# Two-in-one: a circuit optimised for recollection predicts recognition behavior

**Cristina Savin (csavin@nyu.edu)**
Center for Neural Science & Center for Data Science, NYU
4 Washington Place, New York, NY, 10003, USA

**Máté Lengyel (m.lengyel@eng.cam.ac.uk)**
Dept. of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK

Dept. of Cognitive Science, CEU
7 Október 6 utca, Budapest H-1051, Hungary

## Abstract

**Curvilinear receiver operating characteristic (ROC) curves are a classic signature of recognition memory in humans and animals. They have been traditionally explained in terms of two competing models that lead to different interpretations of recognition as either unitary or resulting through the combination of two independent processes. Neither model class specify neural mechanisms or account for all salient aspects of the behavioral data. Here we propose an alternative computational account of recognition memory that can reconcile these seemingly incompatible views. In our model, recognition arises due to two *interacting* subsystems, optimised for efficient recollection. We show that our model yields a hybrid pattern of recognition memory behavior, which combines aspects of both traditional models. Additionally, it provides a functional interpretation of these features rooted in the physiological properties of the neural circuits supporting memory recall.**

**Keywords:** recognition memory, perirhinal cortex, hippocampus, Bayesian decision theory

## Introduction

Introspection provides strong intuitions that recognition can come about either due to recollection ('here comes my old school buddy') or due to a vague sense of familiarity, without any recollective detail ('I'm sure I've met this person before, but I have no idea when and why'). However, it is hotly debated whether or not distinct processes underlie recognition, and what this implies about true and false familiarity judgements, as routinely measured by receiver operating characteristic (ROC) curves (Yonelinas, 2002; Wixted & Squire, 2011).

Dual module (DM) theories posit that recollection and familiarity are independent, with the former an all-or-none process and the latter well described by traditional signal detection theory. In contrast, single module (SM) theories assume that recognition relies on a single graded decision variable, with higher variance for familiar compared to novel items. This variable could arise unitarily, or by combining familiarity and recollection cues. DM is supported by the fact that certain experimental manipulations preferentially affect one of the two components, e.g. localized hippocampal lesions disrupt recollection but leave familiarity relatively unaffected (Fortin et al., 2004), something that SM cannot explain. However, quantitative model comparison favors SM in all but few experiments (with linear ROCs) (Wixted & Squire, 2011). Despite over 30 years of recognition memory research, no consensus exists about which class of models provides a more satisfying account for the data. Notably, neither model has been instantiated at the circuit level.

Here we resolve this impasse by considering a ubiquitous property of memory traces, namely that their strength is not uniform (e.g. due to recency, fluctuations in attention, depth of processing), which puts fundamental constraints on memory recall. We develop a neural circuit architecture that can efficiently operate in the face of trace strength-ambiguity, and show that it naturally combines features of both SM and DM models, thus providing a unified account for a broad range of behavioral and neural observations.

## Methods and results

We cast competing theories within the framework of Bayesian decision theory with the decision variable reflecting the log posterior odds $\ell = \log \frac{\mathrm{P}(y=1|D)}{\mathrm{P}(y=0|D)}$ that a test item is familiar $(y = 1)$ or novel $(y = 0)$ given an internal representation $D$. ROCs, plotting the rate of hits against the rate of false alarms, are traced by varying the relative cost of misses and false alarms, which is equivalent to varying a threshold on $\ell$ (Dayan & Abbott, 2001). This is in contrast to classical signal detection theory-based accounts, which would operate by placing a threshold directly on $D$, which is neither optimal, nor flexible enough to accommodate a non-scalar $D$ – a point to which we return below. By reformulating traditional SM and DM models within this framework, we show that they imply different distributions over $\ell$, $\mathrm{P}(\ell|y = 0)$ and $\mathrm{P}(\ell|y = 1)$ (Fig. A, top), which give rise to the kind of ROCs classically attributed to these models: curvilinear ROCs (that are asymmetric such that high hit rates are possible even for near-zero false alarm rates) for both models, with linear z-ROC for SM and u-shaped z-ROC for DM (where a z-ROC is obtained by transforming both axes of a ROC via the inverse normal c.d.f.; Fig. A, bottom).

Critically, the Bayesian framework also allows us to consider non-scalar internal representations $D$, which in turn opens the way to studying the neural underpinnings of recognition memory. Specifically, we consider memory items represented as patterns of neural activity $\mathbf{x}$, encoded (stored) in the synapses $\mathbf{W}$ of a neural circuit via synaptic plasticity (see generative model in Fig. B, top left). Novel and familiar items differ in memory strength $s$: novel items have $s \approx 0$, while $s$ can vary across familiar items (see above). Recognition then becomes a hierarchical inference process mapping the noisy cues $\tilde{\mathbf{x}}$, and the information stored in synapses $\mathbf{W}$, into a posterior over the binary $y$ (familiar / novel). The optimal neural circuit implementation of this inference proceeds by first inferring the original item, $\mathbf{x}$, and its corresponding strength, $s$ and then computing the posterior log odds ratio, $\ell$. We found that even if only recollection performance is considered, the first step is most efficiently implemented by having two functionally distinct subregions for familiarity and recollection which interact recursively to compute the joint posterior $\mathrm{P}(\mathbf{x}, s|D)$ (Fig. B, top right), rather than an alternative monolithic architecture which only consists of a recollection module, implicitly marginalising out the unknown memory strength to obtain $\mathrm{P}(\mathbf{x}|D)$ (Fig. B, bottom left). In the optimal dual system, memory strength influences pattern completion because it modulates the reliability of synaptic information and thus the relative weighting of the two sources of information $\tilde{\mathbf{x}}$ and $\mathbf{W}$. Conversely, the denoised pattern is used to better estimate $s$, which provides evidence for whether or not an item is familiar or novel (Fig. B, bottom right).

