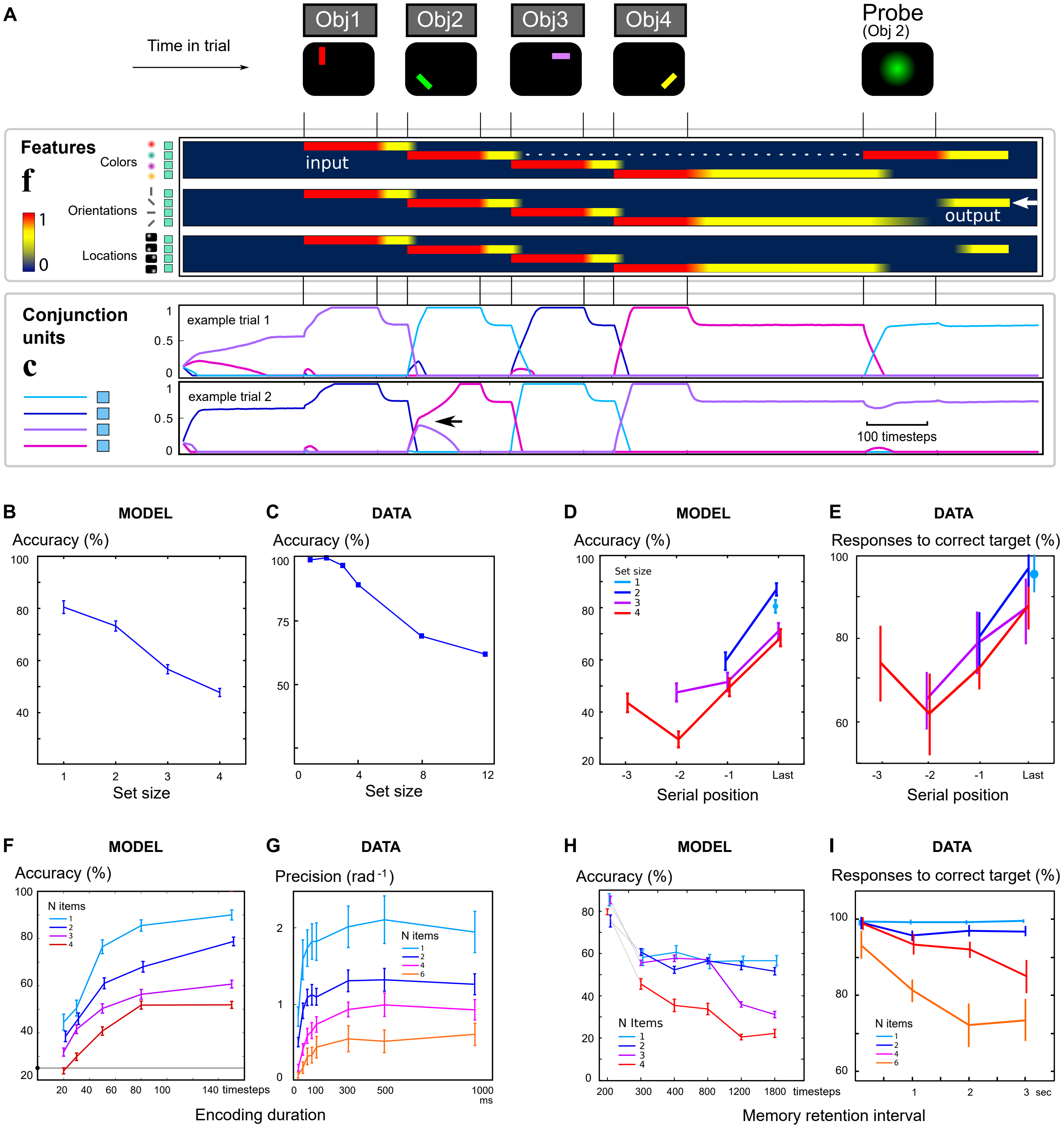

A
Time in trial
Obj1
Obj2
Obj3
Obj4
Probe
(Obj 2)
Features
f
Colors
Orientations
Locations
1
0
input
output
Conjunction units
c
example trial 1
example trial 2
1
0.5
0
100 timesteps
B
MODEL
Accuracy (%)
Set size
C
DATA
Accuracy (%)
Set size
D
MODEL
Accuracy (%)
Set size
1
2
3
4
Serial position
-3
-2
-1
Last
E
DATA
Responses to correct target (%)
Serial position
F
MODEL
Accuracy (%)
N items
20 40 60 80 100 140 timesteps
G
DATA
Precision (rad⁻¹)
N items
1
2
4
6
0 100 300 500 1000 ms
Encoding duration
H
MODEL
Accuracy (%)
N Items
200 300 400 800 1200 1800 timesteps
I
DATA
Responses to correct target (%)
N items
1 2 3 sec
Memory retention interval

Fig.2: Predicting visuospatial WM capacity, encoding and decay

A To simulate WM performance, four objects are presented sequentially, by activating feature neurons (**f**, activity depicted as a heatmap from dark blue to red) indicating the color, orientation and location of each item. Conjunctive units (**c**) are shown below as four differently-colored traces. Conjunctive units compete to become active for each object. One conjunctive unit wins for each object, driving activity that persists even after input is removed (yellow parts of heatmap). At the time of the probe, a single feature is stimulated, triggering pattern completion. Recall is accurate if the orientation of the corresponding item is re-activated. Two example trials are shown; note that different patterns of conjunctive units are activated on different trials even for the same stimuli, depending on trial history. Example 1: good encoding. Example 2: weak encoding of the second item. Two conjunctive neurons with similar recent preferences compete to encode object 2 (arrowhead). When it is probed, item 4 is reported instead.

B & C When more items are encoded in the model, recall accuracy is reduced, as observed in data (adapted from Luck and Vogel, 1997).

D & E The last item encoded in the model is recalled better than others, as it remains active in the focus of attention during the delay period, matching observed serial order curves. Figure adapted from (Gorgoraptis et al., 2011) indicates the probability of reporting the target item as calculated by fitting the distribution of responses in a similar task.

F&G Shorter encoding durations reduce modelled recall accuracy. Data from a similar task (adapted from Bays et al., 2011) where adding items reduced both initial encoding rate and asymptote. The model qualitatively reproduces the interaction observed in human performance.

H & I The model predicts faster memory decay when more items are stored. This matches the empirical interaction between memory-set size and delay. Data adapted from (Pertzov et al., 2016) shows the modelled probability of reporting the target.

3. Serial order effects

When we remember a sequence of objects, we recall the first and last objects better (primacy and recency). Our model can reproduce both of these effects. Simulated performance (**Fig.2D**) matched the serial position curve obtained in WM experiments (**Fig.2E**). The simulation suggests that neutrally, primacy benefits arise because the first object in a trial does not need to compete with ongoing persistent activity from a previous item (**Fig.1B4**). In our model this relies on the fact that, at the start of each trial, feature units are inhibited but previous synaptic weights are not erased – though there is no explicit signal to forget items from the previous trial. Recency benefits arose for two reasons: the finally-encoded item did not incur retroactive interference from subsequent items, and was already in an active rather than silent state at recall.

4. Encoding and maintenance

The time-course of encoding was interrogated by presenting items for brief durations, and demonstrated exponential saturation with an asymptote dependent on the number of items encoded. In a similar empirical study (Bays et al., 2011), memory precision (1/standard deviation of response angular error) followed a similar pattern. In that study, the probability of choosing the target was not calculated, but their reported precision appears to correspond well to our model's probability of reporting the correct target orientation (**Fig. 2F&G**).

Simulations demonstrated that memory deteriorates faster when increasing numbers of items are remembered **(Fig.2H&I)**, as shown in a recent study (Pertzov et al., 2016). This arises because a greater proportion of items are held in an unattended state. Unattended items are more vulnerable to interference, because their synapses are gradually weakened over time according to the plasticity rule. Our model also makes the strong prediction that an item stored in an attended state (e.g. the final item in a sequence) is more robust to decay over time.

5. Shifting the Focus of Attention

An important advance over other models, is the ability of our model to re-activate a previous item by bringing it into the focus of attention. The logic here is that sensory input can guide attention by pattern-completion. In behavioral experiments, an “incidental” task inserted into the memory delay can shift attention to one of the items in memory (**Fig.3A**) (Zokaei et al., 2014a) bringing it into the foreground. We simulated “retro-cueing” one of the items during the memory delay by presenting one of its features for a brief period, which brought that item back into the focus of attention (**Fig. 3B**). The external cue could thus re-activate a memory item which was previously encoded silently. Note that this simulation illustrates how feature-selective units can exhibit task-dependent modulation because they also receive non-sensory input through rapidly-plastic synapses from the conjunctive units.

Recall of the incidentally-cued item improved, compared to the uncued item (**Fig.3C**), matching experimental data (**Fig.3D**). This attentional shifting also explains how cues that indicate which item will be probed (predictive retro-cues, Rose et al., 2016) improve performance, even paradoxically when controlling for the retention interval’s duration (Myers et al., 2017).

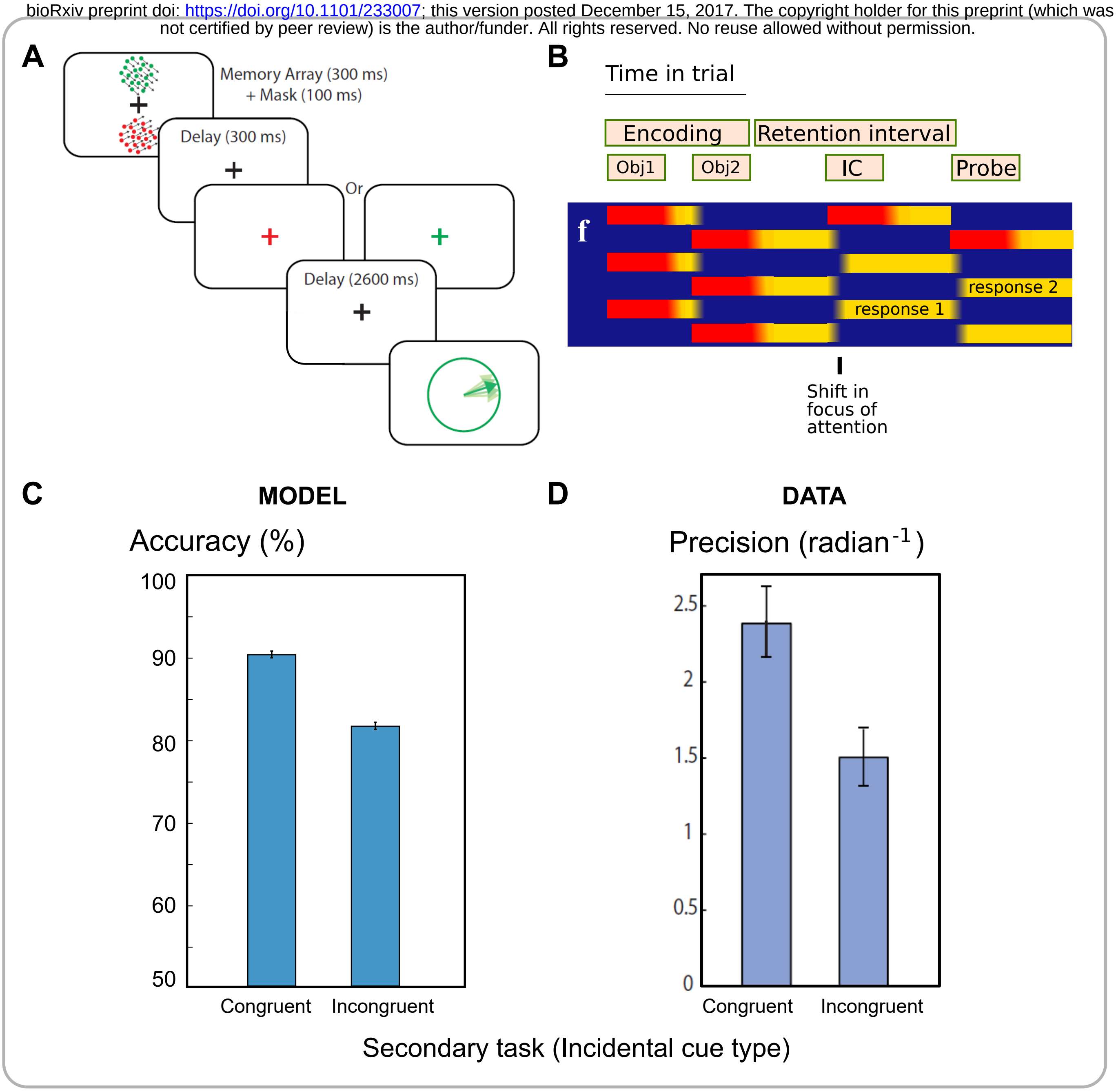

A
Memory Array (300 ms)
+ Mask (100 ms)
Delay (300 ms)
Or
Delay (2600 ms)
B
Time in trial
Encoding
Retention interval
Obj1
Obj2
IC
Probe
f
response 2
response 1
Shift in focus of attention
C
MODEL
Accuracy (%)
100
90
80
70
60
50
Congruent
Incongruent
D
DATA
Precision (radian^{-1})
2.5
2
1.5
1
0.5
0
Congruent
Incongruent
Secondary task (Incidental cue type)

Fig.3: Shifting the focus of attention in WM

A Experiment (Zokaei et al., 2014a) where participants remembered two items, each comprising three features: color, location and orientation. During the retention interval, a color was shown, and as a secondary task, the location of the corresponding object had to be recalled. At the end of the delay, a color was shown which could indicate the same ("congruent") or different ("incongruent") object than the one tested during the delay. Participants then reported the orientation of the corresponding object. Reproduced under the terms of the Creative Commons Attribution 3.0 Unported (CC BY 3.0) license (https://creativecommons.org/licenses /by/3.0) from figure 1A of Zokaei et al. 2014, The Journal of Neuroscience. January 1, 2014. 34(1);158-162..

B Similar events were simulated, with an incidental cue (IC) during the delay. If the first object was cued, then persistent delay activity shifted to the cued item.

C&D The model predicts that the item in the focus of attention before recall is reported more accurately, matching data. Reproduced under CC BY license from figure 1B of Zokaei et al. 2014.

6. Recall

After the probe feature was activated, it took a number of time steps for the conjunction and response feature units to become active. We measured this time to obtain reaction time predictions, which varied inversely with accuracy similar to empirical data (**Fig.S1**).

The process of recall may also be susceptible to interference, because it effectively uses pattern completion to re-activate the other features of the corresponding object. In particular, the memory probe itself can interfere with recall, for example if it contains a feature on the dimension that needs to be reported (**Fig.S2**), in line with empirical probe-interference effects (Souza et al., 2016). Interference of another kind arises when recalling items as a whole series: often the preceding or following item is reported instead (Smyth, 1996; Solway et al., 2012). Although our simulations probe a single item at a time, they still demonstrate such “transposition errors”, where consecutively presented objects are confused (**Fig.S3**).

7. Neural encoding of items in WM

Three major predictions emerge about neural decoding. First, an emergent property of our framework is that sustained activity represents a single item held in memory (Funahashi, 2017), but not multiple items (Lara and Wallis, 2014). We used a linear decoder to extract information about one feature of one of the items in WM, after items had been encoded. The predictions of the model for decodability from feature-selective neurons (**Fig.S4**) are in keeping with human and nonhuman physiological data demonstrating that only the attended WM item is decodable using standard techniques (Konecky et al., 2017; Lewis-Peacock et al., 2012; Sprague et al., 2016). Second, evoking neural activity by stimulation can restore decodability from EEG signals (Rose et al., 2016; Wolff et al., 2017). We simulated transcranial magnetic stimulation (TMS) by an indiscriminate pulse of activation to feature neurons **(Fig.4A)**, and decoded one feature dimension from feature-selective units (**Fig.4B**). If the model's color and orientation feature dimensions are considered as mapping to spatial location and stimulus category respectively, then the simulation matches the effects of TMS on decoding (**Fig.4C**) (Rose et al., 2016), or if they are instead mapped to spatial location and orientation, then the model's results reproduces the effects of a high-energy visual pulse (Wolff et al., 2017). Simulating a stronger pulse of stimulation disrupted attention, but not synapses. This worsened recall of the attended item, yet contrarily improved unattended items (**Fig.4D&E**), precisely as demonstrated empirically (Zokaei et al., 2014a).

