Spikes not slots: noise in neural populations limits working memory

Paul M. Bays

UCL Institute of Neurology, Queen Square, London, WC1N 3BG, UK

This opinion article argues that noise (randomness) in neural activity is the limiting factor in visual working memory (WM), determining how accurately we can maintain stable internal representations of external stimuli. Sharing of a fixed amount of neural activity between items in memory explains why WM can be successfully described as a continuous resource. This contrasts with the popular conception of WM as comprising a limited number of memory slots, each holding a representation of one stimulus – I argue that this view is challenged by computational theory and the latest neurophysiological evidence.

Deterministic and stochastic views of WM

WM refers to the ability of the nervous system to actively internally maintain information over brief intervals [1–4]. It is considered an essential component of most complex behaviours and is closely linked to general intelligence. Critically, WM is strongly limited in its ability to hold multiple representations simultaneously, constraining the complexity of mental operations. In vision, two classes of explanation have dominated attempts to account for this limit.

The first class proposes that the limit is due to a deterministic mechanism: there is a fixed maximum number of representations that can be held in memory at one time; once this limit is reached, further items cannot enter memory [5–7]. These models draw support from theoretical studies of synchronised neural activity [8,9] and imaging studies claiming to show plateaus in the activation function at a particular number [10,11] (see [3] for a critical review). The classic deterministic account of visual WM is the 'slot' model, which posits three or four independent memory slots, each holding a detailed representation of one visual object [6].

The second class asserts that the limit has a stochastic mechanism: representations in memory become increasingly variable as their number increases, until they can no longer be distinguished from random noise [12–19]. These models find support in the inherent variability of neural activity [20,21], and evidence from single-neuron recording and fMRI decoding studies [22–24]. In vision, the stochastic (see Glossary) view is exemplified by the 'resource' model, which proposes that a limited supply of a representational medium is continuously distributed between visual objects; items that receive more resource are stored with less noise [16–19]. Several 'hybrid' models have also been suggested, combining elements of both slot and resource accounts [25–27].

Crucial evidence favouring stochastic models has come from examining the distribution of errors in analogue recall tasks. A typical task, testing memory for orientation, and the observed distributions of error are shown in Figure 1A,B (data from [28]; similar distributions have been obtained for other visual features, e.g., colour and motion direction [15,16,18,29]). The first important observation is that variability (the width of the distribution) increases steadily with the number of items in the memory array (Figure 1B,C). The relationship between variance and set size is well described by a power law (appearing

Glossary

Analogue recall task: an experimental task in which an observer is required to select the value of a remembered stimulus feature from a continuous space of possibilities.
Delay period: the interval in a WM task between presentation and test, during which items must be maintained in memory.
Feature tuning: tendency of a visual neuron to fire maximally in response to stimuli with a particular visual feature, for example, a particular orientation. The full relationship between firing rate and feature value is described by tuning function of the neuron.
Gamma oscillation: high-frequency (30–100 Hz) rhythmic neural activity.
Kurtosis: a measure of the shape of a probability distribution, specifically 'peakedness' of the distribution in comparison to the normal distribution.
Lateral inhibition: the ability of an excited neuron to reduce the activation of other functionally similar neurons within its local environment.
Misbinding: an incorrect recombination of visual features belonging to two or more different objects. For example, recalling a red tractor when you actually saw a red car and a green tractor.
Neural oscillation: large-scale rhythmic activity in the nervous system, resulting from synchronisation of neural firing patterns.
Normal distribution: a bell-shaped continuous probability distribution. Sums or averages of random variables drawn from independent distributions are typically normally distributed. The normal distribution in a circular space is called the von Mises distribution.
Normalisation: rescaling a set of values to have a particular common magnitude.
Poisson process: a process that generates events independently at random with a particular mean rate. The spiking activity of a neuron can be approximated by a Poisson process.
Power law: a relationship in which one quantity varies as a power of another, $y = ax^n$. A power law with exponent $n = 1$ describes a linear relationship.
Set size: the number of distinct items presented for memorisation in a WM task.
Signal-to-noise: the ratio of meaningful signal to background noise in a source of information.
Spatial selectivity: tendency of a neuron to fire only in response to stimuli within a particular region of space (its receptive field).
Stochastic: partly or wholly unpredictable/random.
Theta oscillation: low-frequency (4–7 Hz) rhythmic neural activity.

Corresponding author: Bays, P.M. (p.bays@ucl.ac.uk).

Keywords: working memory; population code; normalisation; attractor network; diffusion to bound; binding.

