Experiencing surprise: the temporal dynamics of its impact on memory

Darya Frank[1,2], Alex Kafkas[1], Daniela Montaldi[1]

[1] Division of Neuroscience and Experimental Psychology, School of Biological Sciences, The University of Manchester, UK
[2] Laboratory for Clinical Neuroscience, Centre for Biomedical Technology, Universidad Politécnica de Madrid, Spain

Acknowledgments:
This work was supported by The University of Manchester President's Doctoral Scholar award (D.F.) and Wellcome Trust Programme Grant Number 094597. We thank Oliver Gray and Paul Thompson for helpful discussions.

Abstract

To efficiently process information, the brain shifts between encoding and retrieval states, prioritising bottom-up or top-down processing accordingly. Expectation violation before or during learning has been shown to trigger an adaptive encoding mechanism, resulting in better memory for unexpected events. Using fMRI we explored (1) whether this encoding mechanism is also triggered during retrieval, and if so, (2) what the temporal dynamics of its mnemonic consequences are. Participants studied object images, then, with new objects, they learned a contingency between a cue and a semantic category. Rule-abiding (expected) and violating (unexpected) targets and similar foils were used at test. When an unexpected event was encountered at test, the triggering of the adaptative encoding mechanism impaired memory. However, when an expected event followed a similar but unexpected event, its performance was boosted, underpinned by activation in the hippocampus, midbrain, and occipital cortex. In contrast, a sequence of two unexpected similar events also triggered occipital engagement, however, this was found to be detrimental to memory performance. Taken together, our findings suggest that encountering surprising events engages an encoding mechanism, supported by bottom-up processing, that undermines memory for that event, even when the goal is to retrieve.

Abstract word count: 194

Introduction

Our environmental setting and critically, how we interact with it, shape how we perceive, encode, and remember our daily life, which is in turn influenced by what we expect to encounter. In some situations, our priority or dominant goal, is to encode information, whereas in others, it is to retrieve information. Indeed, in order to efficiently process the inputs it receives, the brain shifts between top-down and bottom-up information streams to balance the processing of sensory inputs versus the utilisation of stored information. In support of this has been the identification of distinct encoding and retrieval states, orchestrated by hippocampal theta rhythm (Buzsáki, 2002). Bottom-up processing can therefore support the hippocampal encoding state, prioritising the transformation of perceptual inputs into memories, while top-down processing supports a retrieval state by facilitating access to stored information. The specific mechanisms that govern these encoding-retrieval shifts, which may or may not function competitively, are not well understood, but will be driven, at least in part, by the factors that induce each memory state. Whilst there is ample evidence that an adaptive encoding state is triggered upon experiencing discrepancies between expected and encountered events (Axmacher et al., 2010; Bein et al., 2020; Kafkas and Montaldi, 2015; Kumaran and Maguire, 2007; Meeter et al., 2004), it remains unclear whether such an encoding state might also be engaged when the dominant goal is to retrieve, rather than encode. Furthermore, it is often assumed that memory is benefitted by the triggering of this adaptive encoding mechanism, however, it remains to be established whether a memory advantage is still observed when we encounter surprising information while attempting to retrieve. Here, therefore, we first examine whether surprise produced by expectation violation taking place at retrieval (during, but orthogonal to, a recognition task), results in a shift towards an encoding state, accompanied by a gain in bottom-up processing; and second, whether this occurs at the expense of retrieval. We then investigate the extended consequences of this expectation violation, and the potential encoding-retrieval shifts it might generate, by examining how surprise at retrieval, and its ensuing bottom-up processing, modulate the trial-by-trial recognition of subsequent similar events.

Predictive processing frameworks postulate that bottom-up processing is modulated by top-down predictions (Friston, 2005). In memory research, surprise, produced by expectation violation, has been shown to engage hippocampal processing (Duncan et al.,

2012a; Frank et al., 2020; Hindy et al., 2016; Kafkas and Montaldi, 2015; Kumaran and Maguire, 2007; Long et al., 2016), which together with the midbrain dopaminergic system (Kafkas and Montaldi, 2018a; Lisman and Grace, 2005; Shohamy and Wagner, 2008), supports adaptive memory formation. To date, evidence for this adaptive mechanism, showing better memory performance for unexpected information, comes mostly from paradigms in which expectation violation takes place before or during learning (Frank and Kafkas, 2021; Garrido et al., 2015; Greve et al., 2017a; Gruber et al., 2013; Kafkas, 2021; Li et al., 2003; Long et al., 2016). Therefore, any additional resources diverted towards encoding in these scenarios is likely to boost later memory performance, similar to attention enhancement effects (Aly and Turk-Browne, 2017). However, to demonstrate the ubiquity of the beneficial nature of this adaptive encoding mechanism, it is critical to provide evidence of the benefits of a dynamic shift towards encoding when retrieval, instead of encoding, is necessary. This is also more likely to shed light on the effects of the rapid real-life interplay between encoding and retrieval states. Under these circumstances, a shift towards encoding still serves an adaptive purpose, but it might no longer be 'beneficial' in terms of retrieving an accurate memory of the unexpected event. Confirming whether such potentially 'detrimental' encoding effects (e.g., expectation-driven encoding at the expense of retrieval) occur is important to our overall understanding of the modulating role of expectation violation, or surprise, on memory.

Given the malleability, triggered by expectation violation (Kim et al., 2014), of existing memories, it becomes important to consider what the mnemonic consequences of a shift towards encoding may be. Should expectation violation trigger an encoding state, which comes at the expense of a memory search, it might result in poor recognition of the unexpected event at retrieval. This is consistent with the view that the hippocampus continuously shifts between encoding and retrieval states (Buzsáki, 2002; Colgin, 2016; Hasselmo et al., 2002; Ketz et al., 2013). This system can be likened to a pendulum swinging rhythmically between the two memory states, optimising information processing. If, however, situational factors disrupt the on-going swing between states, favouring one over another, the result can be an unpredictable change in memory efficiency. Previous research has shown that an explicit strategy cue (e.g. to encode or integrate) can indeed foster a hippocampal trade-off between the encoding and retrieval of information (Richter et al., 2016). Here, instead, we test whether such shifts can also

be triggered by implicit, task-irrelevant, expectation violations, and we explore the continuous mnemonic consequences of such shifts, in the form of correct recognition of targets and rejection of similar foils.

The engagement of an encoding state should also be demonstrated through enhanced perceptual processing, as encoding relies on sensory inputs. Previous research has found increased activity along the ventral visual stream in response to novelty and expectation violation (Hawco and Lepage, 2014; Kafkas and Montaldi, 2015, 2014; Stoppel et al., 2009). We have also shown that expectation-modulated memory is sensitive to the perceptual similarity between stored representations and current inputs (Frank et al., 2020). Therefore, in order to tap the proposed encoding mechanism engaged by expectation violation, a high level of perceptual load, achieved by using sets (target and similar foils) of highly similar events, is necessary. This allows us to examine how expectation violation might affect trial-by-trial memory representations and recognition decisions (Duncan et al., 2012b; Kafkas and Montaldi, 2018b). For example, encountering an unexpected event might shape the way a subsequent similar event is processed. Therefore, we can better tap the interplay between memory retrieval processes and the hypothesised expectation-driven shift between retrieval and encoding, and how this interplay affects subsequent processing and recognition of similar events.

In two experiments, we examined whether expectation violation spontaneously engages an encoding state, and if so, whether this response is dependent on the perceptual similarity between inputs. We also considered the sustained mnemonic consequences of such shifts on current and subsequent recognition decisions. To do so, we used a paradigm in which expectation was manipulated at retrieval, implicitly and independently of the requirements of the recognition task (see Figure 1 for experimental design). Following the encoding of object images, participants performed a rule-learning task in which they learned a contingency between a cue and the object's category (man-made or natural). Then, at retrieval, the same cues were presented followed by an old (target) or new (parametrically manipulated similar foils, F1, F2, F3) object for which participants were asked to make an old/new recognition decision. In one-third of the retrieval trials, the pre-established cue and object category contingency was violated. This design allowed us to answer two important questions; first whether expectation

violation engages a bottom-up encoding mechanism, even when it is not goal-relevant. Second, what the temporal dynamics of the mnemonic consequences of the expectation status are, for an object and its subsequent similar objects (i.e., interactions between previous and current set events' expectation status). We reasoned that expectation-driven encoding could harm online retrieval processes, but aid recognition of subsequent similar events.

Results

To assess expectation-modulated encoding, and its sensitivity to perceptual similarity, we used mixed-effects logistic regression to model the response to each set event (correct/incorrect recognition) as a function of the preceding event from that set (e.g. if first an F1 is presented, and later on a target, these would be noted $F1_{prev}$ $Target_{curr}$). Furthermore, each set event was associated with an expectation status determined by the preceding cue (rule-abiding events were marked as expected rule-violating as unexpected). Each model also included an interaction between current and previous set events' expectation status to examine dynamic changes (e.g. how a previous unexpected foil affects recognition of a current expected target; see Figure 1c, and Methods for full details).