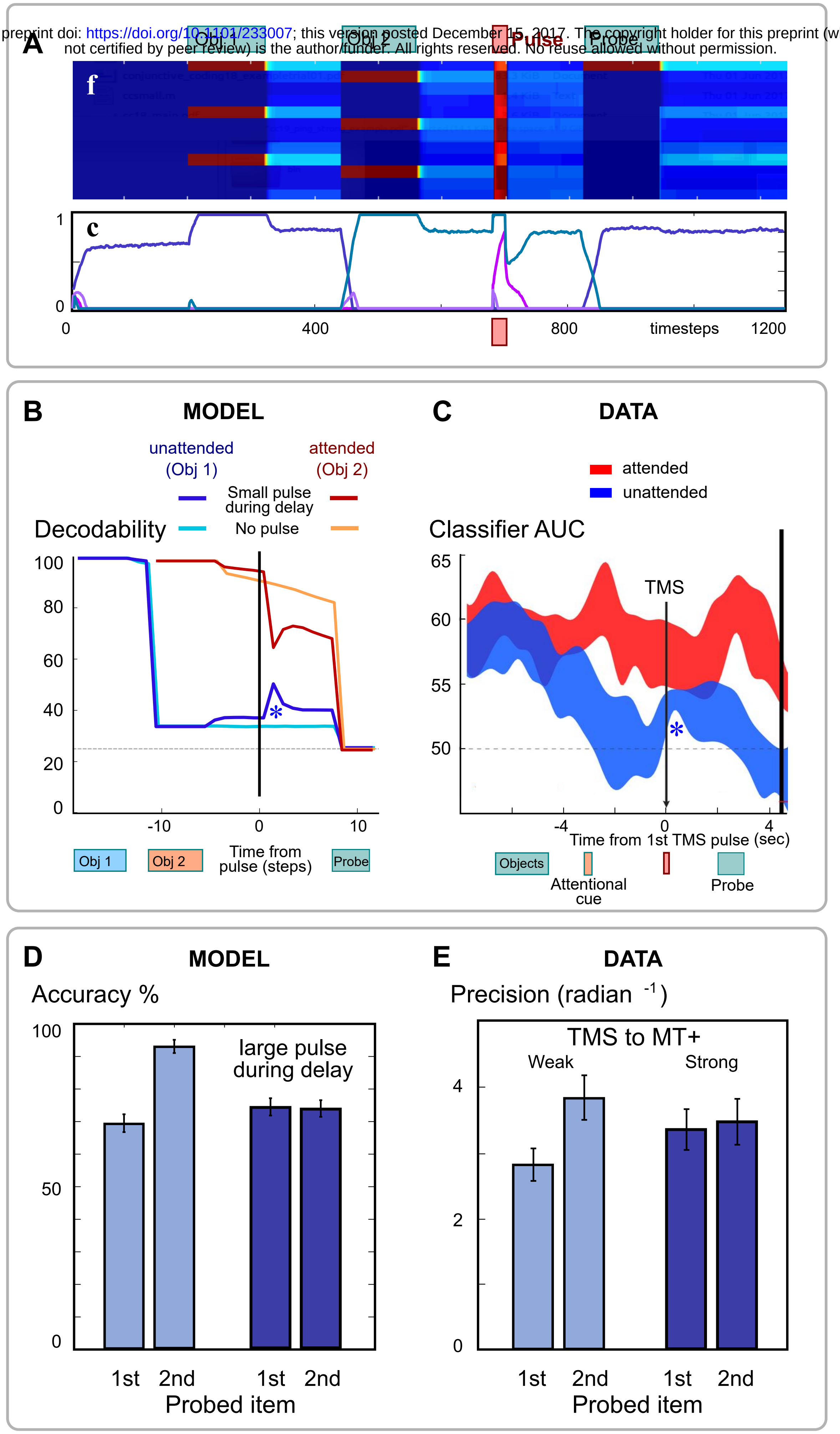

A
Obj 1
Obj 2
Pulse
Probe
f
c
1
0
0
400
800
timesteps
1200
B
MODEL
unattended
(Obj 1)
attended
(Obj 2)
Small pulse during delay
No pulse
Decodability
100
80
60
40
20
0
-10
0
10
*
Obj 1
Obj 2
Time from pulse (steps)
Probe
C
DATA
attended
unattended
Classifier AUC
65
60
55
50
TMS
*
-4
0
4
Time from 1st TMS pulse (sec)
Objects
Attentional cue
Probe
D
MODEL
Accuracy %
100
50
0
large pulse during delay
1st
2nd
1st
2nd
Probed item
E
DATA
Precision (radian -1)
TMS to MT+
Weak
Strong
4
2
0
1st
2nd
1st
2nd
Probed item

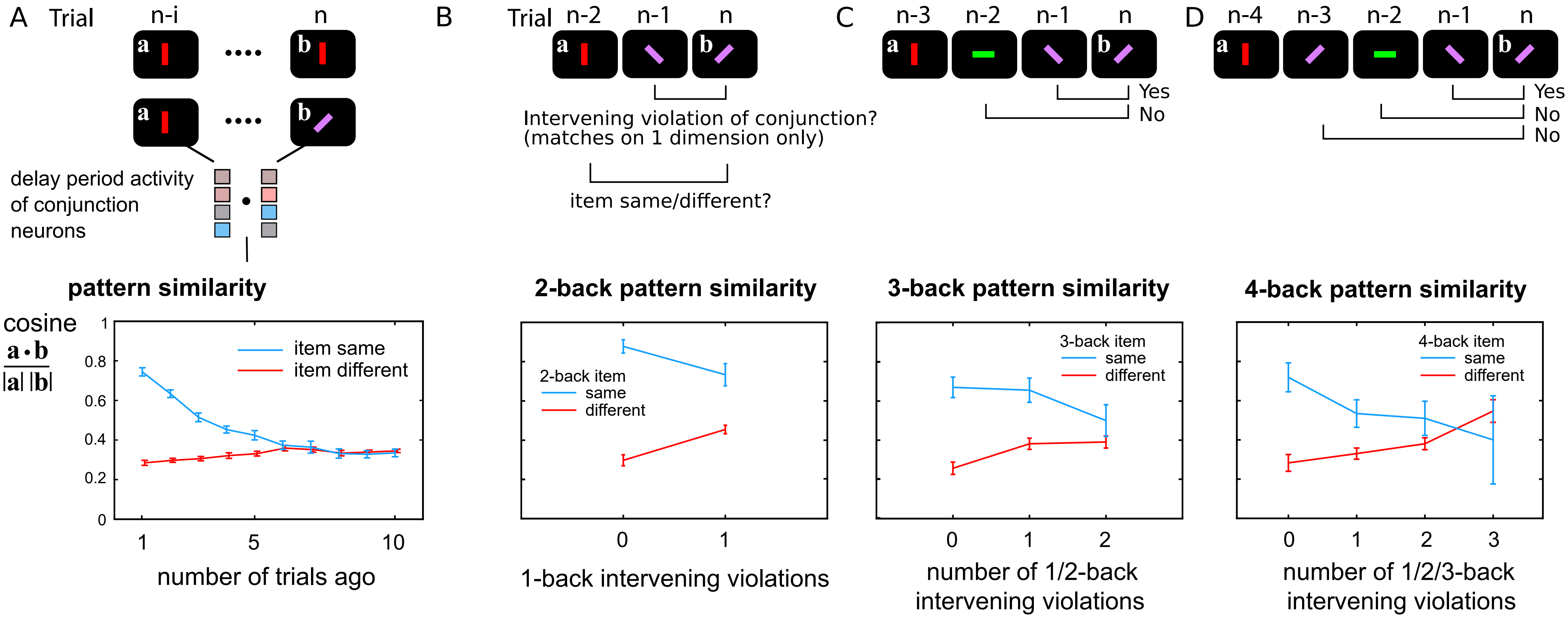
Fig.4: Introducing a pulse of excitation during the delay period

A After presenting two items, during the delay all feature neurons **f** received an excitatory input pulse **i**=+1, consequently activating conjunction neurons.

B&C We tried decoding the identity of each of the two stimuli from feature neuron activity. Although the first object was not decodable without the pulse, it became transiently distinguishable (*) after the pulse. This matches the observed increase in decodability after TMS (Rose et al., 2016).

D&E Stronger pulses altered model performance, abolishing the benefit for the second item, which was in the focus of attention. The pulse disrupted persistent activity, re-instating competition between conjunctive neurons. The prediction matches observed effects of TMS targeting motion-selective cortex (Zokaei et al., 2014a).

Third, the model predicts that decoding from prefrontal cortex is unreliable (Lee and Baker, 2016). This is because the concept of a receptive field breaks down for conjunctive neurons. The same activity can have *different meanings* on different trials, dependent on residual synaptic weights from previous trials. Such neurons should show much stronger representations over short timescales. We predict this will manifest behaviorally, with better recall for a feature combination present on the previous trial (**Fig.S5**), because the same conjunction unit will be re-used. Moreover, neural activity patterns in conjunction neurons predict stimuli strongly if we consider data only from *contiguous* pairs of trials, compared to data from temporally-separated trials (**Fig.5A**), and the pattern similarity should be even lower when intervening stimuli involve a recombination of the features (**Fig.5B-D**). This confirms that each conjunctive neuron's activity represents different things, as its synaptic weights change. Such a system can flexibly encode a broad variety of novel information rapidly, without incurring the combinatorial explosion that haunts previous fixed-selectivity models (Matthey et al., 2015; Postle et al., 2006).



A
Trial
n-i
n
a
b
a
b
delay period activity
of conjunction
neurons
pattern similarity
cosine
$\frac{\mathbf{a} \bullet \mathbf{b}}{|\mathbf{a}|\,|\mathbf{b}|}$
item same
item different
1
0.8
0.6
0.4
0.2
0
1
5
10
number of trials ago
B
Trial
n-2
n-1
n
Intervening violation of conjunction?
(matches on 1 dimension only)
item same/different?
2-back pattern similarity
2-back item
same
different
0
1
1-back intervening violations
C
n-3
n-2
n-1
n
Yes
No
3-back pattern similarity
3-back item
same
different
0
1
2
number of 1/2-back
intervening violations
D
n-4
n-3
n-2
n-1
n
Yes
No
No
4-back pattern similarity
4-back item
same
different
0
1
2
3
number of 1/2/3-back
intervening violations

Fig.5: Conjunctive unit representations are stable over short timescales

Conjunctive units change their selectivity over short periods. If selectivity were stable, neural patterns should be similar when the stimulus is the same. We compared similarity of the pattern of an earlier trial, to trial *n,* during the delay periods of a series of 1-item trials.

A) The similarity of the conjunctive neurons' delay activity pattern is calculated for trials where the stimuli were identical (blue line) or different (red line). Patterns were more similar when stimuli were the same, compared to when stimuli were different, indicating "classical encoding" at least for nearby trials. This classical behavior decreased with the temporal distance between trials. Since we modelled the extreme case where neurons are *purely* conjunctive, with no feature selectivity, consistency of pattern is completely abolished after about 6 trials.

B-D) The model predicts that interference reduces pattern similarity over time by overwriting the synaptic weights. If the objects in intervening trials share one feature with the *n*th trial object, but mismatch on the other feature dimension, then we say the conjunction between the two feature dimensions is "violated".

B) When the intervening trial contained a violation, the patterns on the *n*-2 and *n*th trials reflected the stimuli much more weakly, indicating interference or overwriting of the original conjunction.

C and **D**) Trials 3-back and 4-back were similarly examined, this time asking how many intervening conjunction violations occurred. The more overwriting that occurred between the *n*-3 and *n*th trials, the less classical encoding could be observed.

8. Simulation of task sets

The same system can also implement stimulus-response rules, if some feature neurons represent motor plans. In this case, we encode a *task rule* by attending to a stimulus and a motor plan together. For example, if a left-hand movement plan is activated while a red color-feature is simultaneously activated, they will be encoded together into working memory. The conjunction of sensory features with a motor plan creates a task-set mapping (Duncan et al., 2012). Later, that stimulus can also re-activate the corresponding motor plan by pattern-completion, triggering the movement – so that the stimulus generates a response. Task sets can therefore be rapidly formed by sequentially attending to stimulus-response pairs (Curtis and D'Esposito, 2003), and deciding on an action is simply the motor analogue of WM recall.

To simulate stimulus-response mapping, we presented the task rules sequentially, each consisting of a pairing between one color and one response (**Fig.S6A**). Then on each subsequent trial, a single color from the set was shown, and the response was recorded. The model reproduces Hick's law, in which response times are longer in situations when more response options are possible in the current task set (**Fig.S6B**) (Proctor and Schneider, 2017). It also produces faster reaction times when the response is repeated from the previous trial (**Fig.S6C**), in line with experimental evidence (Schvaneveldt and Chase, 1969).

In this situation, the role of prefrontal conjunctions can be viewed as *controlling* representations in posterior cortex, i.e. routing information from perceptual to motor representation as governed by task sets held in working memory, a role classically assigned to executive/supervisory attention. Critically the model predicts that, because the task rules are held in WM across many trials rather than being repeatedly overwritten, the current stimulus and response (i.e. the active task rule) are consistently decodable from conjunctive neurons, until the rules change (**Fig.S6D**). This contrasts with WM storage, where frequent overwriting leads to poor decoding, and may explain why task rules have generally been easier to decode from PFC (Reverberi et al., 2012; Sakai, 2008).

Discussion

The model of freely-conjunctive neurons presented here accounts for both sustained firing and activity-silent synaptic traces in WM (Silvanto, 2017; Stokes, 2015), and consequently makes a range of testable behavioral and neural predictions (**Table S1**). This neuronal framework provides a parsimonious mechanism for feature binding, general-purpose memory 'slots', and task sets. The model reproduces classical WM effects of capacity, serial order, encoding rate, temporal decay (**Fig.2**), reaction times, and transposition errors (Fig.S1&3), as well as the ability to switch attention between items within memory – a phenomenon that evades most current models (**Fig.3**). At a neural level, it explains why it is difficult to decode memory contents from prefrontal activity, why only the item in the focus of attention can be decoded elsewhere (**Fig.S4**). Further it explains why decodability can be restored by re-focusing an unattended item, or after a perturbation such as transcranial magnetic stimulation (TMS) or bottom-up input (Rose et al., 2016; Wolff et al., 2017), which presumably re-activate the conjunctive neurons and thus an object's features through synaptic traces (**Fig.4**). The model also makes strong novel predictions about probe interference, trial-to-trial effects (Fig.S2&5), and disruption of neural pattern similarity by intervening stimuli (**Fig.5**).

Relation to previous models

Rapid plasticity has long been demonstrated in cortical neuron receptive fields (Edeline et al., 1993) and may arise through a variety of synaptic mechanisms (Zucker, 1989; Tsodyks and Markram, 1997; Fischer et al., 1998; Dittman et al., 2000; Jensen et al., 1996; Malsburg, 1981). It has previously been suggested to underlie short-term retention of information in a network of selective and nonselective neurons (Mongillo et al., 2008), similar to theories of hippocampal binding in long-term memory (Burgess and Hitch, 2005; Rizzuto and Kahana, 2001). Rapidly plastic networks account well for serial recall (Farrell and Lewandowsky, 2002; Fiebig and Lansner, 2017; Sandberg et al., 2003) but cannot distinguish between focused and unfocused items in memory, and do not explain sustained delay-period firing (Funahashi, 2017). Sustained activity models may also account for some attentional effects on decodability (Schneegans and Bays, 2017) but cannot reinstate information that becomes fully undecodable.