1364-6613/

http://dx.doi.org/10.1016/j.tics.2015.06.004

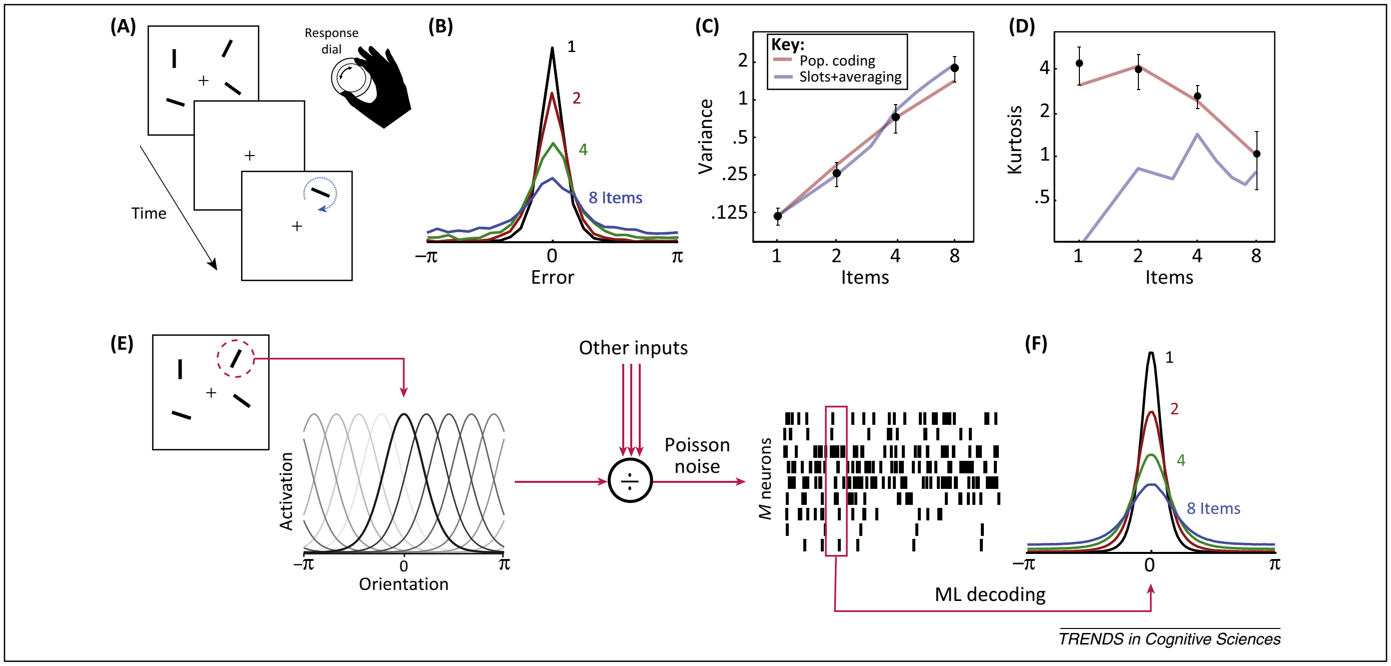


Figure 1. Errors in working memory (WM) reflect noise in neural coding. **(A)** An analogue report task, in which participants adjust a probe stimulus to match the orientation recalled at the same location in the memory array. **(B)** The distribution of responses around the true orientation changes as a function of the number of items in the memory array. **(C)** Error variability (black) increases monotonically with the number of array items. Curves show predictions of the population coding model (red; [28]) and slots + averaging model (blue; [27]) with best-fitting parameters. **(D)** Kurtosis (black) measures deviations from normality of error distributions. Curves as in (C). **(E)** The population coding model. Each stimulus is encoded by a set of orientation-selective neurons with bell-shaped (normal) tuning functions. Normalisation operates across the whole population, scaling summed activity to a fixed level. Neurons generate spikes according to a noisy Poisson process, and recall is modelled as maximum likelihood (ML) decoding of the spiking activity over a fixed time window. **(F)** Error distributions predicted by the population coding model with ML parameters (compare with B). Adapted and reprinted from [28], with permission from the Society for Neuroscience.

linear on log–log axes as in Figure 1C) and there is no evidence for any abrupt discontinuity as might be expected on reaching a deterministic limit. This observation is at the heart of stochastic models of WM and appears incompatible with the classic deterministic view, which predicts no changes in variability so long as all items are stored, that is, until the fixed limit is exceeded.

The second key observation is that WM error distributions deviate significantly from the familiar normal distribution; in particular, the distributions have excess kurtosis (Figure 1D) – they are strongly peaked with long tails. This observation is important because mathematical models in psychology and neuroscience typically assume that internal variables have noise that is normally distributed. On this assumption, the observed deviations from normality have been interpreted variously as evidence for random ‘guessing’ [27] or variability in the precision of memory representations [18,30].

A different perspective on these observations is obtained by considering how visual features are represented in the nervous system. Simple visual features, such as orientation, are encoded in neural activity by population codes (Box 1). A recent study [28] has shown that stimulus estimates obtained by optimal decoding of a noisy population code have non-normal distributions of the type observed in WM experiments (Figure 1E,F). According to this model, the stimulus features in the memory array are encoded in the firing rates of a population of neurons with spatial selectivity and feature tuning. The population activity is normalised, meaning that the total activity level (sum of firing rates) remains constant across changes in set size. The activity pattern persists once the memory array is removed, with neurons firing stochastically according to a Poisson process.

Recall of a probed item is modelled as maximum likelihood decoding (Box 1) of the persistent spiking activity over a fixed time window. Because of the noise in spiking activity, this estimate exhibits variability around the true value; the fewer spikes available to decode, the greater the variability. Because the population as a whole is normalised, the activity coding for each memory item declines with increasing set size. This simple model accurately reproduces the power law relationship between variance and set size observed in experiments (red line in Figure 1C). Furthermore, estimates decoded from a tuned population are not in general normally distributed, and the specific deviations from normality expected fit those observed in experiments (red line in Figure 1D).

Slots and resources

In the population coding model [28], total spiking activity is limited due to normalisation, and is distributed (shared out) between visual items; in this way, the model provides a plausible biological basis for the concept of a limited WM resource. Neural activity can be selectively allocated to stimuli with the greatest behavioural priority, and this quantitatively reproduces empirical observations of enhanced WM precision for such items, as well as corresponding costs for non-prioritised stimuli [17,31,32].

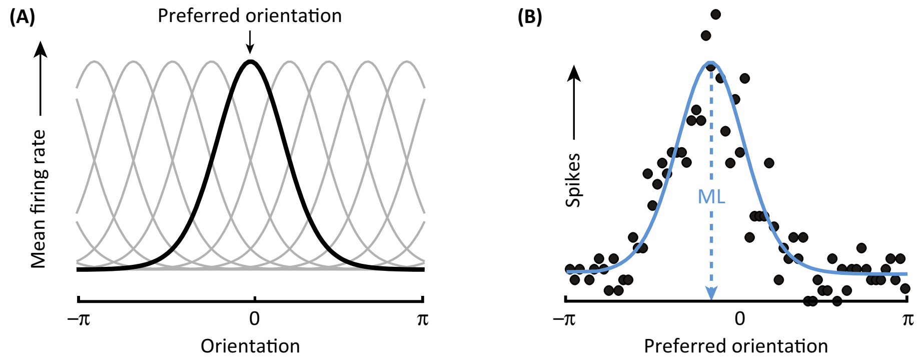
Can slot-based models also account for WM error distributions? An influential study [27] proposed a hybrid model, the ‘slots + averaging’ model that claimed to fit WM errors. The slots + averaging model modifies the classic slot model to allow multiple slots to be allocated

Box 1. Population coding

Population coding is a method of encoding information in the combined activity of a pool of neurons [61,62]. The firing rate of each neuron is determined by a tuning function (Figure IA) with a peak at a 'preferred' value that differs from neuron to neuron. This coding strategy is common throughout the nervous system, including visual cortex, where the encoded values are stimulus features such as orientation, and motor cortex, where values correspond to motor outputs such as an intended movement direction [63,64].

