SI1: Previous biological models of sameness detection

Two broad classes of mechanisms have been proposed for determining whether two stimuli are identical (Grill-Spector, Henson, & Martin, 2006; Kumaran & Maguire, 2007). On the one hand, sequentially presented identical stimuli elicit less activation, due to neuronal “fatigue” or sharpening of the representations. As a result, novel, non-repeated representations have relatively higher levels of activation. However, such models cannot explain why sameness-relations can be generalized: after detecting the repetition in *babagu*, an item with new syllables and the same repetition-pattern (e.g., *wowofe*) will be just as unfamiliar as an item with new syllables and another repetition-pattern (e.g., *wofefe*). As a result, an explicit representation of sameness vs. difference (or match vs. non-match) is required (but see Cope et al., 2018, where generalization is observed under some circumstances).

The second class of mechanisms involves some kind of comparator between memory representations and sensory input, though there are few explicit and biologically realistic models of sameness matches. For example, it has been proposed that the hippocampal CA1 region (and maybe the CA3 region, depending on the studies) are crucial for detecting matches between memory representations and sensory input (while the CA3 regions might have an additional role in retrieving associations; Hasselmo, 2005; Lisman, 1999; Lisman & Otmakhova, 2001).

We will now discuss a number of representative models to illustrate these points.

SI1.1. Hasselmo and Wyble (1997), Carpenter and Grossberg (1987), Wen, Ulloa, Husain, Horwitz, and Contreras-Vidal (2008)

In their simulation of memory retrieval in the hippocampus, Hasselmo and Wyble (1997) provide an explicit model of comparator-based sameness detection, inspired by the anatomy of the trisynaptic circuit. Specifically, the hippocampus receives sensory input from the entorhinal cortex, which in turn projects to region CA3 (via the dentate gyrus). In contrast, CA1 receives input both from CA3 (via the Schaffer collaterals) as well as directly from the entorhinal cortex. If memories are encoded in CA3, the simultaneous input from the entorhinal cortex and CA3 might allow CA1 to detect matches between sensory input (from the entorhinal cortex) and memory representations (from CA3).

Specifically, during encoding of novel items, combined sensory and memory (from CA1) input leads to novel self-organized representations in CA1. To activate these representations during recognition (i.e., to enter the corresponding attractor state), input from both the entorhinal cortex and CA3 is required; sensory input alone does not activate the attractor state. In other words, CA1 enters an attractor state only when the current sensory input matches currently active memory representations in CA3 (see also Ludueña & Gros, 2013, for a model that uses anti-Hebbian learning to configure a *mismatch* detector).

Relatedly, some working memory models detect matches between the contents of working memory and current sensory input by *adding* input from sensory input and WM (Carpenter & Grossberg, 1987; Wen et al., 2008). If an item is in WM, it will provide an additional input. As a result, matches between sensory input and WM can be detected using some threshold (though such a mechanism might not be robust as it depends on the *absolute* fire firing rates;

Engel & Wang, 2011).

SI1.2. Engel and Wang (2011)

While these models detect matches because the combined output of memory representations and matching sensory input trigger the retrieval of other representations, it is also possible to detect matches by *subtracting* sensory input and memory representation. Such a model has been proposed in the context of delayed-match-to-sample tasks. Specifically, Engel and Wang (2011) proposed a biologically realistic model that detects matches through (i) a working memory (WM) sub-network, (ii) a comparator sub-network, and (iii) a decision network receiving input from the comparator network. Neurons in the WM network receive sensory input (but only when attention is directed to the input) and can maintain memory traces through self-excitation. Critically, the comparator network is composed of two distinct populations. One receives both sensory input and input from the WM network (hereafter called sensory+WM neurons). The other receives *only* sensory input but no WM input (hereafter called sensory-only neurons). Engel and Wang (2011) make two other critical assumptions. First, they assume that the total level of excitation should be similar for matches in the sensory+WM population, and for mismatches in the sensory-only population; as the sensory+WM population has an extra excitatory input, they achieved this by scaling down the synaptic excitation targeting the sensory+WM population. Second, the comparator units show center-surround inhibition: there is a (limited level of) excitation from similar stimuli, and much stronger inhibition from dissimilar stimuli.

These assumptions conspire to yield stronger activation in the sensory+WM population for matches, and stronger activation in the sensory-only population for mismatches. As a result, to decide whether a probe matches the target, the decision network just needs to compare the activation of the sensory+WM population and the sensory-only populations. More specifically, in the matching case, the sensory+WM neurons receive input both from the probe and from the matching content of WM; in contrast, the sensory-only neurons receive input only from the sensory representations of the probe. As a result, there is stronger activation in the sensory+WM population. In contrast, in the case of a mismatch, both populations receive input from the sensory representations of the probe.