To support flexible attractor states, we postulated two distinct modes of neural representation (**Fig.1**). First, feature-selective neurons are traditional, place-coded ("labelled-line") units. They are selective because they have some fixed, non-plastic inputs (or in the case of motor units, fixed outputs). But if plasticity modifies both the input and output synapses of a neuron, the meaning or interpretation of a neuron's firing will also change. This is simply because neurons *code* information only in virtue of their inputs and outputs. Plasticity therefore begets a new category of flexibly-coding neurons, where the information signaled by firing is protean and dependent on the history on each trial. Decoding the fine-grained identity of stimuli from prefrontal cortex is unreliable compared to posterior sensorimotor regions (Cogan et al., 2017; Lee and Baker, 2016), because the idea of a receptive field breaks down. Standard decoding methods assume trial-to-trial stability of activation patterns to represent a given feature, and so do not measure the sequential effects we predict. This flexible coding scheme is crucial for our model to generate two phenomena. First, it permits sustained activity that is guided dynamically by task sets or objects in memory, which we postulate corresponds to attentional interactions between frontal and temporo-parietal regions. Second, because individual neurons can encode different things at different times, information must *compete* to be encoded by any conjunctive neuron – thus leading to a capacity limit for general-purpose information storage, observed in both WM and attention. This may help resolve a long-standing theoretical debate on whether working memory consists of pointers, or activated long-term memory (Norris, 2017): conjunctive neurons act as pointers that activate long-term memories.

Relaxing the model's assumptions

In this study we deliberately chose to study the simplest possible model that could support conjunctive neurons. The very small number of neurons, and their simple learning and dynamics, makes it much easier to see how they interact to generate the novel predictions. Moreover it is much more transparent where the model can or cannot match existing data. Naturally there are many directions in which the model needs to be extended, to fully reproduce the phenomena observed in real neurons. A number of its assumptions can plausibly be relaxed.

1. Pure flexible and stable representations

For simplicity we have treated conjunctive neurons as "pure": i.e. that they are homogeneous and domain-general, resulting in inability to decode information across many trials. This is certainly implausible because all-to-all connections between PFC and feature-selective neurons are not feasible. Moreover, how can we then explain studies that *do* demonstrate decoding of WM from prefrontal areas? In reality, we envisage that each conjunctive neuron is likely to receive inputs from only a subset of feature neurons. In order for conjunctive neurons to bind features into objects, these inputs must at least include multiple feature dimensions *and* multiple features in each dimension. The model is therefore potentially compatible with the presence of mixed selectivity (Rigotti et al., 2013), which would provide a background of weak input selectivity based on the presence or absence of connections, upon which rapid plasticity is superimposed.

Further, there may also be significant topography in conjunctive cells connectivity. For example, different regions of prefrontal cortex may be specialized for remembering different kinds of information (Romanski, 2004). This may have two desirable consequences. First, aspects of the attended object – especially information that is highly topographical in posterior areas, such as stimulus category and spatial location – would be consistently decodable from prefrontal cortex (Lee and Baker, 2016) but will be modulated by relevance (Kornblith and Tsao, 2017). Second, conjunctive neurons in different prefrontal subregions may connect preferentially to visual, motor or auditory cortex, which could account for the separability of visuospatial and phonological WM and also their overlap (Morey et al., 2011). We note that stable mixed selectivity, even without plasticity, could in some situations produce binding and capacity limits (Matthey et al., 2015). However without additional mechanisms, it would presumably not account for attentional shifts, activity-silent storage, or apparent control over posterior cortical areas, and moreover makes it challenging to internally ‘read-out’ WM contents.

We treated “features” as just simple perceptual attributes, but we believe that our class of feature-selective neurons could include any aspect of the world that is encoded in a stable way, including those aspects that incorporate long-term knowledge, such as object identity, category, or even linguistic information such as word meanings. These attributes are likely to be encoded stably in posterior cortical areas, in contrast to the temporary combinations of information represented in an ephemeral way – e.g. for online manipulation – as typified by our conjunctive neurons. The current simulations used only a single, rapid learning rate, but it remains to be studied how this could be reconciled with longer-term learning.

2. Internal control over attentional shifts

We have assumed that attentional shifts are externally cued. Endogenous shifts of attention are not modelled. One way of implementing internally-generated attentional modulation would be to de-stabilize the persistent activity by adding delayed suppression, or refractoriness, to the competitive conjunctive neurons. The result would be that, after an object is attended, its activity is extinguished after a delay, leading to a transient and unstable focus of attention. Akin to some models of visual attention guidance (Itti and Koch, 2001), attention will then be successively re-deployed towards the weakest-represented features in WM. This could potentially account for four key phenomena: (a) rehearsal, in which attention moves sequentially between items during a memory delay, (b) the ability to free-recall WM items in order, (c) the guidance of visual search, and (d) in our model, to permit serial encoding of a simultaneously-presented memory array.

Although WM maintenance commonly engages PFC, evidence from neuropsychology and functional imaging suggests PFC's role includes cognitive control, WM manipulation, and response selection, rather than simply WM storage (Bechara et al., 1998; D'Esposito and Postle, 1999; Rowe et al., 2000; Thompson-Schill et al., 2002), and it remains to be tested whether the conjunctive neurons we propose can perform such functions. For example, we cannot account for the ability to "gate out" distractors, and prevent them from being encoded in WM. If sensory input is sufficiently weak, in our model, it can cause transient activation of conjunctive neurons without capturing attention, such that the ongoing attractor state remains stable. But how could *irrelevant* distractors be ignored, while still allowing relevant inputs to capture attention? To achieve this, conjunctive units would themselves need to be under higher-level control. The current model, with only one layer of conjunction units, does not explain higher order control of attention, since sufficiently-strong bottom-up stimuli that match a conjunction will always tend to re-activate that conjunction and thus capture the focus attention. The conjunction and feature neurons together simply act as a "matched filter" amplifying patterns that have recently been active (Chrysikou et al., 2014; Hayden and Gallant, 2013). Perhaps gating *vs* granting access to working memory by preventing this might be controlled by interactions between prefrontal cortex and the basal ganglia (Badre, 2012; Chatham et al., 2014).

3. Location of conjunctive neurons

Conjunctive-coding neurons might not be confined to prefrontal cortex. Other regions that play a role in working memory, such as the hippocampus, basal ganglia or thalamus, might also contain freely conjunctive neurons. Moreover, there may be a continuum or overlap of mechanisms subserving working memory and episodic memory (Fiebig and Lansner, 2014). However the volatile synaptic weights we propose would produce strong but evanescent trial-by-trial selectivity changes (**Fig.5**), quite unlike the rapid but long-lasting associations proposed in the hippocampus. A more intriguing possibility is that both freely-conjunctive and stable-feature neurons are actually present in the *same* brain regions, with a spectrum between highly-plastic and stably-coding neurons.

4. Spiking and synchrony

Neurophysiological evidence points to synchrony of neuronal firing as a key feature of attention (Myers et al., 2017). Since we only simulated firing rates, synchrony and oscillations are not observable. Our reciprocal activation and "reverberatory" sustained firing could lead to spiking synchrony that is cross-region and attention-dependent. The current model makes no specific predictions regarding theta- and gamma-band activity. To generate such predictions, a spiking model, possibly including the thalamus and intracortical oscillations, may be needed. This is far from trivial and further study is required to determine if a spiking implementation of this network would generate the same predictions. An open question is how this might be reconciled with evidence that multiple items may be stored in an electrically active state. For example, could several attractors be simultaneously active, kept distinct through inhibition (Wei et al., 2012) or multiplexed by phase relative to oscillations (Jensen et al., 2002; Lisman and Jensen, 2013)? Our model is incompatible with these possibilities.

In summary, a single architecture captures both persistent activity attractors and silent synaptic memory. We introduce a new scheme of transient flexible neuronal coding, that can support many empirical phenomena (**Tables S1/2**) including the "focus of attention", and generates numerous testable neural predictions.

References

Allen, R.J., Baddeley, A.D., and Hitch, G.J. (2006). Is the binding of visual features in working memory resource-demanding? J. Exp. Psychol. Gen. *135*, 298.

Almeida, R., Barbosa, J., and Compte, A. (2015). Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. J. Neurophysiol. *114*, 1806–1818.

Baddeley, A. (1996). The fractionation of working memory. Proc. Natl. Acad. Sci. *93*, 13468–13472.

BADDELEY, A. (2000). Short-Term and Working Memory. Oxf. Handb. Mem. 77.

Badre, D. (2012). Opening the gate to working memory. Proc. Natl. Acad. Sci. U. S. A. *109*, 19878–19879.

Barch, D.M., Braver, T.S., Nystrom, L.E., Forman, S.D., Noll, D.C., and Cohen, J.D. (1997). Dissociating working memory from task difficulty in human prefrontal cortex. Neuropsychologia *35*, 1373–1380.

Bays, P.M., and Husain, M. (2008). Dynamic Shifts of Limited Working Memory Resources in Human Vision. Science *321*, 851–854.

Bays, P.M., Gorgoraptis, N., Wee, N., Marshall, L., and Husain, M. (2011). Temporal dynamics of encoding, storage, and reallocation of visual working memory. J. Vis. *11*, 6–6.

Bechara, A., Damasio, H., Tranel, D., and Anderson, S.W. (1998). Dissociation Of Working Memory from Decision Making within the Human Prefrontal Cortex. J. Neurosci. *18*, 428–437.

Burgess, N., and Hitch, G. (2005). Computational models of working memory: putting long-term memory into context. Trends Cogn. Sci. *9*, 535–541.

Chatham, C.H., Frank, M.J., and Badre, D. (2014). Corticostriatal Output Gating during Selection from Working Memory. Neuron *81*, 930–942.

Christophel, T.B., Hebart, M.N., and Haynes, J.-D. (2012). Decoding the Contents of Visual Short-Term Memory from Human Visual and Parietal Cortex. J. Neurosci. *32*, 12983–12989.

Christophel, T.B., Klink, P.C., Spitzer, B., Roelfsema, P.R., and Haynes, J.-D. (2017). The Distributed Nature of Working Memory. Trends Cogn. Sci. *0*.

Chrysikou, E.G., Weber, M.J., and Thompson-Schill, S.L. (2014). A Matched Filter Hypothesis for Cognitive Control. Neuropsychologia *62*, 341–355.

Chumbley, J.R., Dolan, R.J., and Friston, K.J. (2008). Attractor models of working memory and their modulation by reward. Biol. Cybern. *98*, 11–18.

Chun, M.M., Golomb, J.D., and Turk-Browne, N.B. (2011). A Taxonomy of External and Internal Attention. Annu. Rev. Psychol. *62*, 73–101.

Cogan, G.B., Iyer, A., Melloni, L., Thesen, T., Friedman, D., Doyle, W., Devinsky, O., and Pesaran, B. (2017). Manipulating stored phonological input during verbal working memory. Nat. Neurosci. *20*, 279–286.

Compte, A., Brunel, N., Goldman-Rakic, P.S., and Wang, X.-J. (2000). Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. Cereb. Cortex *10*, 910–923.

Conway, A.R.A., Kane, M.J., and Engle, R.W. (2003). Working memory capacity and its relation to general intelligence. Trends Cogn. Sci. *7*, 547–552.

Cowan, N. (2010). The Magical Mystery Four How Is Working Memory Capacity Limited, and Why? Curr. Dir. Psychol. Sci. *19*, 51–57.

Curtis, C.E., and D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. Trends Cogn. Sci. *7*, 415–423.

Desimone, R., and Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. Annu. Rev. Neurosci. *18*, 193–222.

D'Esposito, M., and Postle, B.R. (1999). The dependence of span and delayed-response performance on prefrontal cortex. Neuropsychologia *37*, 1303–1315.

Dittman, J.S., Kreitzer, A.C., and Regehr, W.G. (2000). Interplay between Facilitation, Depression, and Residual Calcium at Three Presynaptic Terminals. J. Neurosci. *20*, 1374–1385.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends Cogn. Sci. *14*, 172–179.

Duncan, J., Seitz, R.J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A., Newell, F.N., and Emslie, H. (2000). A Neural Basis for General Intelligence. Science *289*, 457–460.

Duncan, J., Schramm, M., Thompson, R., and Dumontheil, I. (2012). Task rules, working memory, and fluid intelligence. Psychon. Bull. Rev. *19*, 864–870.

Edeline, J.-M., Pham, P., and Weinberger, N.M. (1993). Rapid development of learning-induced receptive field plasticity in the auditory cortex. Behav. Neurosci. *107*, 539.

Eriksson, J., Vogel, E.K., Lansner, A., Bergström, F., and Nyberg, L. (2015). Neurocognitive Architecture of Working Memory. Neuron *88*, 33–46.

Farrell, S., and Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. Psychon. Bull. Rev. *9*, 59–79.

Farrell, S., and Lewandowsky, S. (2004). Modelling transposition latencies: Constraints for theories of serial order memory. J. Mem. Lang. *51*, 115–135.

Farrell, S., Oberauer, K., Greaves, M., Pasiecznik, K., Lewandowsky, S., and Jarrold, C. (2016). A test of interference versus decay in working memory: Varying distraction within lists in a complex span task. J. Mem. Lang. *90*, 66–87.

Fiebig, F., and Lansner, A. (2014). Memory consolidation from seconds to weeks: a three-stage neural network model with autonomous reinstatement dynamics. Front. Comput. Neurosci. *8*, 64.

Fiebig, F., and Lansner, A. (2017). A Spiking Working Memory Model Based on Hebbian Short-Term Potentiation. J. Neurosci. *37*, 83–96.

Fischer, M., Kaech, S., Knutti, D., and Matus, A. (1998). Rapid Actin-Based Plasticity in Dendritic Spines. Neuron *20*, 847–854.

Fries, P., Reynolds, J.H., Rorie, A.E., and Desimone, R. (2001). Modulation of Oscillatory Neuronal Synchronization by Selective Visual Attention. Science *291*, 1560–1563.

Funahashi, S. (2015). Functions of delay-period activity in the prefrontal cortex and mnemonic scotomas revisited. Front. Syst. Neurosci. *9*, 2.

Funahashi, S. (2017). Working Memory in the Prefrontal Cortex. Brain Sci. *7*.

Funahashi, S., Bruce, C.J., and Goldman-Rakic, P.S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. J. Neurophysiol. *61*, 331–349.

Fuster, J.M., and Alexander, G.E. (1971). Neuron activity related to short-term memory. Science *173*, 652–654.

Gayet, S., Paffen, C.L.E., and Van der Stigchel, S. (2017). Visual Working Memory Storage Recruits Sensory Processing Areas. Trends Cogn. Sci.

Gorgoraptis, N., Catalao, R.F.G., Bays, P.M., and Husain, M. (2011). Dynamic Updating of Working Memory Resources for Visual Objects. J. Neurosci. *31*, 8502–8511.

Gregoriou, G.G., Gotts, S.J., Zhou, H., and Desimone, R. (2009). High-frequency, long-range coupling between prefrontal and visual cortex during attention. Science *324*, 1207–1210.

Hansel, D., and Mato, G. (2013). Short-term plasticity explains irregular persistent activity in working memory tasks. J. Neurosci. Off. J. Soc. Neurosci. *33*, 133–149.

Harrison, S.A., and Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. Nature *458*, 632–635.

Hayden, B.Y., and Gallant, J.L. (2013). Working Memory and Decision Processes in Visual Area V4. Front. Neurosci. *7*.

Heinen, K., Feredoes, E., Ruff, C.C., and Driver, J. (2017). Functional connectivity between prefrontal and parietal cortex drives visuo-spatial attention shifts. Neuropsychologia *99*, 81–91.

Howard, M.W., and Kahana, M.J. (2002). A Distributed Representation of Temporal Context. J. Math. Psychol. *46*, 269–299.