Population coding has the advantage that it is robust: because the encoded information is distributed over many neurons, damage to any one neuron has relatively little effect on the representation. Theoretical schemes have been proposed whereby population codes can store multiple inputs and represent uncertainty in the input [65,66]. Methods also exist by which information in a population code can persist over time (Box 2), and persistent activity associated with WM in cortex, for example in prefrontal neurons, exhibits population coding [67].

Neural firing is probabilistic, thus information encoded by neurons can in general only be recovered imperfectly; because outputs can be averaged over many neurons, population codes help reduce this uncertainty. A particularly important method of decoding, which is in many situations statistically optimal, is maximum likelihood (ML). ML decoding [68] consists of identifying the represented value for which the observed activity was most likely to occur. This is also the value at the peak of an idealised (noiseless) response function that best fits the observed firing rates (Figure IB). Biologically plausible methods have been identified for obtaining ML estimates from population codes [69,70]. However, it is unclear whether decoding takes place explicitly in the brain, or is only implicit in the transformation between variables, for example, from sensory to motor coordinates [71].



TRENDS in Cognitive Sciences

Figure I. Standard model of population coding. **(A)** The mean activity of a neuron in response to an input (e.g., the orientation of a visual stimulus) is determined by its tuning function, which is typically bell shaped. Neurons within a population differ in their preferred stimulus value, that is, the value that elicits their maximum firing rate. **(B)** Population activity (black filled dots) plotted against the preferred stimulus value of each neuron. In any finite time period, the spike rate will be a noisy approximation to the idealised value set by the tuning function. Maximum likelihood finds the idealised output (blue curve) that most closely fits the actual activity [61].

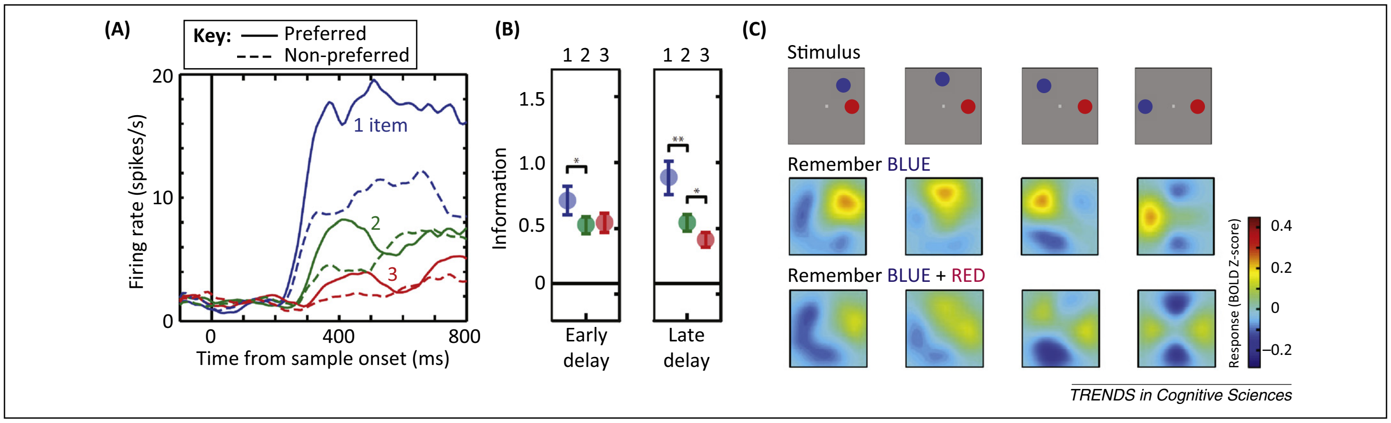
to the same visual item. According to this proposal, the brain maintains multiple independent representations of the same object, each with normally distributed error, and averages them together at recall. As a result, an item that is allocated more slots is recalled with lower variability. In this regard, the slots + averaging model mimics a resource model, with the slots acting as a shared resource (although a quantized one) that is distributed between items; however, it also invokes a deterministic limit: no more items can be stored than there are slots.

The slots + averaging model reproduces the increase in variability with set size observed in experimental data (blue line in Figure 1C). It also predicts some deviations from normality as a result of the mixing together of responses to items with different numbers of slots assigned to them. In particular, estimates of an item that receives no slots will be randomly (uniformly) distributed, which is consistent with the presence of long tails in error distributions at higher set sizes. However, the response distributions predicted by this model do not in fact correspond to those observed in experiments, as indicated by the drastic failure to reproduce the kurtosis of responses (blue line in Figure 1D). In particular, the slots + averaging model – unlike the population coding model – fails to predict the substantial non-normality of errors for small numbers of items, including one item [28]. Currently, no model incorporating a deterministic limit has been shown to reproduce the characteristic deviations from normality observed in WM errors, and this is an important challenge for proponents of this view.

Evidence for normalisation

Normalisation is an important component of the population coding model of WM: it predicts a decline with set size in the strength of the neural signal associated with each item, and hence explains why variability increases with the number of items. Initially identified in primary visual cortex, normalisation has been described as a canonical computation in the brain, implemented through varied mechanisms in multiple neural systems [33,34]. The common principle is that the responses of individual neurons are divided by the summed activity of a larger population of neurons, the 'normalisation pool'.