As Engel and Wang (2011) assume that excitatory input is stronger for the sensory-only population, this population is expected to receive somewhat stronger input than the sensory+WM population. Further, the sensory+WM population also receives input from the (mismatching) WM representation; due to the center-surround inhibition in the network, the probe and the (mismatching) target inhibit each other, further reducing the activation in the sensory+WM population. The decision network just has to decide whether similar orientations have stronger presentation in the sensory+WM population or the sensory-only population.[1]

SI1.3. Johnson, Spencer, Luck, & Schöner (2009)

Another WM model that explicitly incorporates a same/different distinction has been proposed by Johnson et al. (2009). In their model, sensory

[1] This model assumes that WM is mediated by self-sustained activity in a population of neurons. However, it has been questioned whether such self-sustained activation really plays a crucial role in WM (Rose et al., 2016; Stokes, 2015).

input excites (self-sustained) WM representations, which, in turn, inhibits the corresponding sensory activation (with center-surround inhibition in all areas). As a result, upon presentation of the sample stimulus, there is a self-sustained representation of the memory items in WM, but little activation in sensory areas due to the inhibitory input from WM. Hence, if a later sensory input matches the items in memory, the sensory areas will remain largely silent. In contrast, if the sensory input differs from the memory items, sensory input will be uninhibited. Hence, in this model, "decision" neurons that receive excitation from WM will respond to matches, while decision neurons receiving excitation from sensory input will respond to mismatches, at least with mutual inhibition between these decision populations.

However, there are a number of problems with this model. First, it has been questioned whether WM really relies on self-sustained activity (Rose et al., 2016; Stokes, 2015). Second, and crucially, items in (working) memory seem to *attract* attention (Awh & Jonides, 2001; Downing, 2000; Fan & Turk-Browne, 2016) which seems inconsistent with the proposal that memory items suppress perceptual input.

SI1.4. Difficulties of these models with generalizable repetition patterns

In their current instantiations such models are unlikely to account for the generalization of sameness relatoions (nor were these models intended to do so). For example, after exposure to *pupu*, they are unlikely to recognize *baba* over, say, *bapu* when the syllables are novel. In Hasselmo and Wyble's (1997) model, items like *baba* have no memory representation, and thus cannot trigger CA1-like activation any more than *bapu* sequences. That said, a version of Hasselmo and

Wyble's (1997) model might act as a repetition-detector if each item undergoes element-by-element encoding-retrieval cycles. For example, when processing the item *pupu*, the network might first encode the first instance of the syllable *pu*; upon presentation of the second instance of *pu*, a CA1-like structure might enter an attractor state as the current sensory input matches an existing memory representation. In contrast, for items like *bapu*, the second element has no corresponding existing memory representation, and thus does not activate an attractor in a CA1-like structure. Hence, a readout mechanism to a CA1-like structure could, in principle, act as a repetition-detector.

However, there are a number of problems with such an architecture as well. For example, after exposure to both *pupu* and *bapu*, the model might classify *bapu* as a repetition, because the syllable *pu* has an existing memory representation from a previous item. In other words, the model would show proactive interference in sameness-detection, and there is no evidence that this might be the case in real learners.

Likewise, such a model will face difficulties discriminating *ABB* patterns as in *pulili* from *ABA* patterns as in *pulipu*. It is also unclear whether a memory-based repetition-detector can detect the sameness of simultaneously (rather than sequentially) presented items, and whether such models would "recognize" an item in the presence of distractors; after all, a CA1-like region would receive input from the distractors as well, which will bring up the total level of activity. As a result, considerable computational and neuroscientific research is needed to decide whether such an architecture might act as a repetition-detector.

Similar problems arise Engel and Wang's (Engel & Wang, 2011) model.

First, the model would show proactive interference, and falsely detect repetitions in non-repetition sequences if the second item has been placed in working memory on a considerable number of earlier occasions. While this is an empirical prediction, it seems, at first sight, implausible.[2] Second, it is unclear whether a memory-based repetition-detector would be able to detect the sameness of simultaneously presented items. Third, Engel and Wang (2011) use supervised training to teach units to subtract the activation of sensory+WM neurons and sensory-only neurons, respectively. However, in all experiments on repetition learning in infancy, learning is unsupervised; further, the reliance on supervised training prevents the model from generalizing to items that are dissimilar from those it has been trained on (Marcus, 1998a, 1998b).

There is another reason for which such memory-based repetition-detectors are unlikely to support the kinds of generalizations reviewed here. Given how widespread the ability to compute repetition-patterns is, one would expect it to rely on fairly simple circuits. However, these memory-based models rely on the interaction of different brain areas (the entorhinal cortex as well as CA1 and CA3 in the case of Hasselmo and Wyble's (1997) model, and a sensory as well as a working memory system in the case of Engel and Wang's (2011) model).