Itti, L., and Koch, C. (2001). Computational modeling of visual attention. Nat. Rev. Neurosci. *2*, 194–203.

Jensen, O., Idiart, M.A., and Lisman, J.E. (1996). Physiologically realistic formation of autoassociative memory in networks with theta/gamma oscillations: role of fast NMDA channels. Learn. Mem. *3*, 243–256.

Jensen, O., Gelfand, J., Kounios, J., and Lisman, J.E. (2002). Oscillations in the alpha band (9–12 Hz) increase with memory load during retention in a short-term memory task. Cereb. Cortex *12*, 877–882.

Kamiński, J., Sullivan, S., Chung, J.M., Ross, I.B., Mamelak, A.N., and Rutishauser, U. (2017). Persistently active neurons in human medial frontal and medial temporal lobe support working memory. Nat. Neurosci. *20*, 590–601.

Konecky, R.O., Smith, M.A., and Olson, C.R. (2017). Monkey Prefrontal Neurons during Sternberg Task Performance: Full Contents of Working Memory or Most Recent Item? J. Neurophysiol. jn.00541.2016.

Kornblith, S., and Tsao, D.Y. (2017). How thoughts arise from sights: inferotemporal and prefrontal contributions to vision. Curr. Opin. Neurobiol. *46*, 208–218.

Lara, A.H., and Wallis, J.D. (2014). Executive control processes underlying multi-item working memory. Nat. Neurosci. *17*, 876–883.

LaRocque, J.J., Eichenbaum, A.S., Starrett, M.J., Rose, N.S., Emrich, S.M., and Postle, B.R. (2014). The short- and long-term fates of memory items retained outside the focus of attention. Mem. Cognit. *43*, 453–468.

Lavie, N., and Fockert, J.D. (2005). The role of working memory in attentional capture. Psychon. Bull. Rev. *12*, 669–674.

Lee, S.-H., and Baker, C.I. (2016). Multi-Voxel Decoding and the Topography of Maintained Information During Visual Working Memory. Front. Syst. Neurosci. 2.

Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., and Postle, B.R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. J. Cogn. Neurosci. *24*, 61–79.

Lisman, J.E., and Jensen, O. (2013). The theta-gamma neural code. Neuron *77*, 1002–1016.

Litwin-Kumar, A., and Doiron, B. (2014). Formation and maintenance of neuronal assemblies through synaptic plasticity. Nat. Commun. *5*, ncomms6319.

Luck, S.J., and Vogel, E.K. (1997). The capacity of visual working memory for features and conjunctions. Nature *390*, 279–280.

Malsburg, C. von der (1981). The Correlation Theory of Brain Function. In Models of Neural Networks, P.E. Domany, P.D.J.L. van Hemmen, and P.K. Schulten, eds. (Springer New York), pp. 95–119.

Matthey, L., Bays, P.M., and Dayan, P. (2015). A probabilistic palimpsest model of visual short-term memory. PLoS Comput. Biol. *11*, e1004003.

McElree, B. (2006). Accessing Recent Events. B.-P. of L. and Motivation, ed. (Academic Press), pp. 155–200.

McElree, B., and Dosher, B.A. (1989). Serial position and set size in short-term memory: The time course of recognition. J. Exp. Psychol. Gen. *118*, 346–373.

McLean, J.P., Broadbent, D.E., and Broadbent, M.H.P. (1983). Combining attributes in rapid serial visual presentation tasks. Q. J. Exp. Psychol. Sect. A *35*, 171–186.

Merrikhi, Y., Clark, K., Albarran, E., Parsa, M., Zirnsak, M., Moore, T., and Noudoost, B. (2017). Spatial working memory alters the efficacy of input to visual cortex. Nat. Commun. *8*, 15041.

Mi, Y., Katkov, M., and Tsodyks, M. (2017). Synaptic Correlates of Working Memory Capacity. Neuron *93*, 323–330.

Miller, E.K., Erickson, C.A., and Desimone, R. (1996). Neural Mechanisms of Visual Working Memory in Prefrontal Cortex of the Macaque. J. Neurosci. *16*, 5154–5167.

Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic theory of working memory. Science *319*, 1543–1546.

Moore, T., and Armstrong, K.M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. Nature *421*, 370–373.

Morey, C.C., Cowan, N., Morey, R.D., and Rouder, J.N. (2011). Flexible attention allocation to visual and auditory working memory tasks: manipulating reward induces a trade-off. Atten. Percept. Psychophys. *73*, 458–472.

Myers, N.E., Stokes, M.G., and Nobre, A.C. (2017). Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. Trends Cogn. Sci. *0*.

Norris, D. (2017). Short-Term Memory and Long-Term Memory are Still Different. Psychol. Bull. No Pagination Specified.

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. J. Exp. Psychol. Learn. Mem. Cogn. *28*, 411–421.

Oberauer, K., and Lewandowsky, S. (2014). Further evidence against decay in working memory. J. Mem. Lang. *73*, 15–30.

Olivers, C.N.L., Meijer, F., and Theeuwes, J. (2006). Feature-based memory-driven attentional capture: Visual working memory content affects visual attention. J. Exp. Psychol. Hum. Percept. Perform. *32*, 1243–1265.

Parthasarathy, A., Herikstad, R., Bong, J.H., Medina, F.S., Libedinsky, C., and Yen, S.-C. (2017). Mixed selectivity morphs population codes in prefrontal cortex. Nat. Neurosci.

Pearson, B., Raskevicius, J., Bays, P.M., Pertzov, Y., and Husain, M. (2014). Working memory retrieval as a decision process. J. Vis. *14*.

Pereira, J., and Wang, X.-J. (2015). A Tradeoff Between Accuracy and Flexibility in a Working Memory Circuit Endowed with Slow Feedback Mechanisms. Cereb. Cortex N. Y. N 1991 *25*, 3586–3601.

Pertzov, Y., Manohar, S., and Husain, M. (2016). Rapid Forgetting Results From Competition Over Time Between Items in Visual Working Memory. J. Exp. Psychol. Learn. Mem. Cogn.

Postle, B.R. (2016). How does the brain keep information “in mind”? Curr. Dir. Psychol. Sci. *25*, 151–156.

Postle, B.R., Ferrarelli, F., Hamidi, M., Feredoes, E., Massimini, M., Peterson, M., Alexander, A., and Tononi, G. (2006). Repetitive Transcranial Magnetic Stimulation Dissociates Working Memory Manipulation from Retention Functions in the Prefrontal, but not Posterior Parietal, Cortex. J. Cogn. Neurosci. *18*, 1712–1722.

Proctor, R.W., and Schneider, D.W. (2017). Hick’s Law for Choice Reaction Time: A Review. Q. J. Exp. Psychol. *0*, 1–56.

Reverberi, C., Görgen, K., and Haynes, J.-D. (2012). Compositionality of Rule Representations in Human Prefrontal Cortex. Cereb. Cortex *22*, 1237–1246.

Reynolds, J.H., and Desimone, R. (1999). The Role of Neural Mechanisms of Attention in Solving the Binding Problem. Neuron *24*, 19–29.

Rigotti, M., Barak, O., Warden, M.R., Wang, X.-J., Daw, N.D., Miller, E.K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature *497*, 585–590.

Rizzuto, D.S., and Kahana, M.J. (2001). An Autoassociative Neural Network Model of Paired-Associate Learning. Neural Comput. *13*, 2075–2092.

Romanski, L.M. (2004). Domain specificity in the primate prefrontal cortex. Cogn. Affect. Behav. Neurosci. *4*, 421–429.

Rose, N.S., LaRocque, J.J., Riggall, A.C., Gosseries, O., Starrett, M.J., Meyering, E.E., and Postle, B.R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. Science *354*, 1136–1139.

Rowe, J.B., Toni, I., Josephs, O., Frackowiak, R.S.J., and Passingham, R.E. (2000). The Prefrontal Cortex: Response Selection or Maintenance Within Working Memory? Science *288*, 1656–1660.

Sakai, K. (2008). Task Set and Prefrontal Cortex. Annu. Rev. Neurosci. *31*, 219–245.

Sakai, K., Rowe, J.B., and Passingham, R.E. (2002). Active maintenance in prefrontal area 46 creates distractor-resistant memory. Nat. Neurosci. *5*, 479–484.

Sala, J.B., and Courtney, S.M. (2007). Binding of What and Where During Working Memory Maintenance. Cortex *43*, 5–21.

Sandberg, A., Tegnér, J., and Lansner, A. (2003). A working memory model based on fast Hebbian learning. Netw. Bristol Engl. *14*, 789–802.

Schneegans, S., and Bays, P.M. (2017). Restoration of fMRI Decodability Does Not Imply Latent Working Memory States. J. Cogn. Neurosci. *29*, 1977–1994.

Schvaneveldt, R.W., and Chase, W.G. (1969). Sequential effects in choice reaction time. J. Exp. Psychol. *80*, 1.

Scolari, M., Seidl-Rathkopf, K.N., and Kastner, S. (2015). Functions of the human frontoparietal attention network: Evidence from neuroimaging. Curr. Opin. Behav. Sci. *1*, 32–39.

Silvanto, J. (2017). Working Memory Maintenance: Sustained Firing or Synaptic Mechanisms? Trends Cogn. Sci. *21*, 152–154.

Smyth, M.M. (1996). Serial Order in Spatial Immediate Memory. Q. J. Exp. Psychol. Sect. A *49*, 159–177.

Smyth, M.M., and Scholey, K.A. (1996). The relationship between articulation time and memory performance in verbal and visuospatial tasks. Br. J. Psychol. *87*, 179–191.

Solway, A., Murdock, B.B., and Kahana, M.J. (2012). Positional and temporal clustering in serial order memory. Mem. Cognit. *40*, 177–190.

Soto, D., Hodsoll, J., Rotshtein, P., and Humphreys, G.W. (2008). Automatic guidance of attention from working memory. Trends Cogn. Sci. *12*, 342–348.

Souza, A.S., and Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. Atten. Percept. Psychophys. 1–22.

Souza, A.S., Rerko, L., and Oberauer, K. (2016). Getting More From Visual Working Memory: Retro-Cues Enhance Retrieval and Protect From Visual Interference. J. Exp. Psychol. Hum. Percept. Perform.

Sprague, T.C., Ester, E.F., and Serences, J.T. (2016). Restoring Latent Visual Working Memory Representations in Human Cortex. Neuron *91*, 694–707.

Sreenivasan, K.K., Curtis, C.E., and D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. Trends Cogn. Sci. *18*, 82–89.

Stanton, P.K., and Sejnowski, T.J. (1989). Associative long-term depression in the hippocampus induced by hebbian covariance. Nature *339*, 215–218.

Stokes, M.G. (2015). 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn. Sci. *19*, 394–405.

Szczepanski, S.M., Pinsk, M.A., Douglas, M.M., Kastner, S., and Saalmann, Y.B. (2013). Functional and structural architecture of the human dorsal frontoparietal attention network. Proc. Natl. Acad. Sci. *110*, 15806–15811.

Thompson-Schill, S.L., Jonides, J., Marshuetz, C., Smith, E.E., D'Esposito, M., Kan, I.P., Knight, R.T., and Swick, D. (2002). Effects of frontal lobe damage on interference effects in working memory. Cogn. Affect. Behav. Neurosci. *2*, 109–120.

Treisman, A.M., and Gelade, G. (1980). A feature-integration theory of attention. Cognit. Psychol. *12*, 97–136.

Tsodyks, M.V., and Markram, H. (1997). The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. Proc. Natl. Acad. Sci. *94*, 719–723.

Wei, Z., Wang, X.-J., and Wang, D.-H. (2012). From Distributed Resources to Limited Slots in Multiple-Item Working Memory: A Spiking Network Model with Normalization. J. Neurosci. *32*, 11228–11240.

Wimmer, K., Nykamp, D.Q., Constantinidis, C., and Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat. Neurosci. *17*, 431–439.

Wolff, M.J., Jochim, J., Akyürek, E.G., and Stokes, M.G. (2017). Dynamic hidden states underlying working-memory-guided behavior. Nat. Neurosci.

Woodman, G.F., Luck, S.J., and Schall, J.D. (2007). The Role of Working Memory Representations in the Control of Attention. Cereb. Cortex *17*, i118–i124.

Xu, Y. (2017). Reevaluating the Sensory Account of Visual Working Memory Storage. Trends Cogn. Sci. *21*, 794–815.

Zanto, T.P., Rubens, M.T., Thangavel, A., and Gazzaley, A. (2011). Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. Nat. Neurosci. *14*, 656–661.

Zenke, F., Agnes, E.J., and Gerstner, W. (2015). Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. Nat. Commun. *6*, 6922.

Zhang, W., and Luck, S.J. (2008). Discrete fixed-resolution representations in visual working memory. Nature *453*, 233–235.

Zipser, D., Kehoe, B., Littlewort, G., and Fuster, J. (1993). A spiking network model of short-term active memory. J. Neurosci. Off. J. Soc. Neurosci. *13*, 3406–3420.

Zokaei, N., Manohar, S., Husain, M., and Feredoes, E. (2014a). Causal evidence for a privileged working memory state in early visual cortex. J. Neurosci. *34*, 158–162.

Zokaei, N., Ning, S., Manohar, S., Feredoes, E., and Husain, M. (2014b). Flexibility of representational states in working memory. Name Front. Hum. Neurosci. *8*, 853.

Zucker, R.S. (1989). Short-Term Synaptic Plasticity. Annu. Rev. Neurosci. *12*, 13–31.

Supplementary Materials:

Methods

Figures S1-S9

Table S1-S2

Movie S1

References

Methods

The present model considers a minimal arrangement for three feature dimensions, each with four possible feature values, allowing 12 features to be encoded. Each feature unit receives input when a particular feature is present in the stimulus. Four conjunctive units are fully connected reciprocally to the 12 feature units. These connections are all excitatory, and initialized to be random. Four conjunctive neurons is the minimum possible number that can bind information from four objects. The fully-connected network therefore required 48 weights to conjunctions from features ($\mathbf{W}^{cf}$) and 48 to features from conjunctions ($\mathbf{W}^{fc}$). The activity of all units **f** and **c** were initialized to zero at the start of simulation, and weights $\mathbf{W}^{fc}$ and $\mathbf{W}^{cf}$ were randomly assigned from a uniform distribution over the interval [0, 1]. In its simplest form, the activity update equation was:

$$\begin{bmatrix}\mathbf{c}\\\mathbf{f}\end{bmatrix} \leftarrow \mathrm{sigmoid}\left(\begin{bmatrix}\mathbf{W}_{cc} & \mathbf{W}_{cf}\\\mathbf{W}_{fc} & \mathbf{W}_{ff}\end{bmatrix}\begin{bmatrix}\mathbf{c}\\\mathbf{f}\end{bmatrix} + \begin{bmatrix}\mathbf{0}\\\mathbf{input}\end{bmatrix}\right)$$

The conjunctive-feature synapses $\mathbf{W}^{cf}$ and $\mathbf{W}^{fc}$ were updated by a Hebbian covariance rule, whereas the fixed inter-conjunction ($\mathbf{W}^{cc}$) and inter-feature ($\mathbf{W}^{ff}$) synapses each comprised two components:

1) blanket lateral inhibition between conjunction neurons or between features within the same dimension, to implement competition, and
2) self-excitation, so that firing does not stop suddenly when external input is removed, but rather decays exponentially with time.