In many of the classic examples of normalisation (e.g., [35]), this pool is relatively local, that is, the response of a neuron is divided by the summed activity of neurons with similar tuning functions or receptive fields. In the population coding model, in contrast, normalisation operates over all stimuli held in memory. There is growing evidence for such broad normalisation in multiple brain regions. For example, lateral intraparietal (LIP) neurons, which become active when a saccade target is in their receptive field, have been shown to decrease their firing rate as the number of potential targets increases, even if the



TRENDS in Cognitive Sciences

Figure 2. Evidence for normalisation in working memory (WM)-related neural activity. **(A)** Firing rate of an example prefrontal neuron with persistent WM activity. Firing rate declines with increasing memory load, whether the stimulus in the receptive field corresponds to a preferred (unbroken lines) or non-preferred (broken lines) feature of the neuron. **(B)** Correspondingly, the information about a stimulus that can be extracted from delay-period neural activity declines as the number of items increases, consistent with normalisation. (A,B) Adapted and reprinted from [22], with permission from the National Academy of Sciences. Data shown are for lateral prefrontal cortex (LPFC); similar results were found in frontal eye field (FEF) and lateral intraparietal (LIP) areas. Note that significant differences between two and three items were observed only in the late phase of the delay in the LPFC; however, this was not a consistent observation across brain areas. **(C)** An fMRI study [24] decoded signals recorded during a spatial WM task (top row). When participants were instructed to remember the location of one of two stimuli, strong delay-period activity was observed corresponding to the remembered stimulus location only (middle row). When both stimuli were to be remembered, activity was observed corresponding to both locations, but with significantly reduced amplitude at each. Adapted and reprinted from [24], with permission from Elsevier. Data shown are from area V4A; similar results were found throughout visual and posterior parietal cortex.

targets are widely spaced [36–38]. Similar results are found for premotor neurons as the number of possible targets for a reach increases [39].

One study [22] recorded from neurons in primate frontal and parietal cortex that were active during the delay period of a colour WM task. Neurons were observed that decreased their stimulus-specific firing rate as the number of competing stimuli increased (example in Figure 2A). Overall, the informativeness of neural activity about the stimuli declined with increasing memory load (Figure 2B), consistent with a decrease in neural signal-to-noise due to normalisation. Interestingly, both behavioural and neural responses indicated that items competed for representation only with others in the same hemifield; a small behavioural advantage of bilateral presentation has also been demonstrated in humans [40], consistent with a similar (although much weaker) division of resources.

A recent study [24] used an image reconstruction technique to recover stimulus information from fMRI signals recorded during the delay period of a spatial WM task. Spatial locations of multiple stimuli held in memory were successfully decoded from signals in occipital, parietal, and frontal sites (Figure 2C). In occipital and posterior parietal areas, an increase in the number of locations in memory from one to two resulted in a decrease in the BOLD amplitude associated with each stimulus representation. This represents the most compelling evidence yet for a broad mechanism of normalisation underlying WM.

Decay of WM representations

WM variability increases the longer items are maintained in memory [41,42]. The rate of this decay can be manipulated by retrospective cues that direct attention to particular items in memory. A cued item becomes relatively resistant to decay, but at a cost to uncued items, whose rate of decay increases [42]. Similarly, decay rate increases with set size [43]. These observations are consistent with allocation of a limited neural resource to memory maintenance.

What is the neurophysiological basis for WM decay? Within the population coding framework, one possibility is that the neural activity corresponding to each stimulus declines in amplitude with time. However, the population coding model does not describe how this activity is sustained or give a basis for its decay. Instead, theoretical studies have tended to focus on the possibility of diffusion within an attractor network (Box 2). A continuous attractor network maintains stimulus information in a self-sustaining pattern of activity (i.e., a persistent population code), which can be visualised as a 'bump' of activity centred on the true stimulus value.

Over time, this bump diffuses (follows a random walk) through the space of possible stimulus values. As a result, recall variability increases steadily with time.

A recent study [44] found evidence for such random diffusion in prefrontal neurons recorded during a spatial WM task. As predicted by diffusion, tuning functions computed from responses at the end of the delay period were biased in the direction of the error in the monkey's response. Because drift of the activity bump towards one parameter value necessarily implies drift away from other values, the diffusion model predicts that neurons with preferred directions falling either side of the target direction will be anticorrelated: this was confirmed. Neither of these results were predicted by a model with declining amplitude instead of diffusion.

A theoretical study of diffusion within attractor networks [45] has shown that the rate of diffusion is directly related to the precision with which an ideal observer can decode the state of the network. Thus, factors that increase the variability of retrieval from a population code, such as decreased mean firing rate, are also expected to increase the rate of diffusion. It follows that normalisation within an attractor network could account for resource-like behaviours, including the increase of decay rate with number of items.

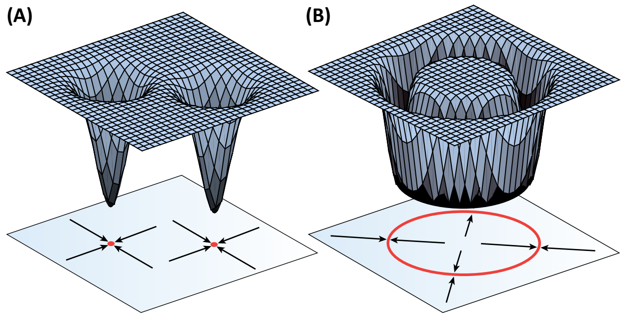
One simulation study of an attractor model with normalisation [46] highlighted other behaviours that

Box 2. Attractor models

An attractor network [72,73] is a neural circuit for which certain patterns of activity are stable and self-sustaining. Regardless of how the network is initialised, activity will settle into one of these stable patterns or 'attractor states'. Such behaviour arises naturally in computational models of networks with recurrent excitation and is considered a possible basis for persistent activity underlying WM.

A discrete attractor network [74,75] has one or more fixed points of attraction (Figure IA). Such a network can maintain the memory of a categorical variable, with each fixed point corresponding to a different discrete value (e.g., a letter of the alphabet). Discrete attractor networks can carry out a type of pattern completion: if the network is initialised with an input that does not correspond exactly to one of the categories, the activity will shift to the nearest stable state corresponding to the closest-matching category.