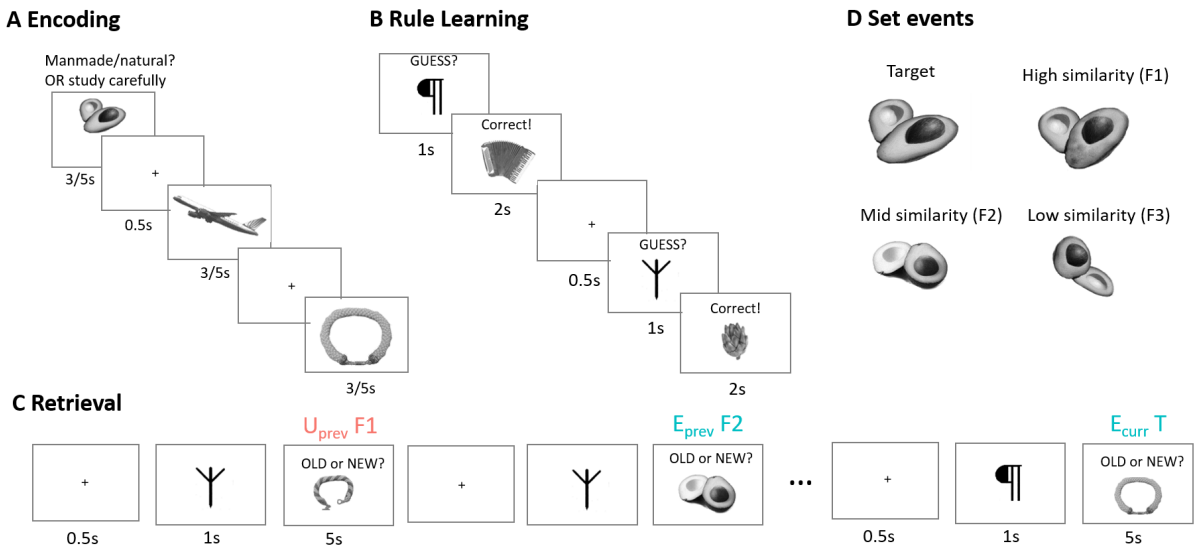


Figure 1 | Experimental design. A) During the first round of encoding, participants responded whether the object was man-made or natural. In the second encoding round, participants are asked to study the object carefully. In Experiment 2, each object appeared three times consecutively; in addition to the two presentations from Experiment 1, participants were asked to respond whether the object was more likely to be found indoors or outdoors. The order of man-made/natural and in/outdoor questions was randomised, the third presentation was always 'study carefully' B) In the rule-learning task participants learned a contingency between a cue and an object's category, man-made or natural. Participants had 1 second to make a decision in Experiment 1, and 3 seconds in Experiment 2. C) In the retrieval task, the same

cues were presented before each set event, old objects (targets) and new similar foils (F1, F2, F3 in Experiment 1; F1 and F2 in Experiment 2). On 30% of these trials there was a mismatch between the cue and the object's category, these are unexpected trials (e.g., a cue for a natural object is followed by a man-made object, marked in red). Participants were instructed to indicate whether the object is old or new. $U_{prev}F1$, *EcurrT* = Unexpected F1 presented before an expected target, from the same set. D) Example of range of perceptual similarity within event sets.

Experiment 1

Memory for targets is modulated by preceding unexpected similar set events

For targets following F1 set events, we found significant main effects of the F1 expectation ($\beta = 0.57$, $X^2(1) = 6.8$, $p = 0.009$), as well as a significant interaction between the target's expectation and the previous F1's expectation status ($\beta = -0.735$, $X^2(1) = 4.14$, $p = 0.042$; Figure 2 left panel). All other main effects were not significant (all p's > 0.585). Subsequent contrast tests revealed that expected targets were more likely to be remembered following an unexpected F1, compared to an expected F1 set event ($z = 2.59$, $p = 0.009$). Furthermore, when the previous F1 set event was unexpected, the subsequent expected targets were more likely to be correctly remembered, compared to unexpected targets ($z = 2.39$, $p = 0.016$). When examining targets that followed F2 and F3 set events, we did not observe any significant predictors (all p's > 0.127).

Correct rejection of similar foils is modulated by the expectation status of preceding targets

For F1 events following targets, we found a trend towards an interaction between the two events' expectation status ($\beta = -0.58$, $X^2(1) = 3.59$, $p = 0.058$), with contrasts showing more correct rejections of F1 events (CR1), for expected F1 than unexpected F1 events following unexpected targets ($z = 2.29$, $p = 0.022$; Figure 2 middle panel). All other effects were not significant (all p's > 0.255). When examining F1 events that followed F2 and F3 set events, we did not observe any significant predictors (all p's > 0.196).

Expectation sequence modulation decreases as perceptual similarity decreases

F2 events were not affected by previous targets (all p's > 0.186), F1s (all p's > 0.142), or F3s (all p's > 0.124) from the same set. Similarlary, responses to F3 events were not modulated by any preceding set events (targets: all p's > 0.523, F1s: all p's > 0.829, F2s: all p's > 0.301).

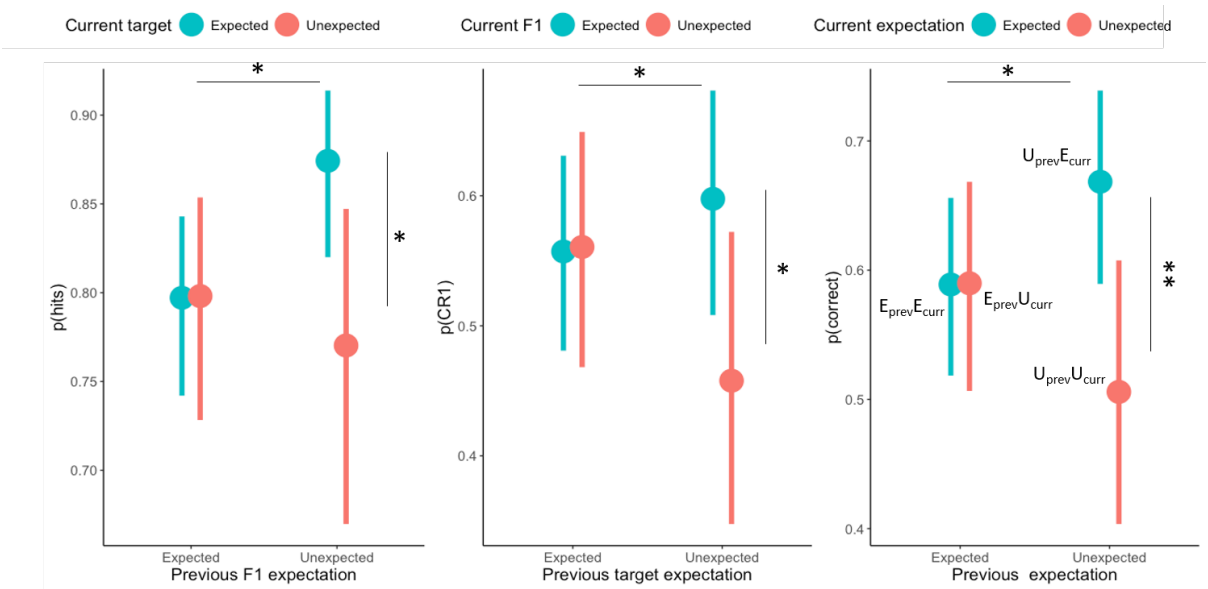


Figure 2 | Experiment 1 Results. Left: Predicting hits. More hits were observed for expected targets following unexpected F1 foils, compared to unexpected targets following unexpected F1, and compared to expected targets following expected F1. **Middle: Predicting CR1.** More CR1 for expected F1 following unexpected targets, compared to unexpected F1 following unexpected targets. **Right: Collapsed hits and CR1.** Similar results showing an interaction between the previous and current set events' expectation status, with $U_{prev}E_{curr}$ events showing the best memory performance, compared to $U_{prev}U_{curr}$ and $E_{prev}E_{curr}$. Unless otherwise stated, all error bars represent 95% confidence intervals. * $p < 0.05$, ** $p < 0.01$.

Targets and their most similar foils show identical expectation sequence effects

Given the similar effects observed for targets and F1 events independently, we collapsed the two, to examine whether these effects are complementary (i.e., whether there is an interaction between current and previous expectation status; Figure 2 right panel). Although hits and correct rejections are not necessarily products of the same mnemonic process, in this paradigm they provide an opportunity to examine how perceptual load (in the form of similarity) interacts with dynamic expectation modulations. The mnemonic comparison between the current set event (target or F1) and its previous set event (F1 or target, respectively), forms the highest load, or interference, in relation to the encoded object, as the participant makes a recognition decision. Therefore, if perceptual processes are engaged, upon encountering an unexpected event, the effects observed for each set event individually should replicate. To eliminate any effects driven by memory strength differences, we only included in this model, events for which the previous response was correct (i.e. for targets following F1 events, we only included targets that followed CR1). Similar results were observed when including all trials in the model (see Supplementary Figure S.3). We found a significant main effect of previous

event's expectation status (β = 0.341, $X^2(1)$ = 4.9, p = 0.027), as well as a significant interaction between expectation status of the current and previous events (β = -0.682, $X^2(1)$ = 6.5, p = 0.01). Subsequent contrast tests revealed that when the previous set event was unexpected, more correct responses were observed for expected compared to unexpected events ($U_{prev}E_{curr} > U_{prev}U_{curr}$; z = 3.1, p = 0.002). Additionally, there were more correct responses for current expected events following unexpected than expected events ($E_{prev}E_{curr} < U_{prev}E_{curr}$; z = -2.21, p = 0.027).

Taken together, these results suggest expectation violation shifts processing away from retrieval and towards encoding, and that the temporal dynamics of the mnemonic consequences of this shift are reflected in the memory for the subsequent set event, as a function of perceptual similarity. When the first (previous) and second (current) events are both unexpected ($U_{prev}U_{curr}$), we observed poor accuracy for targets and F1 foils. On the other hand, when an unexpected event is followed by an expected event ($U_{prev}E_{curr}$) a boost in performance was observed for the current event. To examine whether these effects engage the circuit involved in adaptive memory formation (including the hippocampus and midbrain (Kafkas and Montaldi, 2018a; Shohamy and Adcock, 2010), as well as to test the hypothesis that expectation violation engages an encoding mechanism, supported by the bottom-up information stream (ventral visual pathway), in Experiment 2 a new set of participants performed a similar task while fMRI data was acquired (see Methods for minor task adjustments).