So the update equations can be written out more fully as:

neuron activation:

$$\begin{aligned}\mathbf{c} &\leftarrow \sigma(\beta + (\alpha_1 \mathbf{1} \qquad + \alpha_2 \mathbf{I})(\mathbf{c} - \beta) + \alpha_3 \mathbf{W}^{cf}(\mathbf{f} - \beta) + \varepsilon \cdot \mathcal{N})\\ \mathbf{f} &\leftarrow \sigma(\beta + (\alpha_4 \mathbf{W}^{ff} + \alpha_4 \mathbf{I})(\mathbf{f} - \beta) + \alpha_6 \mathbf{W}^{fc}(\mathbf{c} - \beta) + \mathbf{i})\end{aligned}$$

synaptic updates

$$\begin{aligned}\Delta &= (\mathbf{c} - \beta) \cdot (\mathbf{f}^T - \beta)\\ \mathbf{W}^{cf} &\leftarrow \sigma(\mathbf{W}^{cf} + \gamma_1 \Delta)\\ \mathbf{W}^{fc} &\leftarrow \sigma(\mathbf{W}^{fc} + \gamma_2 \Delta^T)\end{aligned}$$

constants:

$$\begin{aligned}\alpha_i &= \{-0.28, 1.03, 0.05,\ -0.28, 0.75, 0.05\}\\ \beta &= 0.175\\ \gamma &= \{0.02, 0.02\}\\ \varepsilon &= 0.005\\ \mathbf{W}^{ff} &= \begin{cases} 1, & \text{if } i \neq j, \text{ but they are in the same feature dimension}\\ 0, & \text{otherwise}\end{cases}\\ \sigma(x) &= \begin{cases} < 0 : & 0\\ 0 \leqslant x \leqslant 1 : & x\\ > 1 : & 1\end{cases}\\ NC &= 4\\ NF &= 12\end{aligned}$$

In these equations, **c** and **f** are the activities of conjunctive neurons and feature neurons, and **i** is the external input. The six synaptic free parameters α were:

α_1 , α_4 : mutual lateral inhibition between neurons

α_2 , α_5 : self-excitation or temporal decay, and

α_3 , α_6 : synaptic gain for the conjunction-to-feature and feature-to-conjunction synapses, for the conjunctive and feature neurons respectively. β is the baseline neuron activity, γ_i are the learning rates, and ε is the amount of noise added to the conjunctive units. The function σ constrains values lie between 0 and 1 and, for simplicity, was chosen as: $\sigma(y)=\min(1, \max(0, y))$. The identity matrix **I** produces self-excitatory synapses, and a matrix of ones (**1**) produces lateral inhibition. In simulations, the 12 features were arranged into 3 dimensions (color, orientation and location), so that the feature-to-feature inhibition ($\mathbf{W}^{ff}$) was arranged in 3 blocks of 4 units. *N* indicates a gaussian white noise vector with s.d.=1. We constrained learning rates to be identical in both directions ($\gamma_1=\gamma_2$), and noise was in fact not required in order to obtain the typical patterns of errors reported here (ε=0), giving effectively 8 free parameters. A minimal algorithm to reproduce the main result is provided at the end of this section.

Implementation of stimuli and simulations

Model equations were simulated in MATLAB (code available at [www.smanohar.com/wp/wm/download.html]). A web-based interactive simulation can be accessed at http://www.smanohar.com/wp/wm. Each run simulated 200 trials. To test different hypotheses, the simulation setup was varied from a canonical setup. The canonical simulation resembled common working memory experiments (Bays and Husain, 2008). A series of objects was presented, followed by a memory probe that triggers recall. Each object excited one feature in each dimension (colour, orientation and location). The features of the objects presented on each trial were random, with the constraint that no two objects within one trial shared any features. Each trial started with a foreperiod for equilibration lasting 200 time steps, with the features inhibited (input **i** = –1). Then during the encoding epoch, each memory item was presented by activating the corresponding feature units, by maximally activating the features present in the object (i = +1) and inactivating the absent features (i = –1). Stimuli were presented for 120 time steps, followed by an inter-stimulus interval of 50 time steps with no input (**i**=0). After the last item was presented, the memory delay period of 240 time steps followed, with no input, mimicking a retention interval. Then at the probe epoch, the feature acting as the retrieval cue was activated (i=+1), and the other features were inhibited (i = –1). The probe lasted 120 time steps. Finally, the response interval constituted 220 time steps with no input. This permitted re-activation of the features to be recalled. The feature that reached the highest level of activity during this period was selected as the response. In the case of an exact tie-breaker the decision was randomized. An example of this sequence of events shown in **Video S3**.

Parameter selection was performed initially by trial and error to achieve the hypothesized dynamics. First, the decay, inhibition and input weightings α_i for the conjunctive and feature neurons were adjusted to create a sustained activity plateau and to ensure that conjunctive neurons competed in a winner-takes-all manner. Then the learning rates γ_i were adjusted to allow sufficiently rapid weight changes such that over 50 time steps, a trace remained that represented which combination of feature units had been active. The baseline was approximately 20% maximal to permit deactivation of neurons, so that conjunction units that lose the competition will 'unlearn' their associations (Stanton and Sejnowski, 1989). A small amount of noise was added to ensure conjunction units were never identical, to facilitate symmetry-breaking.

Certain effects were more or less sensitive to changes in model parameters. Overall accuracy could be made to range from consistent 100% to consistent 0%, depending on choices of α_i, β, γ_i and ε. We aimed for 70% overall accuracy, allowing a wide dynamic range of performance to examine the predicted effects, so this was the regime in which the main simulations were run.

Simulation 1: set size

Simulation parameters were as above, with $\alpha_i = \{ -0.28, 1.03, 0.05, \ -0.28, 0.75, 0.05 \}$; β=0.175, $\gamma = \{0.02, 0.02\}$; $\varepsilon = 0.005$. Timings were: foreperiod=200 steps, objects=120 steps, inter-stimulus=50 steps, retention delay=240 steps, probe=120 steps, recall=240 steps. Trials proceeded as above, with four items presented sequentially. One, two, three and four items were presented in different trials, and each serial position was probed. For multi-item sequences, each item in the sequence was probed equally often. This gave ten (1+2+3+4) trial types, and 200 trials of each type were simulated. The average accuracy over all serial positions was calculated for each set size (**Fig.2B**). Error bars are the standard error of the accuracy when trials were broken into subsets of 20 trials each. Simulations were performed both using interleaved trials, and also with one condition per block, and comparable results were obtained with both methods.

Capacity limits were naturally limited to four in this simulation because only four conjunctive neurons were present. However the capacity limit does not *directly* relate to the number of conjunctive neurons, but rather, to the *proportion* of conjunctive neurons that will become simultaneously active during winner-takes-all competition (¼ in this case). This is in turn determined by the level of inhibition. A simulation using 8 conjunctive neurons also reproduced the set-size effects, with inhibition tuned to allow two neurons to be active at once.

Note that absolute modelled accuracy was lower than in the empirical data, because participants made same/different judgements (chance=50%) whereas our simulations recalled actual features (chance=25%).

Simulation 2: serial position

The run was identical to simulation 1 above. Trials were grouped according to the serial position of the probed item, and by set size (**Fig.2D**). Accuracy was calculated as per simulation 1. Note that in all simulations, feature unit activity was held at zero at the start of each trial (**i** = –1) to simulate a foreperiod or inter-trial interval. This permitted the first item of the sequence to be encoded faster. We then explored different parameter sets, varying α, β, γ and ε. The recency effect was robust across a wide range of parameters, and was often strong enough to push performance to 100% for the last item. The presence of a primacy effect was more strongly dependent on the specific presentation timings and on the choice of α_i . This matches empirical studies of visual WM, which do not always find primacy effects.

Simulation 3 – encoding duration:

The time that each item was presented for was varied. This ranged from 10 to 120 time steps (in increments of 20 ms), with all items being presented for the same duration. The inter-object duration was fixed at 50 ms as in previous simulations. The average accuracy was collapsed across all serial positions for each set size, and plotted as a function of encoding duration (**Fig.2F**). The simulation demonstrates an interaction between set size and time, such that when more items are stored, the decay is faster. In order to avoid floor effects for this simulation, where accuracy in the 4-item condition could approach chance very quickly as the encoding duration decreases, we increased the 'stickiness' of conjunction units in the model, by reducing the decay factor α_1 and uptitrating their input gain from each other (α_2) and from the feature units (α_3). We therefore set

$$\alpha = \{ -0.5, 1, 0.08, \quad -0.28, 0.7, 0.05\} \quad \text{and} \quad \beta = 0.2$$

for this simulation, and the consequence was that the overall accuracy of the model increased from 75% to around 90% while preserving the set size and serial position effects. This 'higher performance' regime was used for simulations 3, 4 and 5. Other than improved overall accuracy, these parameters produced qualitatively similar effects to the primary simulations (**Fig.S9**).

Simulation 4: effect of memory delay

We varied the number of time steps after the final item was presented, until the onset of the probe. The delay varied between 200 and 1800 steps. For each duration, simulations of 200 trials were run for 1, 2, 3 and 4 items, with each possible position being probed (i.e. total 2,000 trials per duration). In order to avoid floor effects, where accuracy in the 4-item condition could approach chance as the delay increases, we used the same regime as Simulation 3. Data is plotted as a function of the delay (**Fig 2H**).

Simulation 5: transposition errors

Here we studied the tendency to incorrectly report features from items temporally adjacent to the probed item. This arises because occasionally, the same conjunctive unit is activated for two consecutive objects, when the second object fails to sufficiently drive a different conjunction unit. In this case, features of two consecutive objects will be confused.

Data were taken from the standard 4-item condition where four items were presented and one is probed, using parameters of Simulation 3. For this simulation, trials were grouped according to serial position of the probed item. Four responses were possible on each trial, and for each trial, we take the serial position at which the reported feature *actually* appeared on that trial (**Fig.S3B**). This figure shows a histogram of the model's responses. Over 2000 trials, the probability of making each of the four responses was calculated. Each line shows the probability of reporting the orientation of the four items presented in the sequence (x-axis), when a particular serial position was probed (each as a different line). We show logarithms of the mean error rate as in (Farrell and Lewandowsky, 2004), and added a small offset of 10^{-3} since some runs had zero errors. Error bars are standard error of the logarithm of mean error rate. The four possible responses are aligned such that the correct response appears at position zero, at the center of the graph (i.e. the *probed* item's orientation). Sometimes the model erroneously reports an item previous to the one probed (x<0) or an item later than the item probed (x>0). The mean of the four probe conditions is shown in red.

Simulation 6: Incidental retrocue

In the empirical study, the primary task involved recalling the orientation of the item with a given color, and the secondary task ("incidental cueing") required participants to report the location of the item with a given color, which could be congruent or incongruent to the ultimately-probed item. We simulated the empirical task (**Fig.3D**) by presenting two items, as per the 2-item condition in simulation 3. After the items were presented, a 120 time-steps retention period followed, then the probe feature (incidental cue) for one of the two items was activated for 40 time steps ("IC" in **Fig. 3E**). After a further 120 time steps corresponding to recall of the third dimension for this item, a further memory delay of 120 time-steps was included. Then the final memory probe was activated, and recall of the second dimension was measured as previously. This final probe could either be the same item (congruent), or the other item (incongruent), as the one cued for the first response. 200 trials were simulated for each condition. Accuracy was plotted for the valid and invalid incidental cue conditions (**Fig. 3F**).

Simulation 7: Reaction time (RT)

RT was calculated by finding the time at which the winning feature reached its maximal value in the period after the probe. The time to reach 98% of maximum was used, rather than using an absolute threshold, because the final stable value of an activated feature differed when different model parameters were used. Using an arbitrary fixed threshold or the rate of rise yielded qualitatively similar but less consistent RT effects. To examine basic set size and serial position effects, the same trials were used as in simulation 1. Mean RT with standard error is shown (**Fig.S1A**), in comparison to data (McElree and Dosher, 1989) (**Fig.S1B**). RT was generally inversely related to accuracy.

Simulation 8: probe interference:

Many working memory tasks have asked participants to adjust features of the probe to match the remembered features (Zhang and Luck, 2008). In these experiments, the probe contains a feature that is irrelevant to recalling the item. For example, if participants must report the orientation of a bar with a given color, then using a colored bar as a probe introduces an orientation feature that conflicts with the remembered item. The model predicts this will interfere with re-focusing the item (**Fig.S2**). To simulate this, sequences of 1 to 3 items were presented to the model, using an identical setup to simulation 1. However at the time of probe, two features were activated; one in the probe dimension, and one in the recall dimension. As previously, the probe feature input was +1, and other features on that dimension were -1. However an additional input +1 was added for one orientation feature. The additional feature, in the recall dimension, was one that had *not* been presented on that trial. Since only 4 features were present in this model, this constraint meant that we could only test set sizes of one to three items.

Performance was worse when the probe contained an interfering item. There is some empirical support that this might indeed be the case (Souza et al., 2016). The simulation predicts interference across all set sizes. Conversely, we can also predict an improvement in performance for probes containing an additional *helpful* feature, for example if participants must report an object's orientation given both its color *and* its location.

Simulation 9: Weak TMS pulse reactivates delay period decoding

Here we enquired whether the model could reproduce the phenomena where unattended items, which are not normally decodable from brain activity, could be brought back into a temporarily decodable state by applying a pulse of activation. Empirically this has been demonstrated using a nonspecific high-energy visual stimulus pattern (Wolff et al., 2017), or by applying a TMS pulse to sensory cortex (Rose et al., 2016). In the simulation, two items were presented sequentially to the model, exactly as in simulation 1. During the delay period, the feature neurons received a flat high-valued input (**i**=+1) for 10 time steps. This was compared to an identical condition with no stimulus. The delay period therefore comprised wither 120 timesteps + 10 timestep pulse + 120 time steps, or in the no-stimulus condition, 250 timesteps (**Fig 4A**).

2000 trials were simulated, and half the trials were used to construct a linear classifier that predicts the identity of each of the two items presented on that trial. The two classifiers were tested on the remaining trials, to give the decoding accuracy. Decoding was performed across trials at each time point independently, indicating the degree to which neural activity in the feature units predicted the identity of each item. The first object presented was termed the 'unattended' object, since during the delay, attention was focused on the second (final) object. Decodability of the unattended item was transiently restored after the pulse (**Fig.4B** dark blue trace), reproducing the phenomenon in the TMS study (**Fig.4C**) and using a high-energy neutral visual stimulus.

Two further predictions are that 1) stimulation of prefrontal regions should have similar effects, and 2) selectively stimulating specific feature neurons e.g. in sensory cortex will have stronger effects when those features are part of an item currently in memory – i.e. when the synapses from that feature neuron to the conjunctive neurons are already strong.

Simulation 10: Strong TMS pulse disrupts focus of attention

A strong TMS pulse to sensory cortex (MT+) has been shown to disrupt the focus of attention. To reproduce this, we used precisely the same simulation as in 9, but increased the pulse duration to 20 timesteps duration. We compared accuracy for probing the first item, vs the second item that was presented, with and without pulse. The presence of the pulse reduced accuracy for the second item, but paradoxically improved memory for the first item (**Fig.4D**). This is because the pulse indiscriminately re-activated the feature and conjunction neurons, disrupting the focus of attention but not the underlying memory traces. Indeed after the pulse, the attractor state sometimes shifted back to the first item. Precisely this phenomenon was observed in a TMS study (Zokaei et al., 2014a) (**Fig.4E**). The model predicts the same effects if prefrontal neurons are stimulated.

Simulation 11: Delay-period decoding

Three items were presented in sequence to the model. This was identical to simulation 1 except that the delay period duration between items was increased to 100 time steps, to test decoding during the stable attractor state. Each serial position was probed on 2000 trials. As in simulations 9 and 10 above, half the trials were used to construct linear classifiers that predict the identity of the item that was shown at each given serial position on a trial. The three classifiers were then tested on the remaining trials, to give the decoding accuracy. Decoding at each moment in a trial indicates the degree to which neural activity in the feature units predicted the identity of the item presented at each serial position (**Fig.S4A**). We then asked 6 questions, as in (Konecky et al., 2017): can the first item be decoded in the first delay, in the second delay, or in the third delay; can the second item be decoded in the second or third delay; can the last item be decoded in the third delay (**Fig.S4B**)? The feature neurons strongly encoded the most-recently-presented item. Using the activity of conjunction units, however, nothing could be decoded above chance using a linear classifier.