Continuous attractor models [76–78] extend this principle to networks with a continuum of attractor states. For example, memory for a continuous variable such as orientation can be stored in a 'ring attractor' network (Figure IB), where every possible orientation corresponds to a different point on the ring of stable states. Similar to discrete attractors, these networks are relatively resistant to internal noise, in that the system rapidly corrects for perturbations away from the attractor line. However, perturbations along the attractor are not corrected, making the network susceptible to random drift ('diffusion') in its stored variable.



TRENDS in Cognitive Sciences

Figure I. Attractor networks. **(A)** Energy landscape and state-space representation of a discrete attractor network with two fixed points (red dots). Here, two state variables (*x* and *y* axes) stand in for the high-dimensional state of activity in the network. Activity evolves over time (arrows) to settle in the nearest of the fixed attractor states. **(B)** Representation of a ring attractor network. Activity is stable anywhere on the continuum of attractor states (red line).

could contribute to WM decay. When multiple stimulus values were stored as separate activity bumps within a single attractor network, individual bumps were found to spontaneously fade away, and neighbouring bumps were observed to probabilistically merge together. The spontaneous loss of information of a whole item was compared with slot models of WM; however, it should be emphasised that fade out and merging in this model were probabilistic, occurred over the course of maintenance, and did not occur with frequencies appropriate to emulate a limit at a fixed number of items. Unlike diffusion, there is no clear neurophysiological evidence for these behaviours. While an important development, the normalised attractor model has a number of issues as a model of WM. In particular, as a result of merging, this model is unable to store two similar stimuli (e.g., two slightly different colours) as distinct entities. This is because, unlike the population coding model, there is no mechanism for associating features with locations. A corollary is that the model cannot plausibly simulate retrieval of a probed stimulus, as is needed to describe behaviour on analogue recall tasks.

Recall latency

It takes time to retrieve information from WM, and the time taken (latency) increases with the total number of stimuli in memory [43]. Examination of retrieval times reveals that latency has a strongly skewed distribution, but the reciprocal of latency has an approximately normal distribution. This is characteristic of an accumulation process, in which the response is triggered by the threshold crossing of a variable with a normally distributed rate of rise. While early work on such models was in the context of WM [47], they have since become more closely associated with the timing of sensory judgements [48], where there is substantial neurophysiological evidence relating decision latency to cells whose activity drifts towards a boundary value, with the rate of drift determined by the strength (signal-to-noise) of the stimulus input.

In WM retrieval, the accumulation would be of internal evidence provided by a neural population coding for the stimulus. The slower retrieval with larger set sizes is consistent with reduced signal-to-noise and hence with a normalisation account of this storage mechanism; so is evidence of faster responses for prioritised stimuli, and corresponding latency costs for non-prioritised items [43]. While this initial evidence is compatible with the stochastic view, predictions of deterministic models for analogue recall latencies are yet to be elucidated: preliminary consideration suggests a two-component mixture of report times, corresponding to separate latency distributions for items in and out of memory. One important consideration for any model is that decay of the WM representation is expected to continue during the course of retrieval. As a result, it may become suboptimal to accumulate evidence over long periods [45], possibly placing an upper limit on latency.

Failures of binding

Binding refers to the information that groups different visual features together into objects. In perception, binding errors ('illusory conjunctions') are rare and primarily observed when stimuli are unattended or briefly glimpsed [49,50]. In contrast to this, misbinding in WM has a significant impact on our ability to accurately recall what we have seen. In analogue recall tasks (Figure 1A), binding errors are observed as clusters of responses around the values of other, unprobed items in the memory array [16,31,51–54]. Binding errors increase in frequency with increasing set size.

A long-standing proposal [8,9,55–57] is that binding is represented in the brain by synchronised firing of neurons selective for individual visual features (Figure 3A). According to this model, neurons corresponding to features belonging to a single object all fire at the same phase of a common neural oscillation, typically assumed to be in the theta range. This provides a mechanism for linking activity in disparate brain regions specialised for different visual features. To decode this information, a mechanism would need to be sensitive to the relative timing of spikes arriving from different inputs, perhaps by taking advantage of the