Experiment 2

Behavioural Results

Replicating the effects observed in Experiment 1, we found a main effect of the previous set event's expectation status (β = 0.309, $X^2(1)$ = 4.13, p = 0.042), as well as an interaction between the current and previous events' expectation status (β = -0.59, $X^2(1)$ = 4.47, p = 0.034). Subsequent contrast tests revealed better memory performance for expected events following unexpected ones, compared to those following expected events ($U_{prev}E_{curr} > E_{prev}E_{curr}$; z = 2.55, p = 0.011). For set events following unexpected ones, better memory was also found for expected compared to unexpected events ($U_{prev}E_{curr} > U_{prev}U_{curr}$; z = 2.68 p = 0.007). Next, we examined CR2 as a function of the previously seen targets and F1 events, and their expectation status. For F2 following targets, a main effect

of the target's expectation status was observed ($\beta = 0.473$, $X^2(1) = 4.4$, $p = 0.036$), with more CR2 following unexpected targets. All other effects were not significant (all p's > 0.169). Correct responses to F2 events were not affected by preceding F1 events (all p's > 0.263).

<u>fMRI results</u>

Expectation sequence interactions engage hippocampal, midbrain, and occipital regions to support subsequent mnemonic processing.

We first examined the neural correlates of the behavioural contextual expectation interaction reported above. For current expected events that followed unexpected events, compared to those following an expected event ($U_{prev}E_{curr} > E_{prev}E_{curr}$; see Figure 3b), we found increased activation in the hippocampus, bilaterally (right: x = 36, y = -33, z = -12, k = 12, SVC $p_{FWE} = 0.04$; left: x = -36, y = -45, z = -9, k = 10; see Methods for cluster specification), SN/VTA (x = 9, y = -24, z = -12, k = 11, SVC $p_{FWE} = 0.039$), and left inferior occipital gyrus (BA 18; x = -18, y = -87, z = 0, k = 26). For current unexpected events following previous expected ones, compared to those which followed an unexpected event ($E_{prev}U_{curr} > U_{prev}U_{curr}$) we also found increased activation in the right hippocampus (x = 24, y = -33, z = -6, k = 10, SVC $p_{FWE} = .045$) and left parahippocampus (x = -33, y = -45, z = -6, k = 10). Critically, in both contrasts the current set events had the same expectation status and differed only on the expectation status of the previous event.

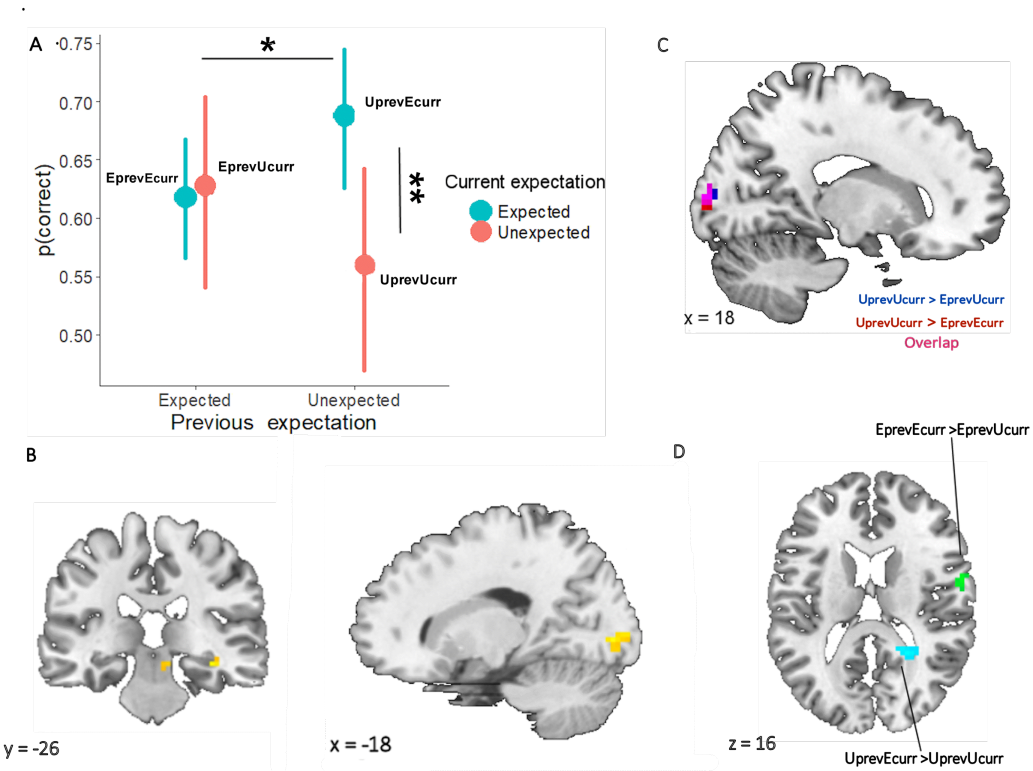


Figure 3 | Behavioural and neural expectation interactions for targets and F1 foils. A) Behavioural results. Replicating the results from Experiment 1, a current by previous expectation status interaction was observed, with $U_{prev}E_{curr}$ showing a boost in memory performance**. B) UprevEcurr > EprevEcurr contrast.** Increased activation in the right hippocampus, SN/VTA and left inferior occipital cortex. **C) Unexpected > Expected interactions.** Increased activation in right occipital cortex (BA 18) was observed for $U_{prev}U_{curr}$ > $E_{prev}E_{curr}$, $U_{prev}U_{curr}$ > $E_{prev}U_{curr}$, despite poor memory performance for $U_{prev}U_{curr}$ events**. D) Expected > Unexpected interactions.** Increased activation in right retrosplenial cortex/precuneus for $U_{prev}E_{curr}$ > $U_{prev}U_{curr}$, and increased activation in right primary motor cortex for $E_{prev}E_{curr}$ > $E_{prev}U_{curr}$.

Comparing current expected vs. unexpected events, following previously unexpected events ($U_{prev}E_{curr}$ > $U_{prev}U_{curr}$; Figure 3d) revealed activation in the right retrosplenial cortex/precuneus (x = 21, y = -45, z = 12, k = 24). The complementary contrast, following expected events, $E_{prev}E_{curr}$ > $E_{prev}U_{curr}$ (Figure 3d), was associated with increased activation in primary motor cortex (BA 4, x = 60, y = 0, z = 15, k = 11). For contrasts in which unexpected events elicited more activations than expected ones, despite showing reduced memory performance ($U_{prev}U_{curr}$ > $E_{prev}E_{curr}$ and $U_{prev}U_{curr}$ > $E_{prev}U_{curr}$; Figure 3c), we found increased activation in right occipital cortex ($U_{prev}U_{curr}$ > $E_{prev}U_{curr}$: BA 18, x = 15, y = -93, z = 9, k =25; $U_{prev}U_{curr}$ > $E_{prev}E_{curr}$: BA 18, x = 18, y = -93, z = 9, k = 36). Comparing the first (previous) event between conditions (first expected vs. first unexpected) did not reveal any significant effects (see Supplementary Figure S.5 for behavioural results of all first set events).

Expectation status differentially engages encoding and retrieval-related regions.

To explore whether expected and unexpected events, across event types, responses, and temporal positions elicited differential activations in a bottom-up (ventral visual stream) or reinstatement (recollection network) manner, we also compared the two conditions using an F contrast (see Figure 4a). We found increased activity for unexpected > expected events in right occipital cortex (BA 19, x = 39, y = -75 z = 15, k = 28; and BA 18, x = 18, y = -93, z = 12, k = 10) and right fusiform gyrus (BA 37, x = 27, y = -48, z = -18, k = 38). For expected > unexpected, we observed increased activation in the right inferior parietal lobe (angular and supramarginal gyri; BA 39 and BA 40, x = 39, y = -51, z = 30, k = 9).

Occipital activation supports the interaction between expectation and memory performance for similar events.

To unpack the overall unexpected > expected effect, we tested how contextual expectation interacted with successful recognition decisions (hits and CR), across presentation order. Whilst no differential neural responses were found for expected and unexpected hits or correct rejections of F2 events, we observed increased activation in the right inferior occipital gyrus (BA 19, x = 27, y = -78, z = -6, k = 20) for unexpected CR1 > expected CR1.

Increased perceptual load interacts with expectation status to engage ventral visual stream regions.

Finally, we examined interactions between expectation status and presentation order (across set events in Figure 4b; see Supplementary Figure S.6 for break-down per set event). We observed again increased activity in two clusters in occipital cortex (BA19, x = 33, y = -78, z = 15, k = 424, and BA 18, x = -21, y =-90, z = 15, k = 72) as well as in right fusiform gyrus (BA 37, x = 21, y = -42, z = -12, k = 25) for unexpected > expected events presented second in the set (no unexpected > expected effects were found for first or third set events).