Simulation 12: Previous trial repetition effect

This effect was examined by using all trials taken from simulation 1. We compared trials in which the probed-item's features on the probe dimension and the recall dimension were the same or different to the probed item of the previous trial. Trials were grouped according to whether the previous trial's probe color was the same as the current trial's probe color, and also whether the probed-item's orientation was the same or different to the previous trial's probed-item orientation (**Fig.S4**). When the identical item was probed, recall was more accurate.

Simulation 13: Pattern similarity after intervening conjunctions

We ran 2000 trials of the 1-item memory condition, and examined delay-period activity in conjunction units. If neurons have classical receptive fields, then when the same stimulus is presented as on a previous trial, the activity pattern will be similar, whereas if the stimulus is different, the patterns will be dissimilar. Two major predictions of flexibly-conjunctive neurons is that the pattern similarity will decrease both with the number of intervening trials (**Fig.5A**), and also when intervening stimuli form different conjunctions with the same features (**Fig.5B-D**).

The similarity between representations on trial *n* and trials *n*-2, *n*-3 up to *n*-10 were examined, in the middle 100 timesteps of the delay period. For each inter-trial distance, trials were divided according to whether the same or different stimulus was shown on those two trials (**Fig.5A**). We considered only 2 feature dimensions in this analysis, so that the each pair of trials had the same stimulus 25% of the time. For *n*-1 to *n*-4, it was possible to further subdivide trials according to the stimuli presented on intervening trials. A “violation” of the conjunction was defined as an intervening stimulus which is similar on one feature dimension but dissimilar on the other feature dimension, to trial *n*. No violation occurs if the intervening stimulus is the same as on trial *n*, or if it contains no features in common with trial *n*. Trials were split according to the number of these intervening violations.

Simulation 14: Implementing simple task rules

To implement execution of actions, we re-labelled the ‘orientation’ dimension as ‘action’, so that there were 4 possible motor actions, which could be coupled to the 4 possible colors (**Fig.S6**). To provide the model with instructions for the task rules, each stimulus-response (S-R) mapping was activated, one at a time. For example, to provide the instruction “red means press button 1”, we activated the “red” colour unit and the “button 1” action unit simultaneously. Analogous to working memory encoding, a conjunction unit became associated with each pairing. In each block, we presented either 1, 2, 3 or 4 rules. After presenting the rules, 12 color cues randomly drawn from those rules were shown sequentially, analogous to memory probes. After each cue, activation of the motor units was measured. We tested the responses to one, two, three and four simultaneous S-R mappings.

The sequence of events was thus very similar to the working memory task, and reaction times and accuracy were calculated in the same way as before (**Fig.S6A**). Two differences in implementation were needed to permit appropriate action selection over many trials. First, we lengthened encoding (240 steps) and shortened the probe duration (80 steps). This increased the stability of the mappings. Second, rather than inhibiting the features during the inter-trial interval, no input was provided (**i** = **0** rather than **-1**), which allowed the units to maintain their ongoing activity. The attentional focus thus remained active throughout the experiment. Without these two changes, interference led to forgetting of the task rules over the first 5 to 10 trials.

Two important empirically observed effects in choice reaction time experiments are Hick's law – the increase in RT with log(number of options) – and the repetition effect. To examine Hick's law, we calculated the mean RT on correct trials, in blocks where there were 1 to 4 rules presented (**Fig.S6B**). To study the effect of stimulus-response repetition on consecutive trials, trials in the 2-rule blocks were categorized according to whether the same stimulus was present on the previous 0, 1 or 2 trials (**Fig.S6C**), and the mean RT was calculated separately for correct and incorrect response trials. Data for correct trials from (Schvaneveldt and Chase, 1969) were re-plotted next to simulations. Decoding from conjunctive units was performed as previously done for feature units, as a function of time. For each trial the classifier was trained on other trials in the same block (**Fig.S6D**). For comparison, accuracy when the classifier was trained on the previous block was also measured. During WM tasks no decoding was possible from conjunctive units. But in this simple stimulus-response task, decoding was possible across a block of trials which shared the same rule. The presentation of new task rules at the start of a block effectively overwrites the conjunctive neurons' weights, precluding decoding.

Simulation 15: Removing the third (task-irrelevant) feature dimension

In previous simulations, each object consisted of three features, but one of the features is task-irrelevant. Simulation 3 was run with standard conditions and timings but on half of trials, the third dimension features were treated as absent (**i**=-1) for every object. We contrasted recall accuracy for conditions where objects consisted of all 3 features vs. only 2 features (**Fig.S7**). The primacy and recency effects were observed to be smaller.

Simulation 16: Exploring the parameter space

We examined 9 free parameters in the model: the baseline activity, lateral inhibition x 2 (for conjunction and feature units), self-excitation x 2, reciprocal excitation x 2, and learning rates x 2 (**Fig.S8**). For each combination of parameters, 50 trials per condition were simulated for set-sizes 1 to 4, for all serial positions (500 trials per parameter set). The canonical model was perturbed along two parameters at a time. For each pair of parameters, a range of values above and below the canonical parameter values were tested, with 10 linearly-spaced levels of each parameter, to give a 10x10 grid. The parameter values used in Simulation 1 thus lie at the center of each grid, and the figure represents all possible cardinal planes through that point in a 9-dimensional hyperspace.

Each simulation resulted in a serial position curve like Fig.2D, from which we could quantify the set size effect (linear slope of accuracy as set size varies, collapsed across serial position), primacy effect (difference in accuracy between final and penultimate items in sequence, averaged across set sizes 2-4) and recency effect (difference in accuracy between first and second items in sequence, averaged across set sizes 2-4). The size of each effect was smoothed using a 3x3 boxcar, and is portrayed by pixel color in the 10x10 grids.

Minimal algorithm to reproduce main results

```
$\alpha$ ← [ -0.28, 1.03, 0.05,    -0.28, 0.75, 0.05  ]
$\beta$ ← 0.175
$\gamma$ ← 0.02
$\mathbf{W}^{ff}$ ← $\alpha_5$ eye(12)  + $\alpha_4$ [
    1  1  1  1  0  0  0  0  0  0  0  0
    1  1  1  1  0  0  0  0  0  0  0  0
    1  1  1  1  0  0  0  0  0  0  0  0
    1  1  1  1  0  0  0  0  0  0  0  0
    0  0  0  0  1  1  1  1  0  0  0  0
    0  0  0  0  1  1  1  1  0  0  0  0
    0  0  0  0  1  1  1  1  0  0  0  0
    0  0  0  0  1  1  1  1  0  0  0  0
    0  0  0  0  0  0  0  0  1  1  1  1
    0  0  0  0  0  0  0  0  1  1  1  1
    0  0  0  0  0  0  0  0  1  1  1  1
    0  0  0  0  0  0  0  0  1  1  1  1
]
$\mathbf{W}^{cc}$ ←  $\alpha_2$ eye(4) + $\alpha_1$ ones(4)

$\mathbf{c}$ ← [0  0  0  0]$^T$
$\mathbf{f}$ ← [0  0  0  0   0  0  0  0   0  0  0  0]$^T$
$\mathbf{W}$ ← rand(12,4)                   // as $\gamma_1$ == $\gamma_2$, $W^{fc}$ == $W^{cf\ T}$
$t$ ← 0
while $t$ < stimulus.length {
     $\mathbf{f}$ ← $\beta$ + $\mathbf{W}^{ff}$ ($\mathbf{f}$ − $\beta$) + $\alpha_3$ $\mathbf{W}$ ($\mathbf{c}$ − $\beta$) + stimulus$_t$
     $\mathbf{f}$ ← max(0,min(1,$\mathbf{f}$))
     $\mathbf{c}$ ← $\beta$ + $\mathbf{W}^{cc}$ ($\mathbf{c}$ − $\beta$) + $\alpha_6$ $\mathbf{W}^T$ ($\mathbf{f}$ − $\beta$) + 0.005 * randn(4,1)
     $\mathbf{c}$ ← max(0,min(1,$\mathbf{c}$))
     $\Delta$ ← ($\mathbf{f}$ − $\beta$)($\mathbf{c}$ − $\beta$)$^T$
     $\mathbf{W}$ ← max(0,min(1, $\mathbf{W}$ + $\gamma$ $\Delta$))
}
choice ← findmax( $\mathbf{f}$[4:8] )

stimulus$_t$ ← {
    200 × [-1 -1 -1 -1   -1 -1 -1 -1   -1 -1 -1 -1]  // foreperiod
    120 × [+1 -1 -1 -1   +1 -1 -1 -1   +1 -1 -1 -1]  // obj1
     50 × [ 0  0  0  0    0  0  0  0    0  0  0  0]
    120 × [-1 +1 -1 -1   -1 +1 -1 -1   -1 +1 -1 -1]  // obj2
    300 × [ 0  0  0  0    0  0  0  0    0  0  0  0]  // delay
    120 × [+1 -1 -1 -1   -1 -1 -1 -1   -1 -1 -1 -1]  // probe
    220 × [ 0  0  0  0    0  0  0  0    0  0  0  0]
}
```

Supplementary Figures

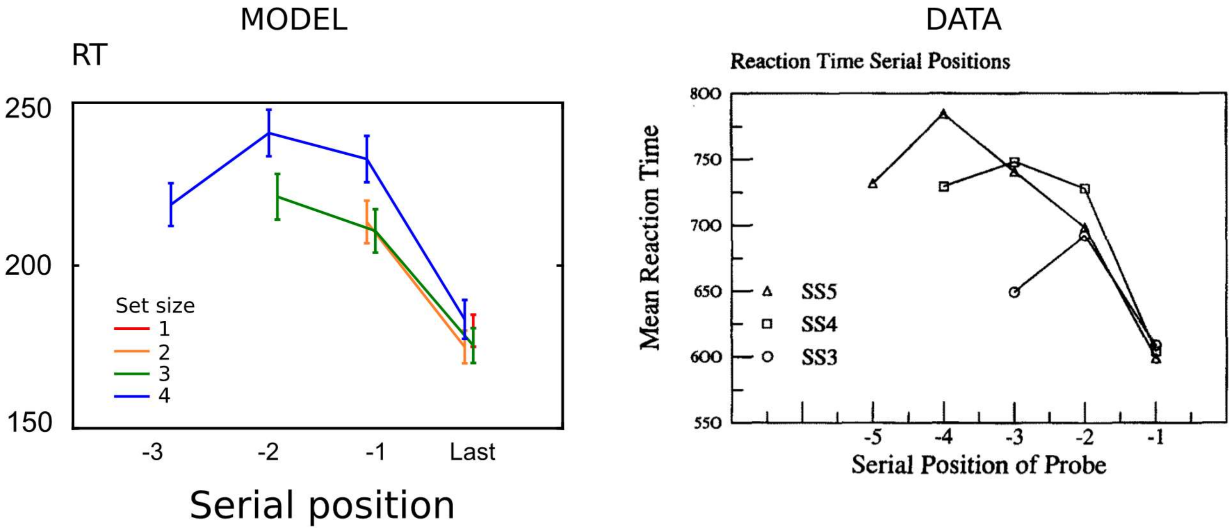


Figure S1: Reaction times (Simulation 7)

The time taken to recall an item can be indexed by how fast the winning feature rises after probe presentation. An inverse relation with accuracy is observed, in keeping with behavior, with items already in the focus of attention showing the fastest responses. (**Simulation 8** below)
A) Reaction times were extracted from trials of simulation 1, quantified as the time after the probe at which the winning feature reached 95% of its maximum value. The faster the attractor settled into a stable winning state, the shorter the reaction time. Each line represents a different set size, and the X-axis indicates serial position of the probed item within the memory set. Reaction times were approximately inversely related to accuracy on each condition.
B) This is consistent with race-models of WM recall (Pearson et al., 2014), and aligns with behavioural data (McElree, 2006; McElree and Dosher, 1989)(Figure adapted from McElree & Dosher 1989).

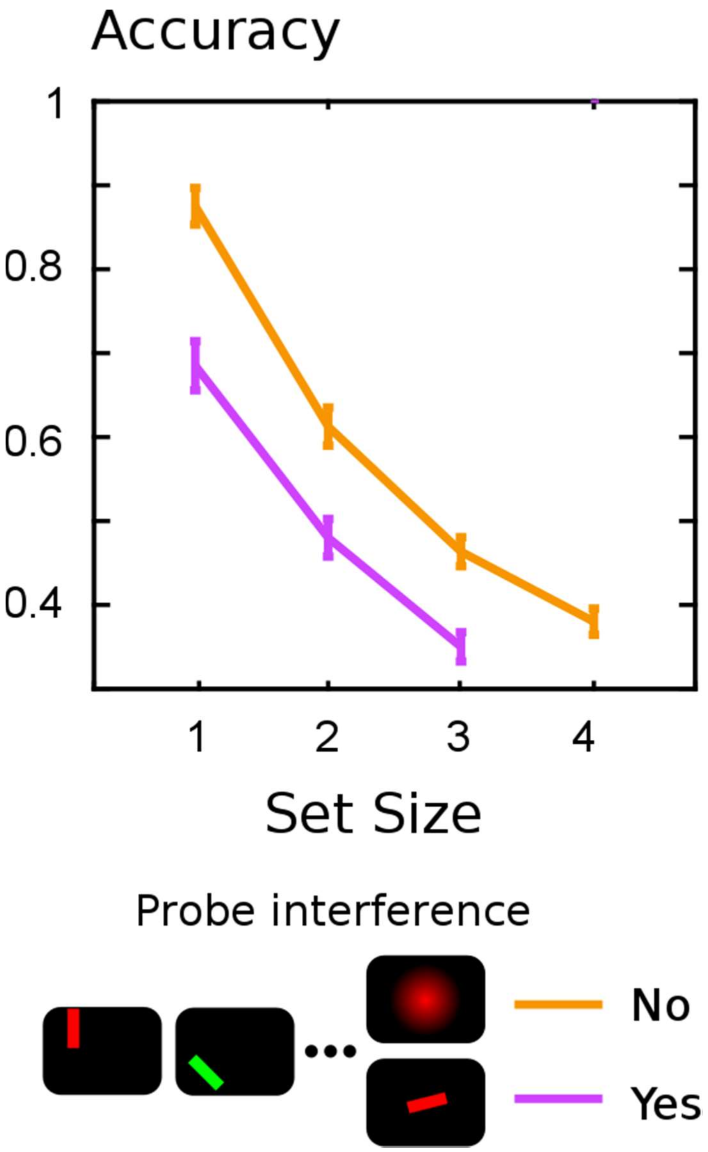


Figure S2: Novel prediction – Probe interference (Simulation 8)
The model predicts that the probe can itself interfere with recall if it contains irrelevant features. The irrelevant features compete with the probe feature, and reduce the probability of correct re-focusing of the item.