For a wide range of model parameters, the neural circuit derived ROCs are curvilinear, with linear z-ROCs (Fig. C, red), similar to SM. The patterns of modulation of the ROC can be traced back to properties of optimal inference in our generative model. The novel item distribution always has negative mean and low variance. In contrast, the mean ($\geq 0$) and variance of the summary statistic $\ell$ increases with memory strength, resulting in a broader $\ell$ distribution for familiar items. $\mathrm{P}(\ell|y=1)$ becomes bimodal when $\mathrm{P}(s|y=1)$ has a heavy tail (Fig. C, blue), similar to DM. In summary, the neural model recapitulates the phenomenology of recognition memory behavior, mechanistically resembling either SM or DM depending on the details of the behavioral paradigm. The model also reproduces changes in ROCs due to lesions (Fig. D), and other puzzling behavioral (e.g. the increase in false recognition after perirhinal lesions; (McTighe et al., 2010), not shown here), and neural observations (e.g. the bi-directional anatomical connectivity between hippocampus and perirhinal cortex, and heterogeneity in the tuning of perirhinal neurons).

## Conclusions

We have shown that a neural circuit implementation of efficient recollection accounts for a rich set of experimental data on recognition memory, reconciling seemingly conflicting, well-entrenched recognition memory models. The critical step is asking why have two systems in the first place: in the face of (unavoidable) trace strength variability two functionally separate but interacting modules are critical for efficient recollection.

## Acknowledgments

This work was supported by the Wellcome Trust (CS, ML).

## References

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience* (Vol. 806). Cambridge, MA: MIT Press.

Fortin, N. J., Wright, S. P., & Eichenbaum, H. (2004). Recollection-like memory retrieval in rats is dependent on the hippocampus. *Nature*, *431*(7005), 188–191.

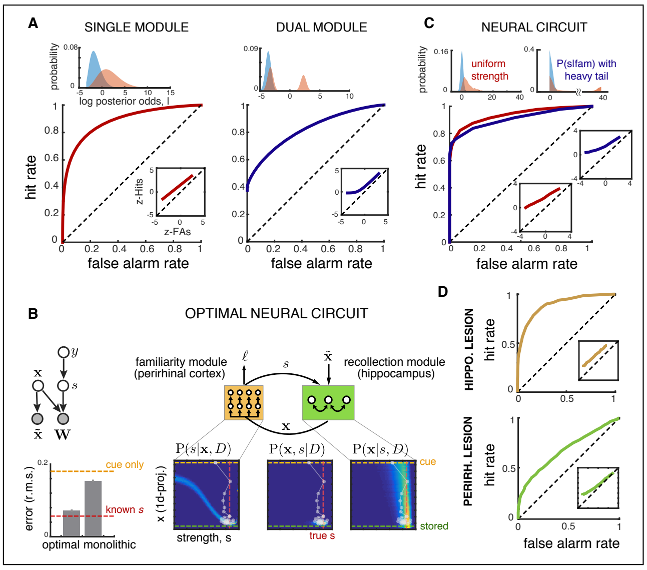

Figure 1: **A.** Bayesian reinterpretation of SM and DM: distribution of summary statistic $\ell$ for targets (red) and lures (blue) and corresponding (z-)ROCs. **B.** Top left: Generative model for recognition memory (top left). Right: optimal neural circuit architecture (top): bi-directionally connected familiarity (2-layer feedforward network, $s$ encoded in second layer) and recollection modules (similar to autoassociative memory network in (Savin et al., 2014)); dynamics sample from the joint posterior $\mathrm{P}(\mathbf{x}, s|D))$ by alternating between sampling from $\mathrm{P}(\mathbf{x}|s, D))$ and $\mathrm{P}(s|\mathbf{x}, D))$ (bottom), with $\ell$ determined from $s$ samples (marginalizing $\mathbf{x}$). Bottom left: recollection performance for the optimal and a monolithic architecture. **C.** Same as **A**, with different assumptions for $\mathrm{P}(s|y=1)$. **D.** Neural circuit (z-)ROCs after different lesions.

Fusi, S., & Abbott, L. F. (2007). Limits on the memory storage capacity of bounded synapses. *Nature Neuroscience*, *10*(4), 485–493.

McTighe, S. M., Cowell, R. A., Winters, B. D., Bussey, T. J., & Saksida, L. M. (2010). Paradoxical false memory for objects after brain damage. *Science*, *330*(6009), 1408–1410.

Savin, C., Dayan, P., & Lengyel, M. (2011). Two is better than one: distinct roles for familiarity and recollection in retrieving palimpsest memories. *Advances in Neural Information Processing Systems (NIPS)*, *24*, 1305–1313.

Savin, C., Dayan, P., & Lengyel, M. (2014). Optimal Recall from Bounded Metaplastic Synapses: Predicting Functional Adaptations in Hippocampal Area CA3. *PLoS Computational Biology*, *10*(2), e1003489.

Wixted, J. T., & Squire, L. R. (2011). The medial temporal lobe and the attributes of memory. *Trends in Cognitive Sciences*, *15*(5), 210–217.

Yonelinas, A. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of memory and language*, *46*(3), 441–517.