[2] Engel and Wang's (2011) model can detect matches between the *A* items in *ABBA* trials. However, they achieve this by assuming that the WM subnetwork receives sensory input only when the input is attentionally encoded. As a result, only the first *A* from the *ABBA* items ever reaches WM. However, this would predict that participants do not notice the repetition of the *B* items. It thus seems that the WM component in Engel and Wang's (2011) has a similar function as (pre) frontal regions in the recent models of inhibition (Egner & Hirsch, 2005; Erika-Florence, Leech, & Hampshire, 2014; Hampshire & Sharp, 2015): it serves to highlight task-relevant representations.

SI1.5. Cope et al. (2018)

Cope et al. (2018) proposed a model to explain the successful performance of bees in delayed-match-to-sample tasks such as in Giurfa, Zhang, Jenett, Menzel, and Srinivasan (2001). They used a model inspired by the architecture of the bee mushroom body. At a conceptual level, the model comprises three populations of neurons: (1) a population of input neurons encoding stimuli (inspired by Kenyon cells); (2) a population of inhibitory neurons (inspired by the protocerebellar tract); (3) and a population of output neurons (inspired by extrinsic neurons), half of which code for a "go" response and half for a "no-go" response.

The input population has excitatory connections (with fixed weights) to both the output neurons and the inhibitory neurons; the inhibitory neurons project to the output neurons as well, but, critically, with weights that are modifiable.

The critical assumption of the model is (an empirically observed) "fatigue effect" in the input neurons: responses to repeated stimuli are weaker than to novel stimuli. As these weaker activations are assumed to be insufficient to drive the inhibitory population, novel and repeated stimuli play different roles in match-to-sample tasks and non-match-to-sample tasks, respectively.

In *match-to-sample tasks*, repeated items fail to activate the inhibitory neurons. As a result, the connection weight between the inhibitory neurons and the output neurons is adjusted only when non-matching, novel items are presented. Given that "go" responses to non-match items are not reinforced in match-to-sample tasks, the strength of the connections between the inhibitory neurons and *go* responses is increased relative to the strength of the connections between the inhibitory neurons and *no go* responses. (The strength of the connections between the inhibitory neurons and *no go* responses does not change as no learning takes place if the bee refuses to "go" for a stimulus to begin with.)

In *non-match-to-sample tasks*, repeated items still fail to activate the inhibitory neurons, so that learning occurs only with non-matching, novel items. However, in *non-match-to-sample tasks*, connections between the inhibitory neurons and *go* responses are weakened, relative to the connections between the inhibitory neurons and *no go* responses.

In other words, the inhibitory population learns to select between go and no-go responses, based on the frequency with which the responses are responses are reinforced when it is activated by novel, non-matching stimuli. It thus detects the correlation between the presence of rewards and input from non-repeated stimuli.

Impressively, these simple computational principles are sufficient to allow the model toe generalize the sameness-relations to untrained items; for example, if the model is trained in delay (non-) match-to-sample task with, say, orientations, it would transfer this learning to a task with, say, colors.

However, there are four situations that raise the question of whether this model would appropriately account for sameness-detection in grammar-learning situations. First, it is unclear to what extent this model can discriminate matching from non-matching pairs when the elements of the pair are presented simultaneously (Martinho & Kacelnik, 2016). This is because the model relies on a decrease in representational strength of items presented repeatedly, and, if identical items are presented simultaneously, no such decrease can occur (though this issue might be solved if organisms attend to the items sequentially).

Second, and critically, humans and some other animals can learn sameness-relations from positive evidence alone, in the absence of reinforcement (Marcus, Vijayan, Rao, & Vishton, 1999).

Third, and relatedly, Cope et al.'s (2018) model learns in a fundamentally different way from humans. Specifically, the model learns about *non*-matching items. In match-to-sample tasks, it learns to increase the inhibition of "go" responses to non-match stimuli; in non-match-to-sample tasks, it learns to decrease the inhibition of "go" responses to non-match stimuli. In contrast, humans learn predominantly about sameness rather than difference relations, and, to the extent that they represent difference relations, they represent them as negations of sameness relations (Hochmann, Carey, & Mehler, 2018; Hochmann, Mody, & Carey, 2016).

Fourth, the model does not produce representations of sameness or differences that can be used for further processing. For example, in Marcus et al.'s (1999) discrimination between *AAB* and *ABB*, the critical distinction was not whether the strings contained a repetition, but rather *where* in the strings the repetition was located. As a result, learners had to bind the output of the sameness detection computations to some kind of representation of sequential positions, which seems beyond the representations produced by Cope et al.'s (2018) model.