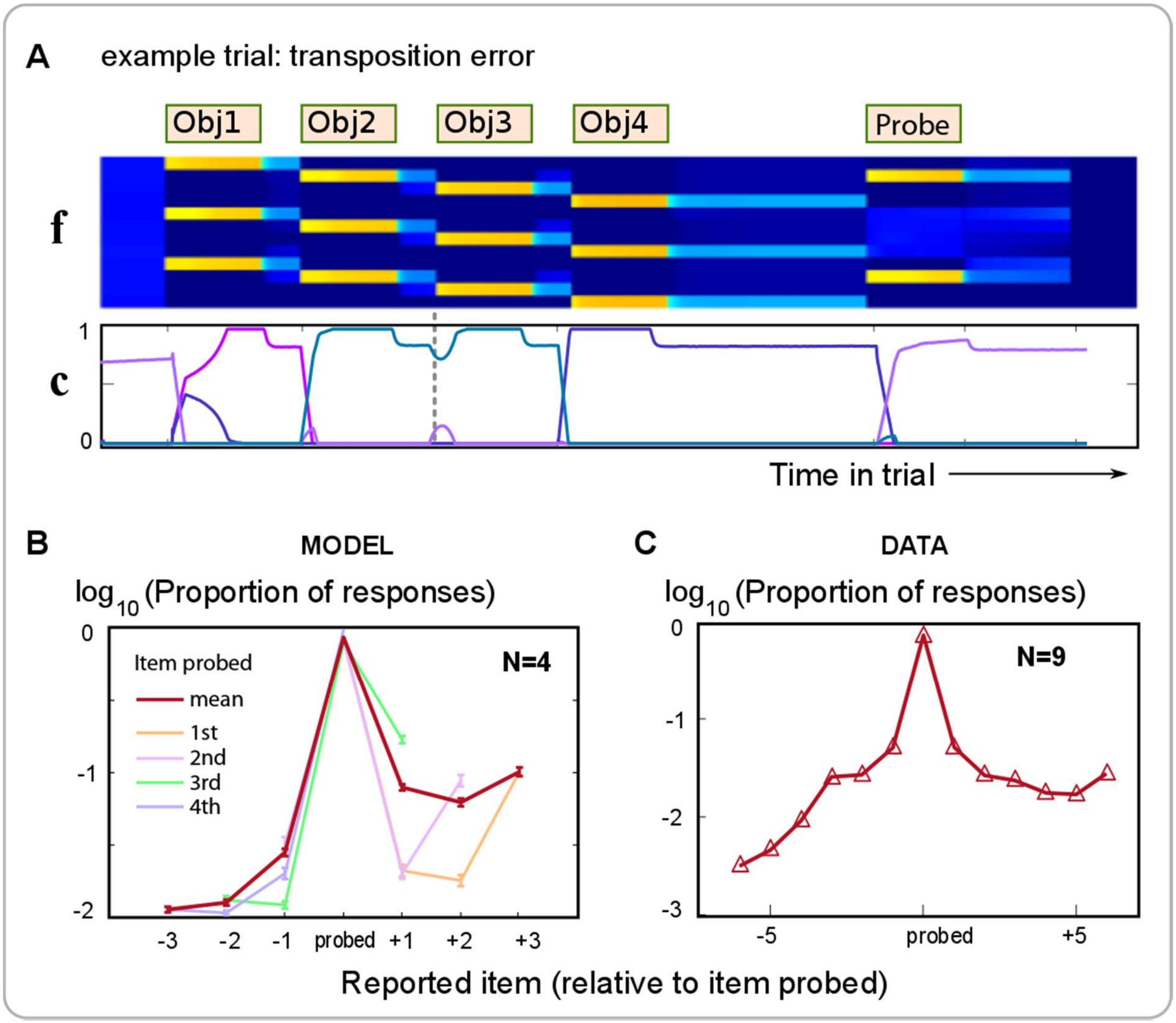


Figure S3: Errors are more likely to be sequentially neighbouring items (Simulation 5)

A Transposition errors (sequence positional errors, or swap errors) arise when features of a temporally adjacent item are instead reported. In this example trial, the conjunction neuron activated by object 2 remained active for object 3 (vertical dashed line), so when item 2 or 3 is probed, the incorrect response feature is sometimes activated.
B Items that were either just before (x<0), or just after (x>0), the probed item tended to be reported more often.
C A qualitatively similar pattern of errors is observed in a verbal WM task in which a list of words had to be recalled in order (adapted from Expt.3 of Farrell and Lewandowsky, 2004). A very similar falloff with inter-item distance is also seen in visuospatial memory using sequence change-detection (Smyth and Scholey, 1996), but their data did not permit distinguishing forward and backward transpositions.

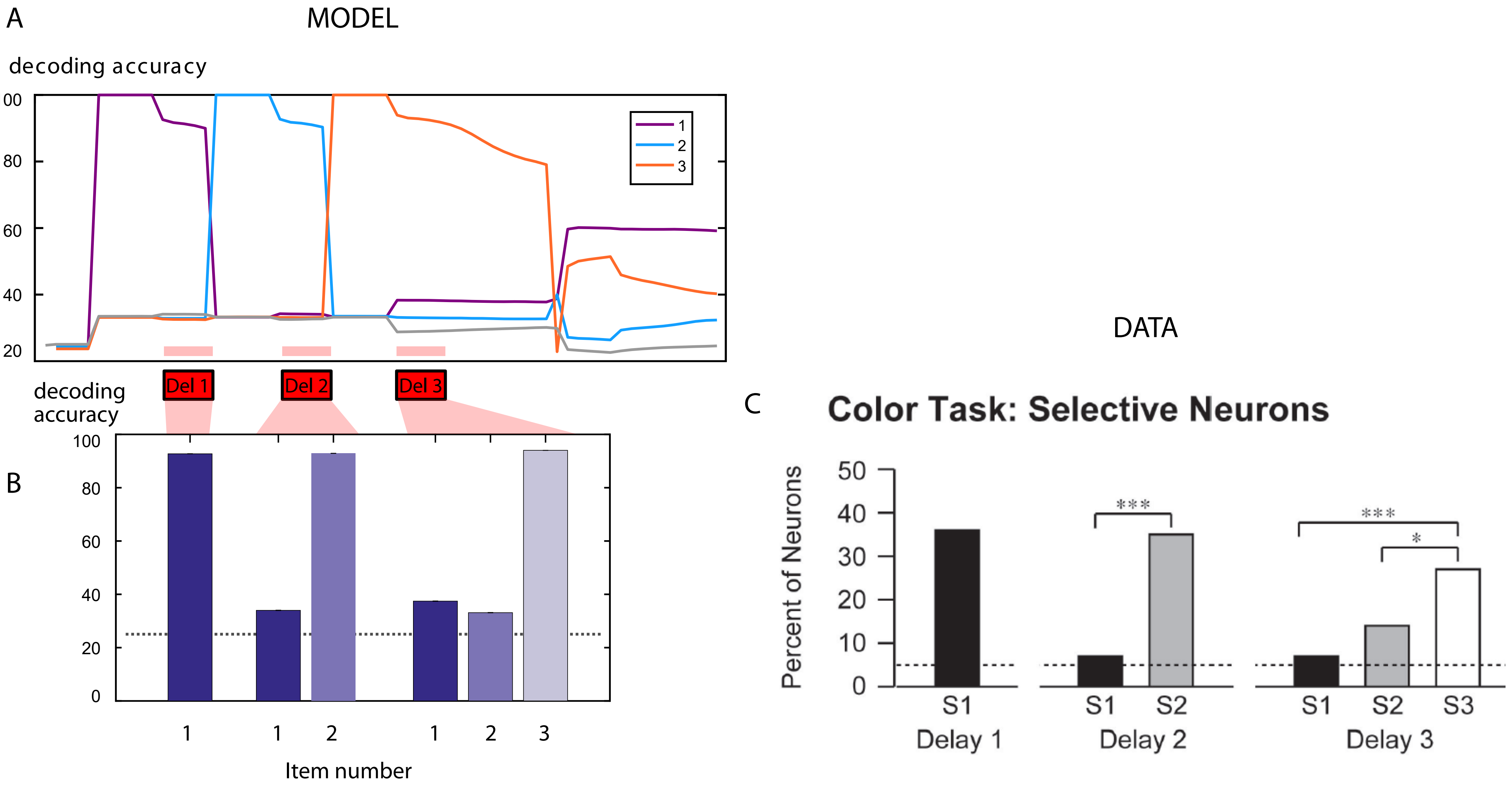

A
MODEL
decoding accuracy
100
80
60
40
20
1
2
3
Del 1
Del 2
Del 3
decoding accuracy
B
100
80
60
40
20
0
1
1
2
1
2
3
Item number
DATA
C
Color Task: Selective Neurons
Percent of Neurons
50
40
30
20
10
0
***
***
*
S1
Delay 1
S1
S2
Delay 2
S1
S2
S3
Delay 3

Figure S4: Decoding from feature units during the delay period (Simulation 11)
Feature-neuron activity during the delay period of the 3-item condition was examined. For each time point, the activity of the four color units was used to decode (across trials) the identity of each presented color. A linear classifier was trained on the activity in the feature units **f** on 50% of trials, and tested on the remaining trials. A) The decoder accuracy for each of the three items is shown, as a function of time. The three delay periods of interest, following the presentation of each item, are shown as pink bars below. B) Average decoder accuracy during each of the delays. During the first delay we could decode the identity of the first item's colour. During the second delay, we could decoded the second items' color but not the first item's color. In the third delay we could decode the third item's color but not the other two colors. Dotted line is chance. C) Data from (Konecky et al., 2017), showing decodability of only the most recent item presented during a monkey WM task. The neurons were in fact from the principal sulcus, and 35% of neurons recorded here were feature-selective, though the remaining neurons were not.

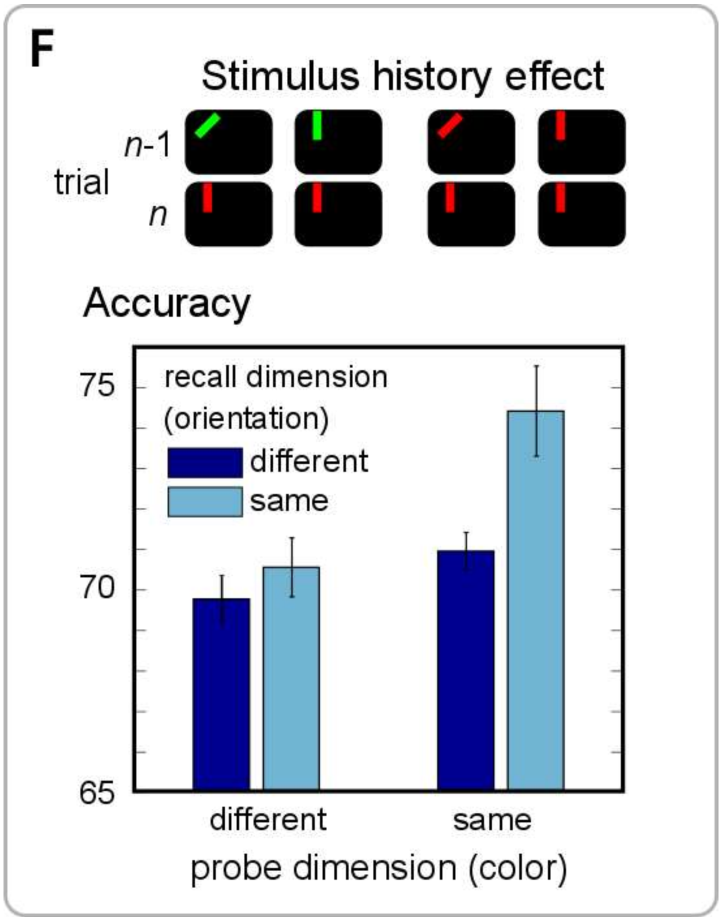


Figure S5: Novel prediction: Trial-to-trial conjunctions effect (Simulation 12)
Since conjunctive neuron selectivities rely upon synaptic traces from the trial history, we predict that the stimuli presented on the previous trial generate interference effects with the current trial. The model predicted that when the probed item's features were identical to those of the item probed on the previous trial, responses were more accurate.

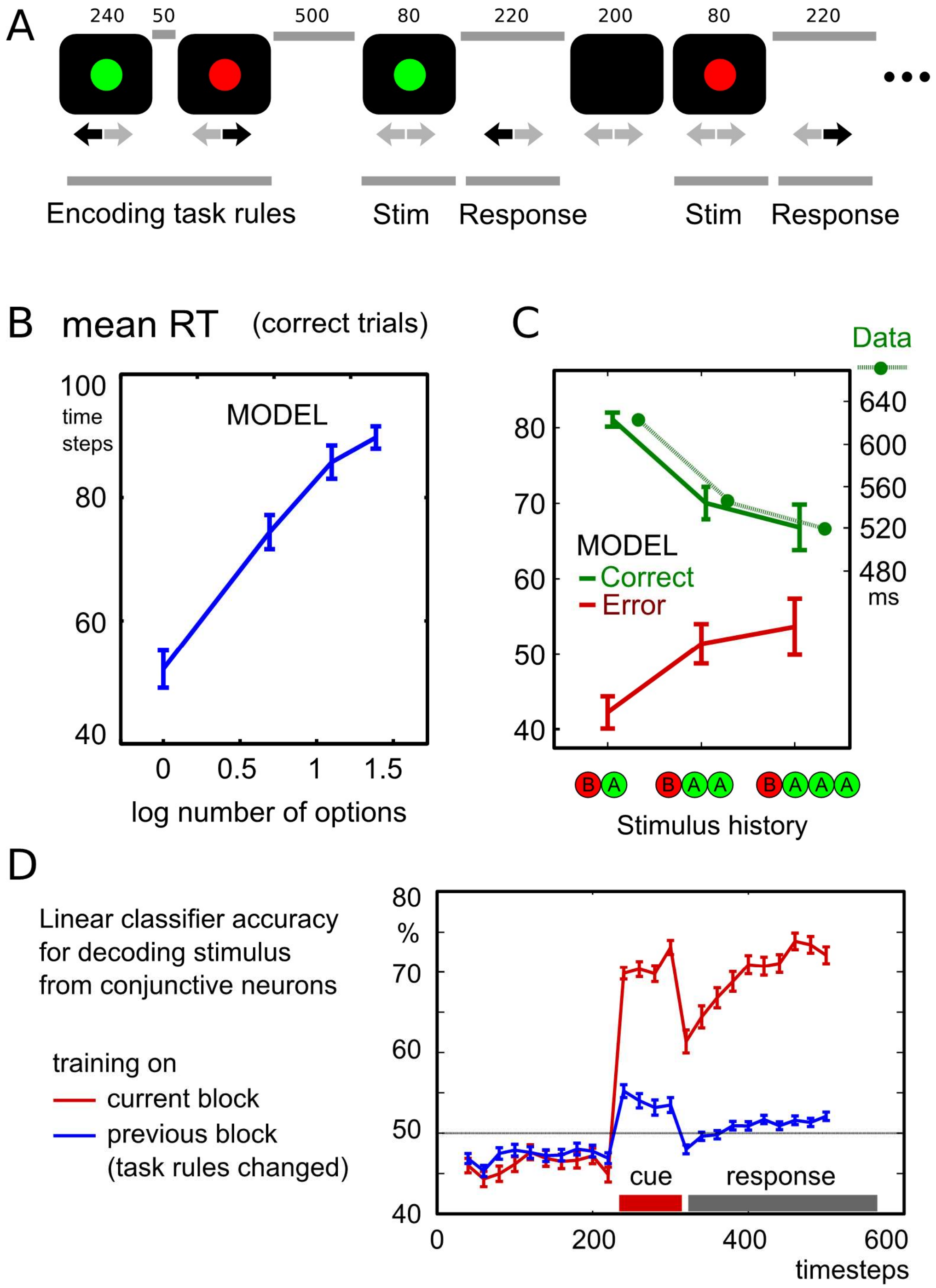


Figure S6: Acting on multiple task rules (Simulation 14)

Here we simulated a simple N-alternative choice task, where each colour indicates that a particular button should be pressed.

A) At the start of each block, a set of task rules was encoded. A single color and motor plan feature were activated simultaneously, indicating the rule for example “if red, press left”. After 1 to 4 rules were presented, a series of test trials followed. One of the colour features that was presented previously was activated, corresponding to presenting a cue. The subsequent response unit activation was measured, to indicate the model's response. Reaction times were calculated as previously from the time of stimulus onset.

B) The RT increased with the logarithm of the number of rules encoded, according to Hick's law.

C) RTs were split according to stimulus repetition history. If the same stimulus was tested on the previous trial ('BAA'), or on the previous two trials ('BAAA'), then the RT on correct trials was faster than if the stimulus was different ('BA'), in keeping with data. Dotted line: RT on correct trials replotted from Expt 4 of (Schvaneveldt and Chase, 1969), in which one of 4 responses was selected after seeing one of 4 stimuli, according to an arbitrary stimulus-response mapping. The model also predicts errors will be faster than correct responses, with an inverted stimulus-repetition effect.