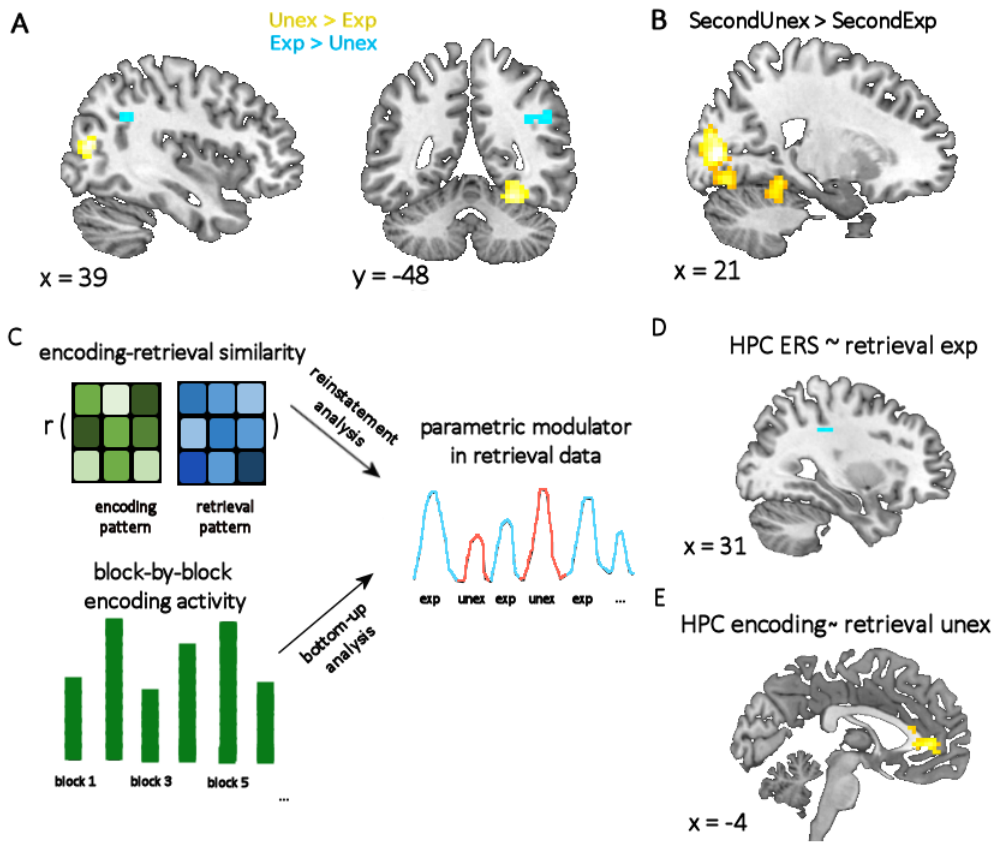


Figure 4 | Overall fMRI expectation effects. A) Main effect of expectation. Unexpected events, compared to expected ones, engaged more activation in right middle occipital cortex and right fusiform. Conversely, excepted events, compared to unexpected ones, engaged activation in right angular and supramarginal gyri. **B) Expectation by presentation order interaction.** Unexpected events presented second in the set engaged regions along the ventral visual stream. No effects were observed for first or third set events**. c) Schematic of reinstatement and encoding analyses.** Encoded-related hippocampal activity and encoding-retrieval similarity were used as parametric modulators (in separate analyses) to query retrieval activation for expected and unexpected events. **D) Reinstatement analysis results.** Activity in the right supramarginal gyrus and left middle temporal gyrus (not shown) during retrieval of expected events correlated with hippocampal ERS. No effects were observed for unexpected events. **E) Bottom-up analysis results.** ACC activation during unexpected events at retrieval correlated with hippocampal encoding activity, no effects were observed for expected events.

Hippocampal encoding activity modulates ACC processing of unexpected events at retrieval.

Adopting hippocampal encoding activity (using the bilateral anatomical mask derived from the Harvard-Oxford atlas, see Methods) as a parametric modulator (Figure 4b), we found that increased hippocampal encoding generated increased involvement of the anterior cingulate cortex (x = 1, y = 39, z = 3, k = 173; Figure 4e) for the retrieval of unexpected events. We did not observe any activation for expected events that was modulated as a function of hippocampal encoding activity.

Hippocampal reinstatement correlates with activation in recollection-related regions for expected events at retrieval.

Increased hippocampal encoding-retrieval similarity (Figure4b) was associated with increased activity in left middle temporal gyrus (x = -42, y = -36, z = -3, k = 11) and right supramarginal gyrus (x = 30, y = -30, z = 36, k = 9) during retrieval of expected events (Figure 4d). We did not observe any activation for unexpected events that correlated with hippocampal ERS.

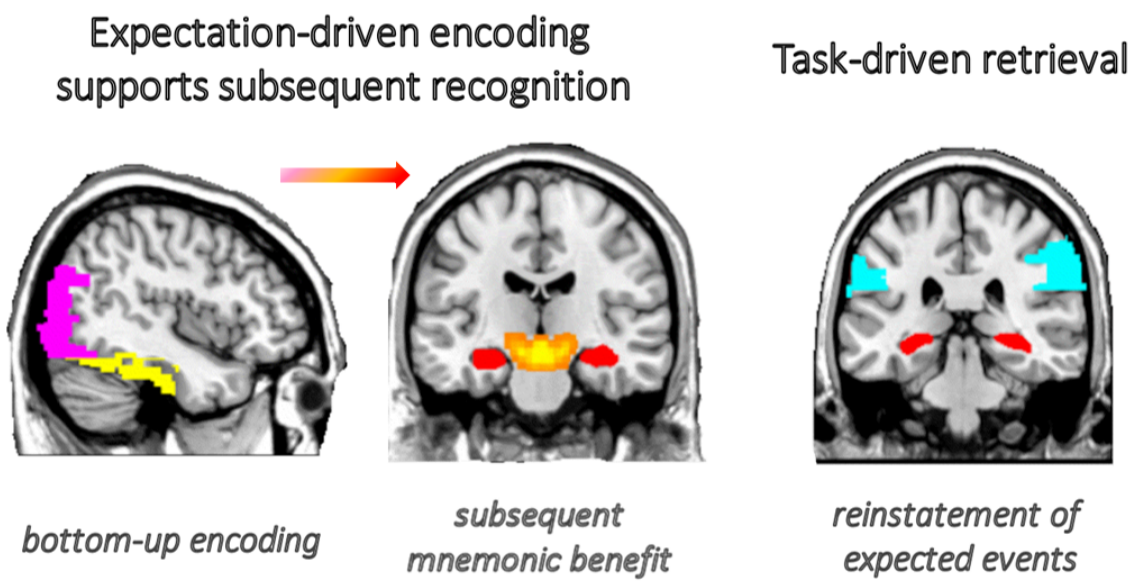


Figure 5 | Illustration of the brain networks involved in processing unexpected and expected events. Left and middle: Expectation-driven, goal-irrelevant encoding. Expectation violation engages bottom-up processing along the ventral visual stream (inferior occipital in pink, fusiform in yellow), regardless of memory performance. The subsequent mnemonic consequences of this shift towards encoding involve the hippocampus (red) and midbrain (orange) dopaminergic regions, underlying subsequent beneficial memory performance. **Right**: **Task-driven retrieval.** In the absence of expectation violation (expected events), engagement of parietal regions (supramarginal gyrus in cyan, angular gyrus not shown) of the core recollection network support reinstatement and memory performance, in accordance with goal to retrieve.

Discussion

The experience of surprise, or expectation violation, has a beneficial effect on learning, but whether it triggers an encoding response even when the dominant goal is to retrieve, remained unclear. In two experiments, we used a contextual expectation manipulation to better understand the dynamic nature of the adaptive memory mechanism triggered by expectation violation during retrieval and its potential mnemonic consequences on hippocampal-dependent memory. We found that encountering unexpected events at retrieval, elicited increased involvement of regions along the ventral visual stream and encoding-related hippocampal activity. However, a sequence of two unexpected, but similar events ($U_{prev}U_{curr}$), resulted in the poorest memory performance, suggesting that

expectation violation can hamper memory performance by preferentially, and inappropriately, engaging a bottom-up encoding state. Interestingly, we also found a later beneficial effect of contextual surprise, such that the presentation of an unexpected event did not support its own recognition, but it did boost the correct recognition of the following, expected, and similar set events (targets and F1). This behavioural effect was associated with increased activity in the hippocampus, midbrain dopaminergic regions (SN/VTA), and occipital cortex. Expected events, on the other hand, were associated with reinstatement-like activity in parietal regions. Taken together, these results provide the first direct evidence for the dynamic engagement of a surprise-driven, bottom-up encoding mechanism (Figure 5), that acts at the expense of concurrent retrieval, and its subsequent mnemonic consequences.

Expectation violation is often associated with improved memory performance which is attributed to adaptive memory formation (Kumaran and Maguire, 2007; Lisman and Grace, 2005; Shohamy and Wagner, 2008). Our results support this view, but critically, also extend it to account for retrieval effects. In line with the idea that increased weight is given to bottom-up inputs upon encountering a prediction error (Kafkas and Montaldi, 2018a; Stoppel et al., 2009), we found increased involvement of visual processing regions in occipital cortex and fusiform gyrus for unexpected events. These regions have been found to increase their activity with stronger levels of unexpected novelty (Kafkas and Montaldi, 2014), resulting in the increased perceptual processing of unexpected events. Although memory formation relies on such bottom-up processing, in order to provide evidence for an encoding mechanism, it is necessary to also demonstrate mnemonic consequences of the increased weight placed on sensory inputs. Indeed, we observed an interaction between current and previous expectation status of similar events (targets and F1), such that memory performance was modulated by a previous unexpected event (improved recognition performance for $U_{prev}E_{curr}$, reduced for $U_{prev}U_{curr}$). Therefore, engaging with unexpected information at retrieval turns on a bottom up-encoding mechanism that while having the potential to enhance current encoding, this spontaneous mechanism can also impair current recognition. Other patterns of expectation modulated memory might be expected under different retrieval conditions (e.g., Kafkas and Montaldi, 2018b).

Whilst occipital involvement was observed for both $U_{prev}E_{curr}$ and $U_{prev}U_{curr}$ events (i.e. independent of mnemonic consequence), only $U_{prev}E_{curr}$ events were associated also with hippocampal and SN/VTA activation. This finding, together with the memory boost for $U_{prev}E_{curr}$ events, pertains to the role these regions play in mediating the processing of perceptual inputs for memory formation. In terms of shifting the weight towards bottom-up inputs, the hippocampus is perfectly placed, at the apex of the sensory hierarchy and connected with midbrain dopaminergic regions, to be in the driving seat. This could also explain why we did not observe hippocampal or midbrain activation for any other unexpected contrasts; their co-activation is likely associated with a boost in memory performance (Kafkas and Montaldi, 2015). Taken together, these findings suggest that a surprise-driven increased weight on bottom-up inputs is goal-independent, but its mnemonic consequences appear to depend on the task at hand. During learning or exploration, further encoding supports later memory for the unexpected event (Garrido et al., 2015; Greve et al., 2017b; Gruber et al., 2013; Li et al., 2003; Long et al., 2016). On the other hand, when retrieval is necessary (as in the current paradigm), the shift towards encoding, despite increased perceptual processing, results in worse memory performance for the to-be-retrieved information (Duncan et al., 2012b; Kim et al., 2014). Therefore, the short- and the long-term consequences of processing unexpected events may be in stark contrast with each other. In the short-term, while another task is performed, unexpected events can be disruptive to our memory, while in the long-term they can be beneficial, especially when our memory for a similar event is tested (as in the present study), and when our memory is tested after a delay (Frank et al., 2020; Kafkas and Montaldi, 2015).