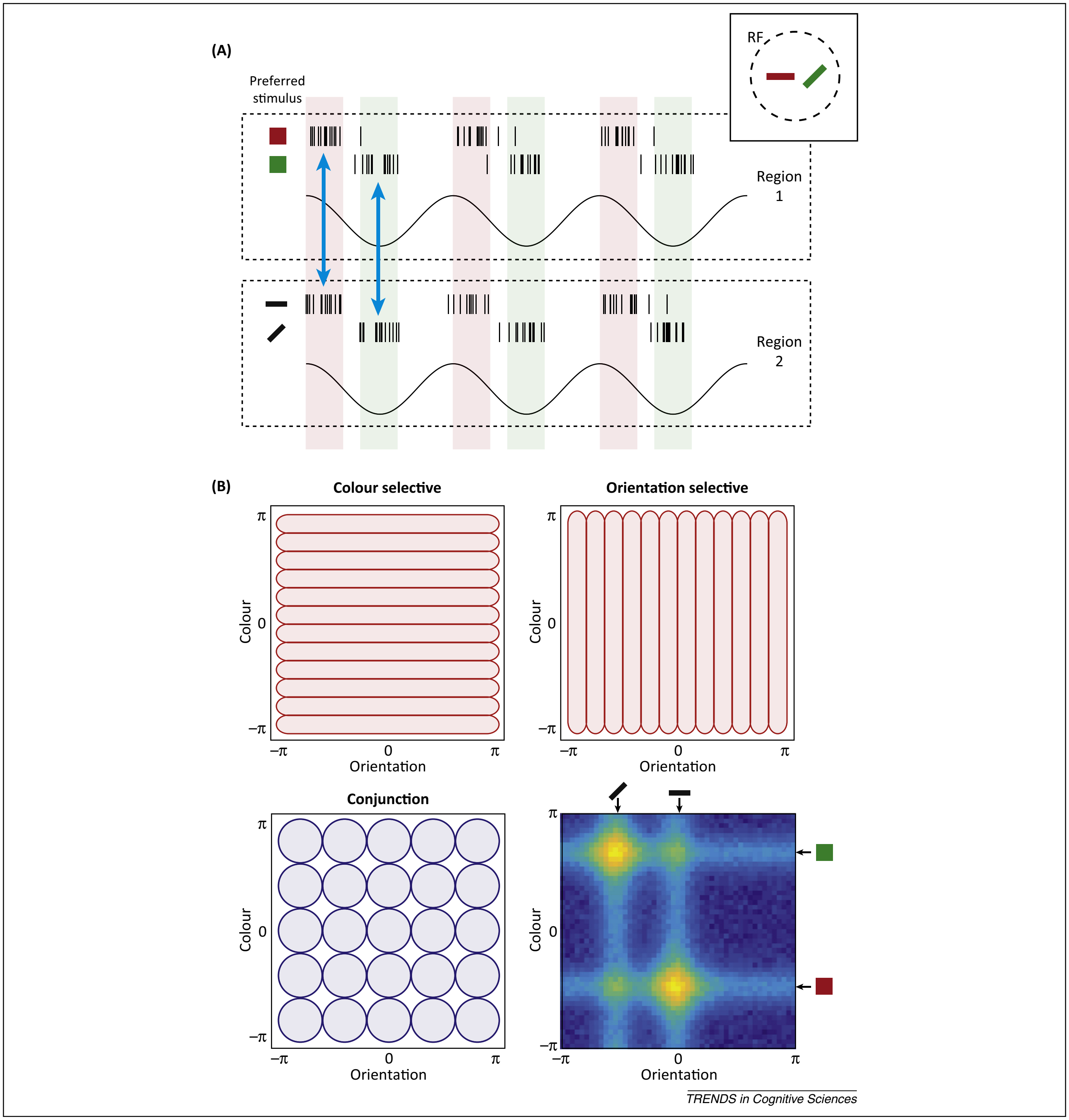


Figure 3. Models of binding and misbinding in working memory (WM). **(A)** Binding by synchrony. Neurons in two brain regions synchronise their activity to a common waveform (e.g., theta rhythm) in order to store in memory (inset) a red horizontal bar and a green oblique bar. A red-selective neuron in colour region 1 fires spikes in the same phase as a horizontal-selective neuron in orientation region 2; likewise, activity of a green-selective neuron is matched with an oblique-selective neuron. Spike timing is probabilistic, and errors of binding (e.g., recall of a green horizontal bar) arise from mistimed spikes. **(B)** Binding by conjunction. Stimulus information is stored in the activity of a mixed population of neurons, comprising colour-selective neurons (response fields shown top-left), orientation-selective neurons (top-right), and conjunction neurons that are activated only by a specific pairing of colour and orientation (bottom-left). Optimal decoding of the population activity recovers the stimulus values and their binding (decoding probability shown bottom-right). Misbinding occurs probabilistically because of noise in firing rate.

rapid depolarising effect of coincident synaptic inputs [58]. Binding errors could arise because spike timing is probabilistic and hence some activity will take place in the 'wrong' phase, creating spurious synchronisations.

An influential study [8] suggested that a deterministic capacity limit could arise from a limit on the number of distinct phases available in the background oscillation. They proposed that each item is associated with a particular cycle of a high-frequency gamma oscillation phase-locked to a slower theta rhythm. They illustrated this concept with a fixed limit of seven items (the prevailing estimate of WM capacity at the time), although the broad range of frequencies typically associated with gamma and theta rhythms would allow almost any number to be selected as the limit. A similar model [9] proposed that the limit on distinct phases arose from mutual inhibition between assemblies of neurons representing different objects. Rather than a deterministic limit, this model predicted that the number of distinct phases fluctuated from trial to trial, with a mean that again depended on the choice of model parameters.

Critically, because misbinding can only occur between items in memory, an oscillation-based model with a limit on the number of items stored makes the strong prediction that, once the limit is reached, any further increase in set size will not increase the frequency with which binding errors occur. No such plateau in misbinding frequency is observed experimentally [16].

Despite the longevity of the idea, direct neurophysiological support for binding-by-synchrony is still lacking. It has also not been established how the correct synchronisations between disparate brain regions would be initialised by sensory input, particularly when stimuli are presented simultaneously. An alternative view [59] is that binding information is maintained by specialised neurons that are selective for the conjunction of elementary visual features. A 'mixed' population code (Figure 3B) includes both conjunction neurons and neurons selective for single features. For example, the population might include neurons that fire in response to a red stimulus, neurons that respond to a horizontally oriented stimulus, and conjunction neurons that respond only to a red horizontally oriented stimulus. Similar populations have been observed in the brain, for example, in area V2 [60].

In this model, decoding can be achieved by standard methods (e.g., maximum likelihood) and binding errors occur for the same reason as variability in recall of a single feature: noise in activity of the population of neurons. This model has been shown to reproduce the qualitative pattern of binding errors observed in analogue report tasks, including the monotonic increase in misbinding with set size [59].

Concluding remarks

Multiple aspects of WM performance degrade smoothly and continuously with increasing memory load, including recall precision, retrieval latency, decay rate, and binding fidelity. In each case, there is evidence that individual items can be afforded protection against degradation, but only at a cost to other unprioritised stimuli. These results are consistent with the characterisation of WM as a limited resource, and difficult to reconcile with the proposal of a deterministic limit at three or four items. More specifically, the evidence in each case appears consistent with the flexible allocation of a limited quantity of neural signal between memorised stimuli. As the neural signal decreases, representations become increasingly dominated by noise, and this accounts for the degradation in WM fidelity.