References

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in Cognitive Sciences*, *5*(3), 119–126.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*(1), 54–115. http://dx.doi.org/10.1016/S0734-189X(87)80014-2

Cope, A. J., Vasilaki, E., Minors, D., Sabo, C., Marshall, J. A. R., & Barron, A. B. (2018). Abstract concept learning in a simple neural network inspired by the insect brain. *PLoS Computational Biology*, *14*(9), e1006435. https://doi.org/10.1371/journal.pcbi.1006435

Downing, P. E. (2000). Interactions between visual working memory and selective attention. *Psychological Science*, *11*(6), 467–473. https://doi.org/10.1111/1467-9280.00290

Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, *8*(12), 1784–1790. https://doi.org/10.1038/nn1594

Engel, T. A., & Wang, X.-J. (2011). Same or different? A neural circuit mechanism of similarity-based pattern match decision making. *Journal of Neuroscience*, *31*(19), 6982–6996. https://doi.org/10.1523/JNEUROSCI.6150-10.2011

Erika-Florence, M., Leech, R., & Hampshire, A. (2014). A functional network perspective on response inhibition and attentional control. *Nature Communications*, *5*, 4073. https://doi.org/10.1038/ncomms5073

Fan, J. E., & Turk-Browne, N. B. (2016). Incidental biasing of attention from visual long-term memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(6), 970–977. https://doi.org/10.1037/xlm0000209

Giurfa, M., Zhang, S., Jenett, A., Menzel, R., & Srinivasan, M. V. (2001). The concepts of "sameness" and "difference" in an insect. *Nature*, *410*(6831), 930–933. https://doi.org/10.1038/35073582

Grill-Spector, K., Henson, R. N., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn Sci*, *10*(1), 14–23. https://doi.org/10.1016/j.tics.2005.11.006

Hampshire, A., & Sharp, D. J. (2015). Contrasting network and modular perspectives on inhibitory control. *Trends in Cognitive Sciences*, *19*(8), 445–452. https://doi.org/10.1016/j.tics.2015.06.006

Hasselmo, M. E. (2005). The Role of Hippocampal Regions CA3 and CA1 in Matching Entorhinal Input With Retrieval of Associations Between Objects and Context: Theoretical Comment on Lee et al. (2005). *Behavioral Neuroscience*, *119*(1), 342–345.

Hasselmo, M. E., & Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus: simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, *89*(1–2), 1–34.

Hochmann, J.-R., Carey, S., & Mehler, J. (2018). Infants learn a rule predicated on the relation same but fail to simultaneously learn a rule predicated on the relation different. *Cognition*, *177*, 49–57. https://doi.org/10.1016/j.cognition.2018.04.005

Hochmann, J.-R., Mody, S., & Carey, S. (2016). Infants' representations of same and different in match- and non-match-to-sample. *Cognitive Psychology*, *86*, 87–111. https://doi.org/10.1016/j.cogpsych.2016.01.005

Johnson, J. S., Spencer, J. P., Luck, S. J., & Schöner, G. (2009). A dynamic neural field model of visual working memory and change detection. *Psychological Science*, *20*(5), 568–577. https://doi.org/10.1111/j.1467-9280.2009.02329.x

Kumaran, D., & Maguire, E. A. (2007). Which computational mechanisms operate in the hippocampus during novelty detection? *Hippocampus*, *17*(9), 735–748. https://doi.org/10.1002/hipo.20326

Lisman, J. E. (1999). Relating hippocampal circuitry to function: recall of memory sequences by reciprocal dentate-CA3 interactions. *Neuron*, *22*(2), 233–242.

Lisman, J. E., & Otmakhova, N. A. (2001). Storage, recall, and novelty detection of sequences by the hippocampus: elaborating on the SOCRATIC model to account for normal and aberrant effects of dopamine. *Hippocampus*, *11*(5), 551–568. https://doi.org/10.1002/hipo.1071

Ludueña, G. A., & Gros, C. (2013). A self-organized neural comparator. *Neural Computation*, *25*(4), 1006–1028. https://doi.org/10.1162/NECO_a_00424

Marcus, G. F. (1998a). Can connectionism save constructivism? *Cognition*, *66*(2), 153–182.

Marcus, G. F. (1998b). Rethinking eliminative connectionism. *Cognit Psychol*, *37*(3), 243–82.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80.

Martinho, A., & Kacelnik, A. (2016). Ducklings imprint on the relational concept of “same or different.” *Science*, *353*(6296), 286–288. https://doi.org/10.1126/science.aaf4247

Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136–1139. https://doi.org/10.1126/science.aah7011

Stokes, M. G. (2015). “Activity-silent” working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Wen, S., Ulloa, A., Husain, F., Horwitz, B., & Contreras-Vidal, J. L. (2008). Simulated neural dynamics of decision-making in an auditory delayed match-to-sample task. *Biological Cybernetics*, *99*(1), 15–27. https://doi.org/10.1007/s00422-008-0234-0