Most previous research exploring shifts between encoding and retrieval states has been done in rodents (Hasselmo and Stern, 2014), or in humans where explicit instructions were utilised (Richter et al., 2016). A recent study (Bein et al., 2020) showed that functional connectivity between hippocampal subfield CA1 and entorhinal cortex or CA3 (reflecting a bias towards an encoding or a retrieval state, respectively,) was modulated by mnemonic prediction errors. The authors postulated that the bias they observed towards one hippocampal state or another could be modulated by dopaminergic inputs (Kafkas and Montaldi, 2018a; Lisman and Grace, 2005). In line with these results, using a manipulation that was independent of the memory task (and completely implicit to most

participants), we found expectation-driven hippocampal and midbrain activation, which we suggest reflects an encoding response triggered by expectation violation. The mnemonic consequences of this shift are manifested in more accurate recognition of later highly similar expected events. Further support for the engagement of a hippocampal encoding state can be found in the increased activity along the ventral visual stream for unexpected compared to expected events, reflecting increased bottom-up processing. We propose the increased gain on bottom-up inputs, together with the hippocampal-midbrain activations, are markers of an encoding state triggered by expectation violation, even in the absence of explicit reward or instructions (only two participants in Experiment 1 and five participants in Experiment 2 realised the expectation manipulation). Whilst unexpected events in which you experience a high level of surprise are particularly memorable, our results suggest that, as in other areas of research, it remains to be determined to what extent explicit awareness of surprise modulates this mechanism.

Although a shift towards encoding, and away from retrieval, upon encountering an unexpected event can explain why we do not observe a memory boost for these events, it is worth considering why increased encoding of the U_{prev} event results in better memory performance for subsequent expected events. One possibility is that the initial expectation-violation related shift towards encoding results in a sharper representation of U_{prev}. This representation is then optimised for pattern completion and retrieval, when the subsequent expected similar event is presented. Therefore, the current expected event would be more readily compared to the previous unexpected event, as well as to the originally encoded object. The contrasting delta similarity between these representations can thus, inform and aid the current recognition decision for the expected event. An alternative interpretation for this finding is a prospective memory enhancement (Dunsmoor et al., 2015; Oyarzún et al., 2016). Aversive (Dunsmoor et al., 2015) and rewarding (Oyarzún et al., 2016) stimuli have been shown to up-regulate encoding and consolidation of temporally- or semantically-close stimuli, supporting prospective memory enhancement for these inconsequential items. Whilst prospective enhancement could explain the boost of E_{curr} memory performance, this account does not explain the poor memory performance for $U_{prev}U_{curr}$ events, nor does it predict the

selectivity of these effects to highly similar inputs in a relatively short period of time (i.e. without longer-term consolidation).

Critically, the interactions between current and previous events' expectation were observed only for targets and F1 (i.e. the foil with the highest degree of similarity to the encoded object). Moreover, these effects were unchanged by interfering events from the same set (F2, and in Experiment 1 also F3) or events from different sets presented during the task (i.e. there could be up to 200 trials between targets and F1s from the same set). The selectivity of the expectation interactions to the higher end of the perceptual similarity scale, and their robustness with respect to interference from other stimuli, suggests that a high perceptual and memorial load is required to trigger this encoding mechanism, extending previous findings (Bein et al., 2020). In such situations, the ability to process and compare the fine details of current inputs and recently stored representations underpins correct recognition decisions (Yassa and Stark, 2011). Therefore, the triggering of enhanced perceptual processing by expectation violation serves an adaptive purpose (Hawco and Lepage, 2014; Stoppel et al., 2009). For less similar events, which are more easily recognised as new (F2, F3), the effect of a sharper representation elicited by expectation violation is less critical (Frank et al., 2020).

Further support for the role played by perceptual load in engaging an encoding state, can be seen in the occipital and fusiform effects for those unexpected events presented second within the set sequence, compared to their expected counterparts. Recognition decisions for these events have to overcome interference from the first set event, which may lead to increased engagement of perceptual processing in order to better compare the current sensory input to the stored representations. Whilst the ventral visual stream encoding effects observed for unexpected events can be explained by the bottom-up account (Kafkas and Montaldi, 2018a; Kim, 2011), we also found that hippocampal encoding activity correlated with ACC activation for unexpected events at retrieval. Recent findings suggest that the ACC interacts with the dopaminergic system to instigate the encoding of rewarding information (Aberg et al., 2020; Elston and Bilkey, 2017). Our results are consistent with these findings, suggesting the ACC is not only involved in error monitoring (Carter and van Veen, 2007), but also in the encoding of an unexpected outcome even when reward is not explicitly present in the task.

As the expectation manipulation took place during retrieval, it remains unclear whether expected events resulted in task-relevant retrieval (Figure 5), or in the active engagement of a retrieval state, irrespective of task demands. Although the observed engagement of parietal regions of the core recollection network (Hayama et al., 2012) is indicative of reinstatement, it does not differentiate between the two alternatives. Future studies could orthogonalise expectation and memorial state, therefore allowing a factorial design of goal (encoding/retrieval) and expectation status. Such examinations of shifts towards a retrieval state, perhaps coupled with designs optimised for functional and effective connectivity, will contribute to on-going efforts to explain how the hippocampus shifts between memory states (Bein et al., 2020; Colgin, 2016; Kafkas and Montaldi, 2018a). Finally, further characterising the temporal dynamics of the relationship between expectation violation and bottom-up processing would shed light on the role the ventral visual processing stream plays in modulating memory circuits, and how memory demands might direct activity in the visual system.

In conclusion, here we report novel evidence for a ubiquitous encoding mechanism, triggered at retrieval by expectation violation, resulting in differential effects on recognition performance. A sequence of unexpected followed by expected similar events, was associated with a boost in recognition performance and activation in the hippocampus, SN/VTA and occipital cortex. On the other hand, a sequence of two unexpected events was associated with poor recognition, and activation in occipital cortex. The complex temporal dynamics of the effects of this mechanism on memory, clearly demonstrate that the expectation-driven shift towards an encoding state, supported by enhanced bottom-up processing, is detrimental to memory performance in real time, but beneficial for the correct recognition of subsequent similar events. These findings have important implications for our understanding of how our processing of sequential events, similar or different, expected or unexpected, is modulated by the temporal dynamics of the event sequence.

Methods

Experiment 1

Participants. 30 participants (4 males) between the ages 18-22 (M = 19, SD = 1.04) gave informed consent and took part in the experiment. The sample size was selected based on previous studies utilising a similar design and analysis approach (Frank et al., 2020; Kafkas and Montaldi, 2018b). Participants had normal or corrected-to-normal vision and no history of neurological or psychiatric disorders. All procedures were approved by the University of Manchester Research Ethics Committee. Four participants were excluded from any further analysis due to failure to learn the cue-outcome association (3) or recognition performance below chance (1). Data from 26 participants are reported.

Materials. 78 images of natural (39) and man-made (39) objects were selected from the Similar Objects-Lures Image Database (SOLID (Frank et al., 2019)). These images were used as the target objects, presented during encoding. Using the dissimilarity index from SOLID, three foils of decreasing levels of similarity (F1 – most similar, F2 – mid-level F3 - lowest similarity) were selected for each target image. Additionally, similarity was parametrically manipulated by keeping the average distance between the levels constant (average dissimilarity 2100 DI; see Figure 1 for examples and Frank et al., 2019 for further explanation on DI values). Therefore, 78 object sets (one target and three foils; total 312 stimuli) were utilised.

Procedure. The experiment was controlled using PsychoPy version 1.82 (Peirce, 2007) and consisted of four parts (similar to design used in (Kafkas and Montaldi, 2018b)). First, at encoding, participants were presented with the target objects twice to ensure sufficient exposure. During the first presentation, the object was shown on the screen for 3 seconds, and participants were asked to decide whether the object was man-made or natural; pressing the left arrow key for natural, and the right arrow key for man-made. In the second presentation participants were asked to pay close attention to the images (shown for 5s each) and were informed that they would be asked to distinguish between the presented ('old') object and similar ('new') objects later; they did not have to make any response during the second encoding presentation. The order of image presentation was randomised across participants. The next phase involved a 5-minute filled delay task, during which participants solved arithmetic problems.

In the third phase, a rule-learning task was used to allow manipulation of expectation later at retrieval. Participants were asked to learn the contingency between a symbol (acting as a cue) and an object's category, man-made or natural. Four symbols were used

in total, two for each category, and these were counterbalanced between participants. Each cue was presented 14 times. Trials began with a 500ms fixation point, followed by the cue, presented for 1 second. During this time participants were instructed to guess the following object's category by using the same keys as in the encoding phase. Feedback and a new object (not tested) were then presented together for 2 seconds. This task established participant's expectation regarding the cue-object sequences, and critically this was manipulated at retrieval. To ensure expectations were set, only data from participants who reached criterion (above 50% accuracy in the first half, and 75% accuracy in the second half of the task) were analysed (See Supplementary Figure S.1 for accuracy and RT in the rule-learning task).