The inherent stochasticity of neural activity means that theoretical models based on neural principles are broadly incompatible with a deterministic limit on number of items stored. While some models are capable of generating randomly fluctuating limits, the principles underlying them, such as spontaneous fade out of activity bumps, do not have clear support in neurophysiological observations. No model predicts a limit specifically in the three- to four-item range. In contrast to this, the mechanisms of graded degradation discussed here are all based on established neurophysiological principles: population coding [61], normalisation [33], diffusion [44], and accumulation-to-bound [48]. Nonetheless, these links between behavioural observations and neurophysiology are to varying degrees speculative or theoretical at this time, and further computational, behavioural, and neurophysiological studies will be required to establish them firmly. Some outstanding questions are listed in Box 3.

Box 3. Outstanding questions

- How can the stochastic view of WM be extended to memoranda that are categorical in nature, for example, letters, shapes, or high-level objects? Can discrete attractor networks (Box 2) provide a useful description?
- How are time intervals and the temporal order of events represented in WM?
- Can the population coding model of visual WM be adapted for other modalities, for example, tactile or auditory WM?
- Are there limits on flexible allocation? A system based on binary spiking events could be practically limited in how finely activity can be distributed between representations. Furthermore, under certain conditions it may be statistically optimal to limit allocation to a subset of available items, rather than store all with very low precision. Such effects might be captured by a stochastic upper bound, as in [54].
- What is the neurophysiological basis for the representation in WM of ensemble statistics [79], such as the average colour of a group of objects?
- What is the mechanism underlying the broad normalisation observed in WM-related activity, and how is it distinct from those proposed for perception and attention [33]? Possible mechanisms include lateral inhibition within posterior areas, or a resource-limited descending signal from, for example, prefrontal cortex.

Acknowledgements

This work was supported by the Wellcome Trust.

References

1 Baddeley, A. (2012) Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* 63, 1–29
2 D'Esposito, M. and Postle, B.R. (2015) The cognitive neuroscience of working memory. *Annu. Rev. Psychol.* 66, 115–142
3 Ma, W.J. *et al.* (2014) Changing concepts of working memory. *Nat. Neurosci.* 17, 347–356
4 Brady, T.F. *et al.* (2011) A review of visual memory capacity: beyond individual items and toward structured representations. *J. Vis.* 11, 4
5 Cowan, N. (2001) The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* 24, 87–114
6 Luck, S.J. and Vogel, E.K. (1997) The capacity of visual working memory for features and conjunctions. *Nature* 390, 279–281
7 Miller, G.A. (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 81
8 Lisman, J. and Idiart, M. (1995) Storage of 7±2 short-term memories in oscillatory subcycles. *Science* 267, 1512–1515
9 Raffone, A. and Wolters, G. (2001) A cortical mechanism for binding in visual working memory. *J. Cogn. Neurosci.* 13, 766–785
10 Todd, J.J. and Marois, R. (2004) Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature* 428, 751–754
11 Vogel, E.K. and Machizawa, M.G. (2004) Neural activity predicts individual differences in visual working memory capacity. *Nature* 428, 748–751
12 Frick, R.W. (1988) Issues of representation and limited capacity in the visuospatial sketchpad. *Br. J. Psychol.* 79, 289–308
13 Lakha, L. and Wright, M.J. (2004) Capacity limitations of visual memory in two-interval comparison of Gabor arrays. *Vis. Res.* 44, 1707–1716
14 Palmer, J. (1990) Attentional limits on the perception and memory of visual information. *J. Exp. Psychol. Hum. Percept. Perform.* 16, 332–350