D) Unlike in the WM task, decoding is possible from conjunctive units, since the task set is maintained rather than overwritten on each trial. For each test trial, the stimulus identity was decoded using the other trials in the same block for training (red), or the trials in the previous block (blue). Decoding was possible within the block, because a consistent conjunction unit pattern – representing the task rule corresponding to the current stimulus – was activated.

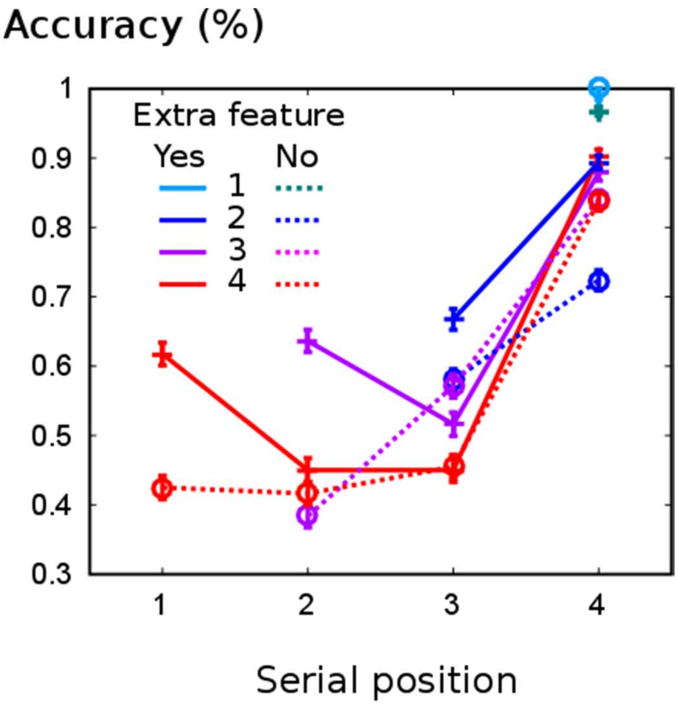


Fig. S7: Novel prediction: Benefit with an extra feature dimension (Simulation 15)
Previous simulations had three features per object (colour, orientation and location), of which only two were task-relevant. The third, irrelevant, feature was distinct for each object. Removing the task-irrelevant feature (e.g., presenting sequential items all at one location, rather than at different locations) worsens model performance, in particular by reducing primacy and recency benefits.

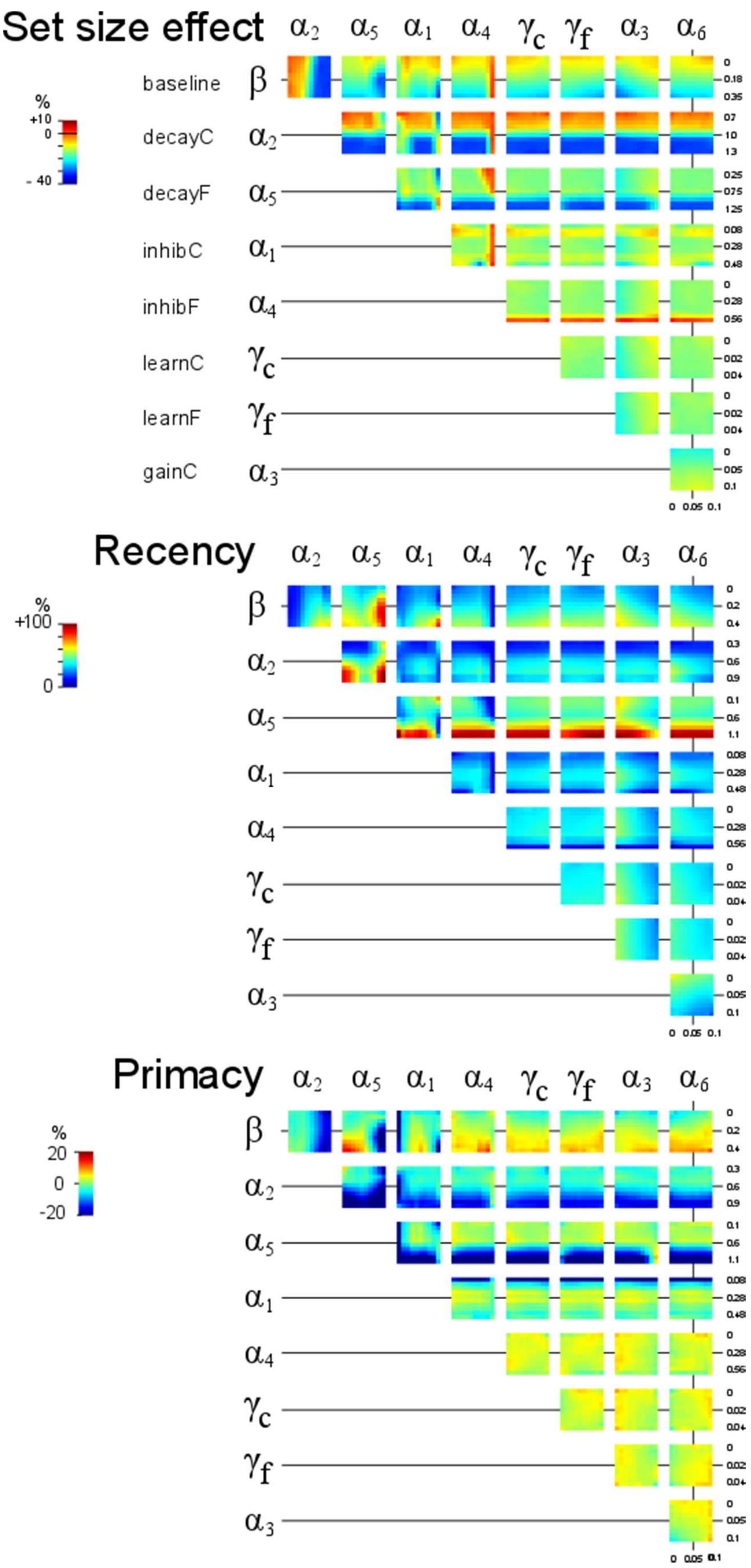

Set size effect
α2 α5 α1 α4 γc γf α3 α6
%
+10
0
- 40
baseline β
decayC α2
decayF α5
inhibC α1
inhibF α4
learnC γc
learnF γf
gainC α3
0 0.18 0.35
0.7 1.0 1.3
0.25 0.75 1.25
0.08 0.28 0.48
0 0.28 0.56
0 0.02 0.04
0 0.02 0.04
0 0.05 0.1
0 0.05 0.1
Recency
α2 α5 α1 α4 γc γf α3 α6
%
+100
0
β
α2
α5
α1
α4
γc
γf
α3
0 0.2 0.4
0.3 0.6 0.9
0.1 0.6 1.1
0.08 0.28 0.48
0 0.28 0.56
0 0.02 0.04
0 0.02 0.04
0 0.05 0.1
0 0.05 0.1
Primacy
α2 α5 α1 α4 γc γf α3 α6
%
20
0
-20
β
α2
α5
α1
α4
γc
γf
α3
0 0.2 0.4
0.3 0.6 0.9
0.1 0.6 1.1
0.08 0.28 0.48
0 0.28 0.56
0 0.02 0.04
0 0.02 0.04
0 0.05 0.1
0 0.05 0.1

Fig.S8: Influence of model parameters upon central behavioral effects (Simulation 16)

We examined the effect of varying 9 free parameters in the model, varying two of them at a time. The parameter values used in the main paper lie at the centre of each 10 x 10 grid, and the figure represents all possible pairs of free parameters. For each parameter combination, we quantified the set size effect (reduction in accuracy as set size increases), primacy effect (difference in accuracy between final and penultimate items in sequence) and recency effect (difference in accuracy between first and second items in sequence. Warm pixels indicate larger effects.

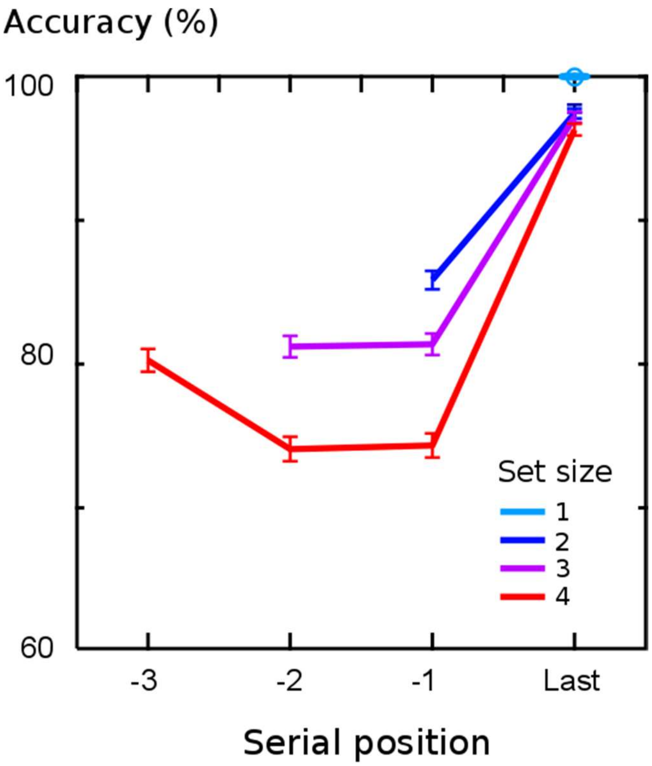


Fig.S9: Basic results for ‘high-accuracy’ parameter regime (used for simulations 3 to 5)
To prevent floor effects when simulating encoding and delay effects, we required a higher initial performance level, and adjusted the parameters accordingly. This figure demonstrates the equivalent of Figure 2C, using this new set of parameters. Performance is overall higher but demonstrates qualitatively similar effects.

Table S1: Empirical findings explained by the model

Phenomenon explained	Mechanism in model	Refs
Set size reduces recall accuracy	Competition between conjunction units → interference	(Zhang and Luck, 2008)
Primacy – first item benefit	No competition with previous items at encoding	(Baddeley, 1996)
Recency – last item benefit	Features retained in active state; no retrograde interference	(Gorgoraptis et al., 2011)
Only last item **encoded actively** in firing rates	Subsequent items capture focus of attention	(Konecky et al., 2017)
Shifting attention to item improves recall	Re-activation by pattern completion focuses attention on item by sustained activity. Subsequently probing that item is faster and more accurate, as it is already active.	(Myers et al., 2017; Souza et al., 2016; Zokaei et al., 2014b)
Additional features in an object remembered without extra cost	During encoding, synaptic weights increase in parallel to all concurrently active features.	(Allen et al., 2006; Luck and Vogel, 1997; Sala and Courtney, 2007)
Frontoparietal activation during WM maintenance	Conjunctive and feature neuron firing required for focus of attention and thus for encoding	(Rowe et al., 2000; Sakai et al., 2002)
Frontoparietal connectivity is signature of shifting attention	Synapses between conjunctive and feature units are bidirectionally activated to drive persistent activity	(Heinen et al., 2017; Scolari et al., 2015; Szczepanski et al., 2013)
Synchrony between posterior and frontal regions during WM	Reciprocal excitation between conjunctive and feature neurons crucial for stable attractor of focus of attention.	(Fries et al., 2001; Gregoriou et al., 2009)
Fine-grained decoding from PFC during WM is elusive	Conjunctive neurons encode information flexibly across trials, with selectivity depending on recent history	(Cogan et al., 2017; Harrison and Tong, 2009; Lara and Wallis, 2014; Sprague et al., 2016)
Working memory operates as an **attentional template**	Unattended items in WM maintained synaptically so stable attractor is reactivated by partial information	(Lavie and Fockert, 2005; Woodman et al., 2007)
Memory contents lead to obligatory **capture of attention**	Partial information pertaining to items in silent WM is amplified, leading to persistent activity / focus of attention	(Olivers et al., 2006; Soto et al., 2008)
WM capacity correlates with **complexity of task set**	One WM object corresponds to one stimulus-response pairing	(Conway et al., 2003; Duncan, 2010)
PFC subserves both WM and task set maintenance	Conjunctive neurons can flexibly bind actions to stimulus features, or groups of features, together	(Barch et al., 1997; Duncan et al., 2000)

Neural decoding of **unattended items** in WM is weak	Unattended items encoded in synaptic traces	(Lewis-Peacock et al., 2012; Sprague et al., 2016)
TMS to feature areas, or indiscriminate sensory stimuli, can **re-activate** representations	Neurons reactivated by TMS pulse that are connected synaptically to conjunctive neurons (i.e. unattended WM items) are selectively amplified by reciprocal synapses	(Rose et al., 2016; Wolff et al., 2017)
TMS to posterior cortex disrupts benefit conferred by focus of attention	Electrical activity i.e. persistent activation is disrupted by electrical stimulus, but synapses are not	(Zokaei et al., 2014a)
rTMS to PFC disrupts WM by altering posterior activity	Reducing conjunctive unit excitability reduces stability of attentional attractors	(Zanto et al., 2011)
Attention operates by frontal **amplification** of posterior feature-selective neurons	The attractor basin formed by mutual conjunctive and feature synapses allows conjunctive units to control gain in feature-selective neurons.	(Desimone and Duncan, 1995; Merrikhi et al., 2017; Moore and Armstrong, 2003)
Primacy effect not robust	Depends on ITI and events before trial	(BADDELEY, 2000)
Transposition errors	New conjunction unit not always activated for new items	(Farrell and Lewandowsky, 2004)
Binding must be **serial**	Conjunctive neurons encode all simultaneously active features.	(McLean et al., 1983; Treisman and Gelade, 1980)
More items mean faster decay	Focus of attention is robust to decay	(Pertzov et al., 2016)
Errors report succeeding items more than prior items	More recently encoded items have stronger synaptic weights, and more likely to intrude.	(Farrell and Lewandowsky, 2004)
Probe can interfere with recall	Irrelevant features in probe suppress reactivation of items	(Souza et al., 2016)
Reaction times inversely related to accuracy	Time taken for conjunctive unit to activate features depends on synaptic strength and competition from other units	(McElree and Dosher, 1989; Pearson et al., 2014)
Encoding rate slower when more items stored	Items encoded briefly are more susceptible to interference from other items in memory	(Bays et al., 2011)

Table S2: Psychological concepts corresponding to the model

Binding	Simultaneous activation of two feature neurons causes a single conjunction neuron to become active and form bidirectional connections to those feature neurons.
Focus of attention	Active representation in feature-selective neurons, coupled with a conjunctive unit that is simultaneously active. The two types of neuron are mutually excitatory and generate persistent activity.
Unfocused item in WM	Bidirectional increases in synaptic weights between a combination of place-coded feature neurons, and one conjunction neuron
Recall	Associative re-activation of a pattern of activity in feature neurons.
Task set	One of the feature dimensions represents motor-plan neurons. During encoding, the simultaneous activation of a motor plan and a perceptual feature causes a conjunctive unit to associate the two. Reactivating the perceptual feature triggers the motor program.
Top-down control	Conjunction neurons effectively amplify feature neurons that were previously encoded as WM items.
Forgetting from WM	Activation of feature neurons by an external stimulus provides new input. This input competes to be represented by conjunction neurons. The conjunction neuron with the most-similar connections will be activated and re-wire to encode the stimulus. This interfering stimulus thus overwrites and displaces an item previously in WM.

Movie S1: **Timecourse of activity during working memory encoding and recall**. (A) top left panel shows the instantaneous rate of change in weights Δ, (B) below is shown the current synaptic weights. (C) top right illustrates the object currently encoded by the feature neurons. The drawn intensity of each possible stimulus is the product of the activity of the corresponding feature neurons. During encoding this corresponds to the objects presented to the model. (D) lower panels show the feature neuron activity as a heatmap, as a function of time, and the activities of the four conjunctive neurons as traces. Delay period activity generally corresponds to the final item presented. Errors occur when one conjunctive neuron is active for two objects, or when one conjunctive unit fails to win the competition.