The final phase was an old/new recognition task, during which all set events (targets, F1, F2 and F3) were presented. The experimenter informed participants that they will be presented with old (target) objects, and similar new ones (foils). Each retrieval trial began with a fixation point (500ms), followed by a cue (1 second) and an object (up to 5 seconds). Participants were told to focus on the object and press 'old' if they thought it was exactly the same as the target they had previously seen. Participants were instructed to press 'new' if they noticed anything different in the object. Following 12 practice trials, the main task began. The key manipulation at retrieval was the validity of the cue. One-third of the cues were misleading, making the object to-be-recognised unexpected. Valid trials (e.g. cue natural followed by a natural object) were marked as expected. As an old/new recognition task was used, the four old/new decisions per object set were independent of each other, critically meaning that an 'old' response could be given for multiple set events (i.e. for target and one or more foils).

Data analysis. To assess expectation-modulated dynamic encoding, we collated object sets (apples, scissors etc.). In each set, there were four set events (target, F1 - most similar, F2 – mid similarity, and F3 – least similar), each with a randomly assigned (a) expectation condition, (b) presentation position within the set, and (c) order of presentation at retrieval (1-312 trials). We ran a mixed-effects binary logistic regression model on these ungrouped data, using expectation status of the current and previous set events, and the presentation position at retrieval as a covariate. Models were computed using the lme4 package (Bates et al., 2015) in the R environment (R Development Core Team, 2008). The parameters of such models can be used to assess the probability of

giving a correct response ('old' for targets, 'new' for foils) and also account for each participant's unique intercept (response bias). To assess the slope of each predictor in the model (H_0: $\beta = 0$), we used an omnibus χ^2 Wald test (West et al., 2014), as implemented in the car package (Fox and Weisberg, 2018). Extraction and plotting of the effects reported below was conducted using the effects (Fox, 2003), emmeans (Searle et al., 1980) and ggplot2 (Wickham, 2009) packages. To examine the dynamic interaction between perceptual similarity and expectation status, for each set event we devised three models of interest, modelling each event separately as a function of the preceding event sequence. For example, targets preceded by F1 were modelled separately from targets preceded by F2. Each model thus included the current and previous set events' expectation status, as well as the order of presentation at retrieval as a covariate. Code and data are available here: https://github.com/frdarya/DynamicExpectation.

Experiment 2

Experiment 2 was Identical to Experiment 1 except the following details.

Participants. 25 participants (8 males, ages 18-33, M = 25, SD = 4.2) gave informed consent and took part in the study. Participants had normal or corrected-to-normal vision and no history of neurological or psychiatric disorders. All procedures were approved by the University of Manchester Research Ethics Committee. One participant was excluded from all analyses due to failure to learn the cue-outcome contingency during the rule-learning task.

Procedure. A similar paradigm and expectation manipulation was used in Experiment 2, with the following exceptions; in the encoding phase, each object was presented three times consecutively. Each object was on the screen for 2000ms, with a jittered fixation cross (250-750ms) between each presentation. During the first and second presentations participants were asked to make a semantic decision about the object, whether it is man-made or natural, and whether it is more likely to be found indoors or outdoors. The order of these questions was random. During the third presentation participants were always asked to study the object carefully focusing on the details. Following the third presentation there was another jittered fixation cross, for a longer period of time (800-1200ms), to create mini-blocks separating each object.

The rule-learning task was identical to that used in Experiment 1, except for a longer response presentation time of the cue (3s instead of 1s) and a jittered fixation cross (250-750ms) between each trial (see Supplementary Figure S.1 for behavioural and Figure S.2 for fMRI results from the rule learning task). Before the retrieval task, participants solved arithmetic problems for two minutes (not scanned). At retrieval, we used two levels of foil similarity (F1 and F2) in each set, instead of three. F3 objects were not used, as they did not yield any effects of interest in Experiment 1. Each retrieval trial started with a jittered fixation cross (250-750ms), followed by a presentation of the cue for 1000ms and then the set event (target, F1 or F2) for 3000ms. In all scanned tasks, we used implicit baselines (fixation crosses for 3500ms in encoding and rule-learning tasks, 4500ms in retrieval) in 30% of trials.

Behavioural data analysis. Following the analysis and results from Experiment 1, we collapsed targets and F1 events and modelled the probability of making a correct decision (hits and correct rejections) based on the current set event's expectation status and the previous set event's expectation status. As in Experiment 1, to eliminate the memory strength confound, only correct responses were included (See Supplementary Figure S.3 for similar results when including all trials, and Supplementary Figure S.4 for separate analyses of targets and F1). Correct rejections of F2 foils were also modelled as a function of previous set events, as was done for Experiment 1.

fMRI acquisition and analysis. MR scanning was carried out on a 3T MRI scanner (Philips, Achieva). To minimise movement during the scan, foam wedges and soft pads were used to stabilise the participant's head. First, T1-weighted images (matrix size: 256x256, 160 slices, voxel size 1mm isotropic) were collected while participants rested in the scanner. A gradient echo-planar imaging (EPI) sequence was used to collect T2* images for the BOLD signal. 40 slices parallel to the AC-PC line, covering the whole brain (matrix size 80 x 80, voxel size 3 x 3 x 3.5mm^3), were obtained for each volume (TR = 2.5s, TE = 35ms). Participants performed three tasks in the scanner (encoding 313 volumes; rule-learning 143 volumes; retrieval 534 volumes) and a distractor task, which was not scanned.

fMRI data were pre-processed and analysed using SPM12 (Statistical Parameter Mapping, Wellcome Centre for Human Neuroimaging, University College London https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). Images were realigned to the mean image using a six-parameter rigid body transformation, resliced using sinc

interpolation and slice-time corrected to the middle slice. T1 anatomical images were co-registered to the corresponding mean EPI image. Spatial normalisation to the Montreal Neurological Institute (MNI) template was carried out using the DARTEL toolbox implemented in SPM12 (Ashburner, 2007). An isotropic 8mm FWHM Gaussian kernel was used for smoothing the normalised EPI data for univariate analyses. To remove low-frequency noise the data was high-pass filtered using a cut-off of 128s. The same pre-processing pipeline, without smoothing, was used for the encoding-retrieval similarity analysis (see *reinstatement analysis below*). Two a-priori regions of interest (ROIs), the bilateral hippocampus, and a midbrain ROI including only the substantia nigra (SN) and ventral tegmentum area (VTA) were used. The hippocampus mask was taken from the Harvard-Oxford anatomical atlas (threshold at 25% probability; Desikan et al., 2006). The midbrain mask was taken from the probabilistic atlas of the midbrain (Murty et al., 2014).

Univariate analyses. Each participant's functional data from the retrieval session was analysed using the general linear model (GLM) framework within an event-related design modelling the canonical hemodynamic response function. The six motion parameters produced at realigmnet for each session were used as nuisance regressors. To minimise residual motion artefacts the ArtRepair toolbox (http://cibsr.stanford.edu/tools/human-brain-project/artrepair-software.html) was used to produce additional nuisance regressors for each participant. The time series were high-pass filtered to remove low frequency noise (128s cut-off). All activations reported are significant at the uncorrected level $p < 0.001$ with at least 9 contiguous voxels. Given our a priori hypothesis for the ROIs introduced above (bilateral hippocampus and SN/VTA), a small volume correction (SVC) approach was adopted. When SVC was applied, activations were also corrected for family-wise error (FWE) for the ROI volume. Unthresholded group-level T-maps are available here: https://neurovault.org/collections/TTDMPHLE/.

To test the behavioural effect of an interaction between a previous event's contextual expectation and the current event's contextual expectation, we collapsed targets and F1 and classified them based on presentation order (which came first) and their expectation status. F2 events were modelled as a separate condition. In this analysis, we compare the current item between current and previous expectation status (e.g. $E_{prev}\mathbf{U_{curr}} > E_{prev}\mathbf{E_{curr}}$; see Supplementary materials for analysis of the previous item). Given our experimental

design, there are four parameters whose interactions could be further explored: set event (target, F1, F2), contextual expectation (expected or unexpected), memory response given (correct or incorrect) and presentation order within the set (first, second or last). This results in 36 conditions, however, these could not be modelled together due to insufficient number of trials per bin (n < 7) for the majority of participants. Therefore, we devised two separate models; in the first model, we examined expectation x set event x successful memory interactions by collapsing trials across presentation order, and modelling only correct responses for each event (all incorrect responses were modelled as a separate regressor). In the second model, we explored the interaction of expectation, set event and presentation order, irrespective of recognition responses.

Bottom-up encoding analysis. In this analysis we examined regions in which retrieval activity correlated with block-by-block estimates of encoding activity. To do so, we extracted hippocampal univariate encoding activity (beta estimates) during each encoding mini-block (three repetitions of the same item). These estimates were used as parametric modulators applied to retrieval data in a GLM analysis splitting retrieval trials by expectation status. Next, we converted the beta estimates of encoding activity within this mask into t-maps (Misaki et al., 2010) and used them as parametric modulators applied to retrieval data (using the beta estimates did not change the results). This approach allowed us to query retrieval data looking for regions whose retrieval activity correlated with hippocampal encoding activity for expected vs. unexpected events.

Reinstatement analysis. To identify which regions showed top-down retrieval-oriented activity, we conducted an encoding-retrieval similarity (ERS) analysis using CosmoMVPA (Oosterhof et al., 2016) in Matlab R2018a (MathWorks Inc.). For this purpose, we modelled each encoding mini-block as a condition of interest. For each of the 78 blocks, we converted the beta estimate map into a t-statistics (Dimsdale-Zucker and Ranganath, 2019). As one participant's last trial at encoding could not be estimated, only 77 sets were analysed for this participant. Pairwise Pearson's correlations between all trials were calculated and were then Fisher z-transformed (values ranging between -2 and 2). This analysis was carried out in bilateral hippocampus, as an a-priori ROI. Hippocampal ERS scores were then used as parametric regressors that were applied to retrieval data in a model splitting trials based on their expectation status and set event. This was done to examine which regions exhibited retrieval activation that correlated with hippocampal

ERS, capturing reinstatement of the encoded object (a *t*-contrast on the parametric event).