15 Wilken, P. and Ma, W.J. (2004) A detection theory account of change detection. *J. Vis.* 4, 1120–1135
16 Bays, P.M. *et al.* (2009) The precision of visual working memory is set by allocation of a shared resource. *J. Vis.* 9, 7
17 Bays, P.M. and Husain, M. (2008) Dynamic shifts of limited working memory resources in human vision. *Science* 321, 851–854
18 van den Berg, R. *et al.* (2012) Variability in encoding precision accounts for visual short-term memory limitations. *Proc. Natl. Acad. Sci. U.S.A.* 109, 8780–8785
19 Huang, L. (2010) Visual working memory is better characterized as a distributed resource rather than discrete slots. *J. Vis.* 10, 8
20 Faisal, A.A. *et al.* (2008) Noise in the nervous system. *Nat. Rev. Neurosci.* 9, 292–303
21 Tolhurst, D.J. *et al.* (1983) The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vis. Res.* 23, 775–785
22 Buschman, T.J. *et al.* (2011) Neural substrates of cognitive capacity limitations. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11252–11255
23 Emrich, S.M. *et al.* (2013) Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. *J. Neurosci.* 33, 6516–6523
24 Sprague, T. *et al.* (2014) Reconstructions of information in visual spatial working memory degrade with memory load. *Curr. Biol.* 24, 2174–2180
25 Alvarez, G.A. and Cavanagh, P. (2004) The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychol. Sci.* 15, 106–111
26 Anderson, D.E. *et al.* (2011) Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. *J. Neurosci.* 31, 1128–1138
27 Zhang, W. and Luck, S.J. (2008) Discrete fixed-resolution representations in visual working memory. *Nature* 453, 233–235
28 Bays, P.M. (2014) Noise in neural populations accounts for errors in working memory. *J. Neurosci.* 34, 3632–3645
29 Zokaei, N. *et al.* (2011) Precision of working memory for visual motion sequences and transparent motion surfaces. *J. Vis.* 11, 2
30 Fougnie, D. *et al.* (2012) Variability in the quality of visual working memory. *Nat. Commun.* 3, 1229
31 Bays, P.M. *et al.* (2011) Temporal dynamics of encoding, storage, and reallocation of visual working memory. *J. Vis.* 11, 6
32 Gorgoraptis, N. *et al.* (2011) Dynamic updating of working memory resources for visual objects. *J. Neurosci.* 31, 8502
33 Carandini, M. and Heeger, D.J. (2012) Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* 13, 51–62
34 Reynolds, J.H. and Heeger, D.J. (2009) The normalization model of attention. *Neuron* 61, 168–185
35 Heeger, D.J. (1992) Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* 9, 181–197
36 Balan, P.F. *et al.* (2008) Neuronal correlates of the set-size effect in monkey lateral intraparietal area. *PLoS Biol.* 6, e158
37 Churchland, A.K. *et al.* (2008) Decision-making with multiple alternatives. *Nat. Neurosci.* 11, 693–702
38 Mirpour, K. and Bisley, J.W. (2012) Dissociating activity in the lateral intraparietal area from value using a visual foraging task. *Proc. Natl. Acad. Sci. U.S.A.* 109, 10083–10088
39 Pastor-Bernier, A. and Cisek, P. (2011) Neural correlates of biased competition in premotor cortex. *J. Neurosci.* 31, 7083–7088
40 Umemoto, A. *et al.* (2010) A bilateral advantage for storage in visual working memory. *Cognition* 117, 69–79
41 White, J.M. *et al.* (1994) Saccades to remembered target locations: an analysis of systematic and variable errors. *Vis. Res.* 34, 79–92
42 Pertzov, Y. *et al.* (2013) Rapid forgetting prevented by retrospective attention cues. *J. Exp. Psychol. Hum. Percep. Perform.* 39, 1224–1231
43 Pearson, B. *et al.* (2014) Working memory retrieval as a decision process. *J. Vis.* 14, 2
44 Wimmer, K. *et al.* (2014) Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci.* 17, 431–439
45 Burak, Y. and Fiete, I.R. (2012) Fundamental limits on persistent activity in networks of noisy neurons. *Proc. Natl. Acad. Sci. U.S.A.* 109, 17645–17650
46 Wei, Z. *et al.* (2012) From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *J. Neurosci.* 32, 11228–11240
47 Ratcliff, R. (1978) A theory of memory retrieval. *Psychol. Rev.* 85, 59–108
48 Shadlen, M. and Kiani, R. (2013) Decision making as a window on cognition. *Neuron* 80, 791–806
49 Treisman, A. and Schmidt, H. (1982) Illusory conjunctions in the perception of objects. *Cogn. Psychol.* 14, 107–141
50 Cohen, A. and Ivry, R. (1989) Illusory conjunctions inside and outside the focus of attention. *J. Exp. Psychol. Hum. Percept. Perform.* 15, 650–663
51 Bays, P.M. *et al.* (2011) Storage and binding of object features in visual working memory. *Neuropsychologia* 49, 1622–1631
52 Emrich, S.M. and Ferber, S. (2012) Competition increases binding errors in visual working memory. *J. Vis.* 12, 12
53 Rerko, L. *et al.* (2014) Spatial transposition gradients in visual working memory. *Q. J. Exp. Psychol.* 67, 3–15
54 van den Berg, R. *et al.* (2014) Factorial comparison of working memory models. *Psychol. Rev.* 121, 124
55 Eckhorn, R. *et al.* (1988) Coherent oscillations: a mechanism of feature linking in the visual cortex? *Biol. Cybern.* 60, 121–130
56 Gray, C.M. *et al.* (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338, 334–337
57 Siegel, M. *et al.* (2009) Phase-dependent neuronal coding of objects in short-term memory. *Proc. Natl. Acad. Sci. U.S.A.* 106, 21341–21346
58 Knig, P. *et al.* (1996) Integrator or coincidence detector? The role of the cortical neuron revisited. *Trends Neurosci.* 19, 130–137
59 Matthey, L. *et al.* (2015) A probabilistic palimpsest model of visual short-term memory. *PLoS Comput. Biol.* 11, e1004003
60 Gegenfurtner, K.R. *et al.* (1996) Processing of color, form, and motion in macaque area V2. *Vis. Neurosci.* 13, 161–172
61 Pouget, A. *et al.* (2000) Information processing with population codes. *Nat. Rev. Neurosci.* 1, 125–132
62 Salinas, E. and Abbott, L. (1994) Vector reconstruction from firing rates. *J. Comput. Neurosci.* 1, 89–107
63 Vogels, R. (1990) Population coding of stimulus orientation by striate cortical cells. *Biol. Cybern.* 64, 25–31
64 Georgopoulos, A.P. *et al.* (1982) On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J. Neurosci.* 2, 1527–1537
65 Zemel, R.S. *et al.* (1998) Probabilistic interpretation of population codes. *Neural Comput.* 10, 403–430
66 Ma, W.J. *et al.* (2006) Bayesian inference with probabilistic population codes. *Nat. Neurosci.* 9, 1432–1438
67 Goldman-Rakic, P.S. (1995) Cellular basis of working memory. *Neuron* 14, 477–485
68 Seung, H. and Sompolinsky, H. (1993) Simple models for reading neuronal population codes. *Proc. Natl. Acad. Sci. U.S.A.* 90, 10749–10753
69 Deneve, S. *et al.* (1999) Reading population codes: a neural implementation of ideal observers. *Nat. Neurosci.* 2, 740–745
70 Jazayeri, M. and Movshon, J.A. (2006) Optimal representation of sensory information by neural populations. *Nat. Neurosci.* 9, 690–696
71 Pouget, A. and Sejnowski, T. (1997) Spatial transformations in the parietal cortex using basis functions. *J. Cogn. Neurosci.* 9, 222–237
72 Amit, D.J. (1992) *Modeling Brain Function: The World of Attractor Neural Networks*, Cambridge University Press
73 Brody, C.D. *et al.* (2003) Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. *Curr. Opin. Neurobiol.* 13, 204–211
74 Hopfield, J.J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79, 2554–2558
75 Amit, D.J. and Brunel, N. (1997) Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb. Cortex* 7, 237–252
76 Seung, H.S. (1998) Continuous attractors and oculomotor control. *Neural Netw.* 11, 1253–1258
77 Zhang, K. (1996) Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *J. Neurosci.* 16, 2112–2126
78 Compte, A. *et al.* (2000) Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex* 10, 910–923
79 Brady, T.F. and Alvarez, G.A. (2011) Hierarchical encoding in visual working memory: ensemble statistics bias memory for individual items. *Psychol. Sci.* 22, 384–392