Supplementary data

Rule learning performance

Experiment 1 (Figure S.1A). We split rule-learning trials into first and second halves, and tested accuracy and reaction time (RT) between halves as a measure of learning the cue-category contingency. On the second half of the task, performance was significantly better than the first half ($t(25) = 8.05$, $p < 0.001$) and RT was faster ($t(25) = -2.88$, $p = 0.008$). *Experiment 2* (Figure S.1B). Similar results were obtained in Experiment 2, with better performance during the second half of the task ($t(23) = 5.24$, $p < 0.001$) and faster RT ($t(23) = -3.85$, $p < 0.001$).

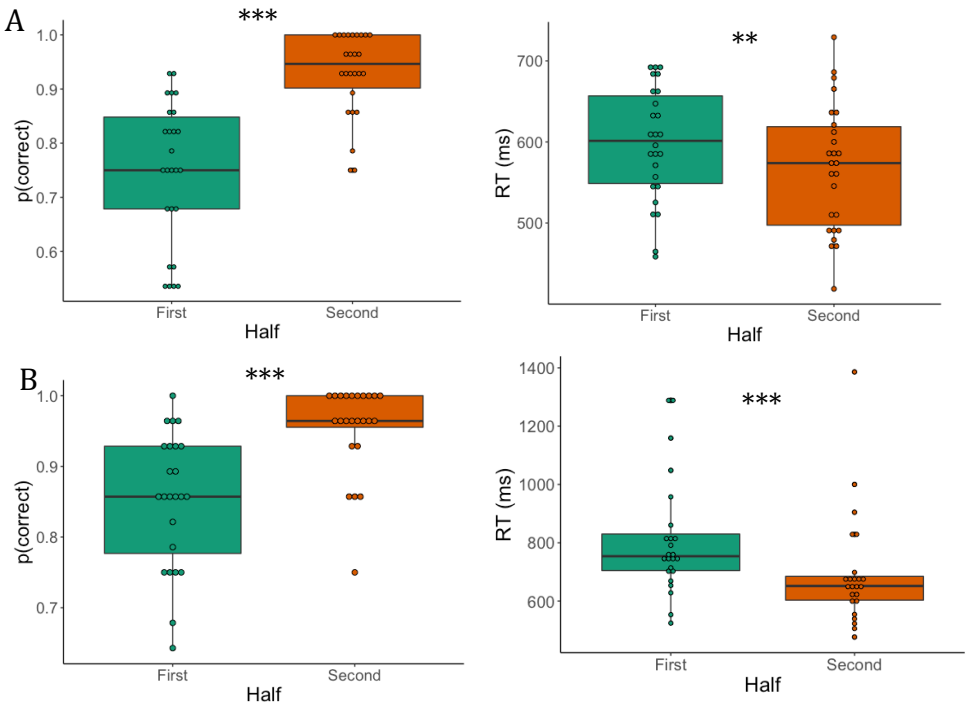


Figure S.1 | Rule learning performance. A) Experiment 1 – accuracy (left) and RT (right) on rule learning trials, split into first and second half of the task. B) Experiment 2 – accuracy (left) and RT (right).

Rule learning fMRI activation – Experiment 2

To identify regions involved in learning the contingency (and therefore setting the contextual expectation) between the cue and an object's category , we examined BOLD activity during the rule learning task using a correct > incorrect response contrast. As one participant did not have any incorrect responses, this analysis was done for the remaining 23 participants. We found increased activity in bilateral caudate ($x = 18$, $y = 12$, $z = 21$, $k = 19$; $x = -18$, $y = 18$, $z = 18$, $k = 73$) and right putamen ($x = 27$, $y = 3$, $z = 12$, $k = 63$).

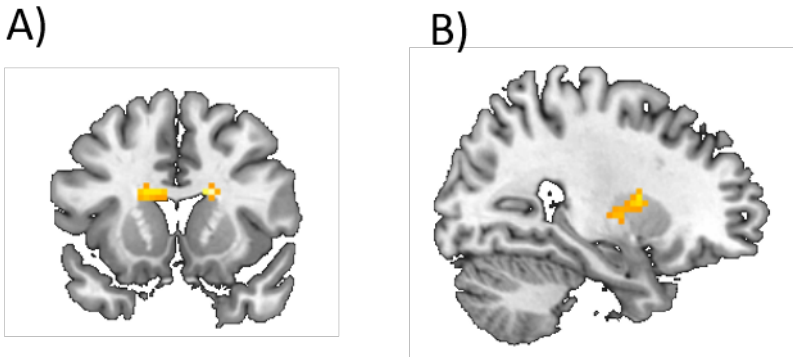
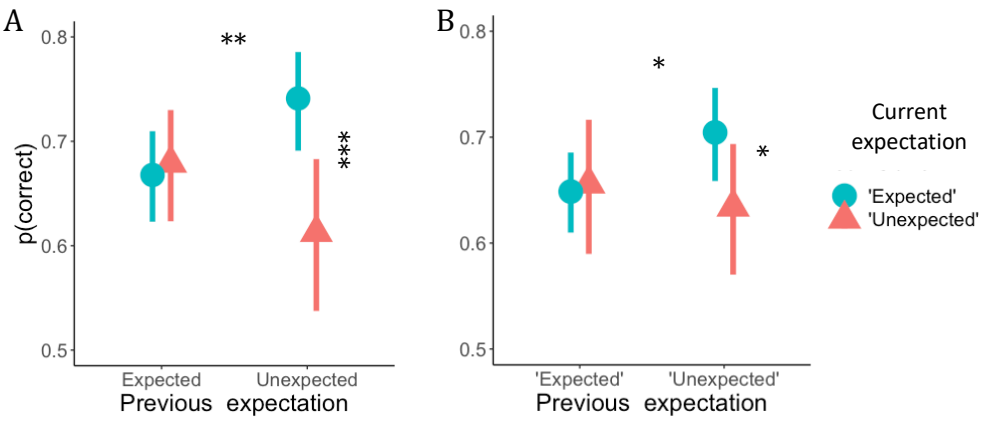


Figure S5.2 | Rule learning task. Activations for correct > incorrect responses. A) Bilateral caudate B) Right putamen.

Current by previous expectation status interaction (including all previous trials)

Experiment 1 (Figure S3.A). A significant main effect of previous expectation was observed ($\beta = 0.35$, $X^2(1) = 7.75$, $p = 0.005$), as well as a significant interaction between previous and current expectation status ($\beta = -0.64$, $X^2(1) = 8.81$, $p = 0.003$). Follow-up contrasts revealed better memory performance for $U_{prev}E_{curr}$ compared to $E_{prev}E_{curr}$ ($z = -2.78$, $p = 0.005$), and $U_{prev}U_{curr}$ ($z = 3.36$, $p < 0.001$).

Experiment 2 (Figure S3.B). A significant main effect of previous expectation was observed ($\beta = 0.25$, $X^2(1) = 4.41$, $p = 0.035$), and a trend towards an interaction between previous and current expectation status ($\beta = -0.35$, $X^2(1) = 2.47$, $p = 0.11$). A comparison



of the direct contrasts revealed better memory performance for $U_{prev}E_{curr}$ compared to $E_{prev}E_{curr}$ ($z = -2.1$, $p = 0.035$), and $U_{prev}U_{curr}$ ($z = 1.97$, $p = 0.048$).

Figure S.3 | Expectation interaction effects when including all previous trials. A) Experiment 1. B) Experiment 2.

Predicting hits and CR1 separately – Experiment 2

Predicting hits (Figure S.4A). A significant interaction between previous (F1) and current (target) expectation status (β = -1.26, $X^2(1)$ = 6.09, p = 0.013). Follow-up examination of contrasts revealed more hits for expected targets following unexpected F1, compared to those unexpected targets (z = 2.85, p = 0.004), and more hits for unexpected targets following expected F1, compared to those following unexpected F1 (z = 2.19, p = 0.029). For targets following F2 events, a significant interaction between the target's and F2's expectation status was observed (β = -0.99, $X^2(1)$ = 4.75, p = 0.029), with more hits for expected targets, compared to unexpected ones, following unexpected F2 (z = 2.17, p = 0.029).

Predicting CR1 (Figure S.4B). A trend towards a main effect of previous target's expectation status was observed (β = 0.33, $X^2(1)$ = 2.93, p = 0.087), with more CR1 following unexpected targets compared to expected ones. Despite the fact that the interaction between previous (target) and current (F1) expectation was not significant (β = -0.31, $X^2(1)$ = 0.78, p = 0.375), the data followed the same pattern observed above and in the main text, with most CR1 for expected F1 following unexpected targets (see Figure S.4B). For F1 following F2, a main effect of F1 condition was observed (β = -0.48, $X^2(1)$ = 4.84, p = 0.027), with more CR1 for expected compared to unexpected F1.

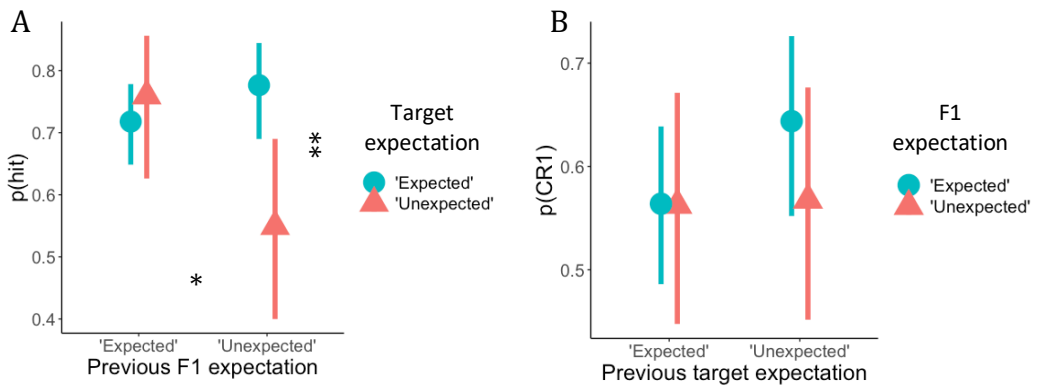


Figure S.4 | Experiment 2 - Predicting hits and CR1. A) Predicting hits. B) Predicting CR1

<u>First set events</u>

Experiment 1 (Figure S.5A). When examining first targets within the set, we found more hits for expected targets compared to unexpected ones ($X^2(1)$ = 5.6, p = 0.018). When examining correct rejections of foils (F1, F2, F3), we found a significant main effect of foil level ($X^2(2)$ = 57.8, p < 0.001), with better memory for less similar foils. The main effect of expectation $X^2(1)$ = 0.4, p = 0.524) and the interaction between foil and expectation ($X^2(2)$ = 2.54, p = 0.281) were not significant.

Experiment 2 (Figure S.5B). When examining first targets within the set, we found more hits for expected targets compared to unexpected ones ($X^2(1) = 4.7$, $p = 0.03$). When examining correct rejections of foils (F1, F2), we found a significant main effect of foil level ($X^2(1) = 19.2$, $p < 0.001$), with better memory for less similar foils. The main effect of expectation $X^2(1) = 0.18$, $p = 0.668$) and the interaction between foil and expectation ($X^2(2) = 0.004$, $p = 0.949$) were not significant.

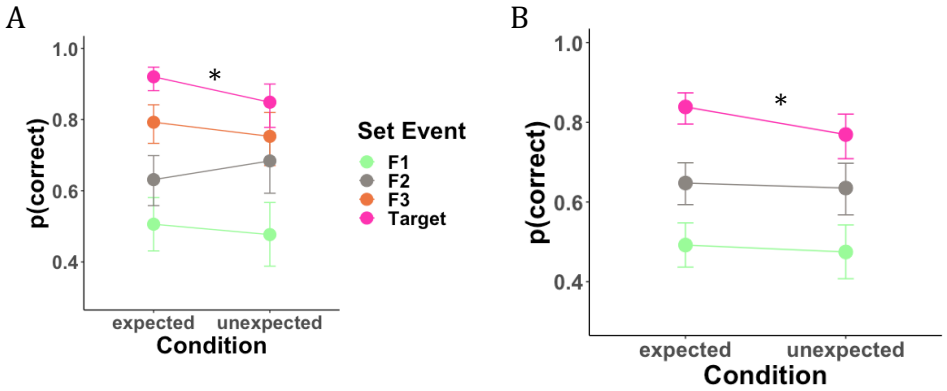


Figure S.5 | First set events. A) Experiment 1. B) Experiment 2.

<u>Expectation by presentation order interaction, by set event – Experiment 2</u>

We found increased engagement of middle occipital cortex (BA 18, x = 30, y = -87, z = 12, k = 124) and right posterior parietal cortex (BA 5, x = 6, y = -54, z = 66, k = 77) for second unexpected F1 > second expected F1 events, and increased activity in the right fusiform gyrus (BA 37, x = 30 y = -42, z = -15, k = 63) for second unexpected F2 > second expected F2 events.

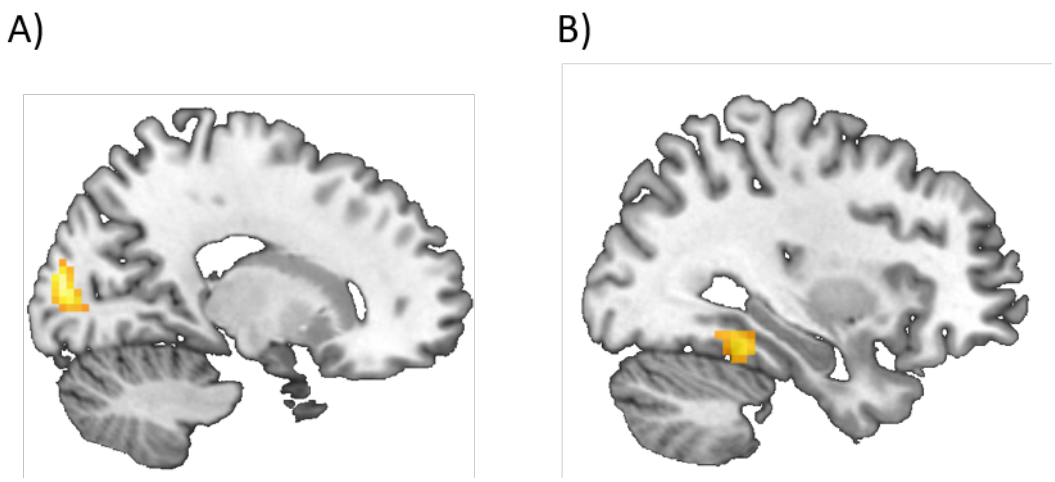


Figure S.6 | Unex>Exp in second set events. A) Second unexpected F1 > second expected F1, increased activity in middle occipital cortex (BA 18) and right posterior parietal cortex (BA 5) (not shown). B) Second unexpected F2 > second expected F2, increased activity in right fusiform (BA 37).

References

Aberg KC, Kramer EE, Schwartz S. 2020. Interplay between midbrain and dorsal anterior cingulate regions arbitrates lingering reward effects on memory encoding. *Nat Commun* **11**:1829. doi:10.1038/s41467-020-15542-z

Aly M, Turk-Browne NB. 2017. How Hippocampal Memory Shapes, and Is Shaped by, AttentionThe Hippocampus from Cells to Systems. Cham: Springer International Publishing. pp. 369–403. doi:10.1007/978-3-319-50406-3_12

Ashburner J. 2007. A fast diffeomorphic image registration algorithm. *Neuroimage* **38**:95–113. doi:10.1016/j.neuroimage.2007.07.007

Axmacher N, Cohen MX, Fell J, Haupt S, Dümpelmann M, Elger CE, Schlaepfer TE, Lenartz D, Sturm V, Ranganath C. 2010. Intracranial EEG Correlates of Expectancy and Memory Formation in the Human Hippocampus and Nucleus Accumbens. *Neuron* **65**:541–549. doi:10.1016/j.neuron.2010.02.006

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* **67**. doi:10.18637/jss.v067.i01

Bein O, Duncan K, Davachi L. 2020. Mnemonic prediction errors bias hippocampal states. *Nat Commun* **11**:3451. doi:10.1038/s41467-020-17287-1

Buzsáki G. 2002. Theta Oscillations in the Hippocampus. *Neuron* **33**:325–340. doi:10.1016/S0896-6273(02)00586-X

Carter CS, van Veen V. 2007. Anterior cingulate cortex and conflict detection: an update of theory and data. *Cogn Affect Behav Neurosci* **7**:367–379. doi:10.3758/cabn.7.4.367

Colgin LL. 2016. Rhythms of the hippocampal network. *Nat Rev Neurosci* **17**:239–249. doi:10.1038/nrn.2016.21

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ. 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**:968–980. doi:10.1016/j.neuroimage.2006.01.021

Dimsdale-Zucker HR, Ranganath C. 2019. Representational Similarity Analyses: A Practical Guide for Functional MRI Applications. *Handb Behav Neurosci* **28**:509–525. doi:10.1016/B978-0-12-812028-6.00027-6

Duncan K, Ketz N, Inati S, Davachi L. 2012a. Evidence for area CA1as a match/mismatch

detector: A high-resolution fMRI study of the human hippocampus. *Hippocampus* **22**:389–398. doi:10.1002/hipo.20933.Evidence

Duncan K, Sadanand A, Davachi L. 2012b. Memory's Penumbra: Episodic Memory Decisions Induce Lingering Mnemonic Biases. *Science (80-)* **337**:485–487. doi:10.1126/science.1221936

Dunsmoor JE, Murty VP, Davachi L, Phelps EA. 2015. Emotional learning selectively and retroactively strengthens memories for related events. *Nature* **520**:345–348. doi:10.1038/nature14106

Elston TW, Bilkey DK. 2017. Anterior Cingulate Cortex Modulation of the Ventral Tegmental Area in an Effort Task. *Cell Rep* **19**:2220–2230. doi:10.1016/j.celrep.2017.05.062

Fox J. 2003. Effect Displays in R for Generalised Linear Models. *J Stat Softw* **8**. doi:10.18637/jss.v008.i15

Fox J, Weisberg S. 2018. An R Companion to Applied Regression. Sage.

Frank D, Gray O, Montaldi D. 2019. SOLID-Similar object and lure image database. *Behav Res Methods*. doi:10.3758/s13428-019-01211-7

Frank D, Kafkas A. 2021. Expectation-driven novelty effects in episodic memory. *Neurobiol Learn Mem* **183**:107466. doi:10.1016/j.nlm.2021.107466

Frank D, Montemurro MA, Montaldi D. 2020. Pattern Separation Underpins Expectation-Modulated Memory. *J Neurosci* **40**:3455–3464. doi:10.1523/JNEUROSCI.2047-19.2020

Friston KJ. 2005. A theory of cortical responses. *Philos Trans R Soc B Biol Sci* **360**:815–836. doi:10.1098/rstb.2005.1622

Garrido MI, Barnes GR, Kumaran D, Maguire EA, Dolan RJ. 2015. NeuroImage Ventromedial prefrontal cortex drives hippocampal theta oscillations induced by mismatch computations. *Neuroimage* **120**:362–370. doi:10.1016/j.neuroimage.2015.07.016

Greve A, Cooper E, Kaula A, Anderson MC, Henson RN. 2017a. Does prediction error drive one-shot declarative learning? *J Mem Lang* **94**:149–165. doi:10.1016/j.jml.2016.11.001

Greve A, Cooper E, Kaula A, Anderson MC, Henson RN. 2017b. Does prediction error drive one-shot declarative learning ? *J Mem Lang* **94**:149–165. doi:10.1016/j.jml.2016.11.001

Gruber MJ, Watrous AJ, Ekstrom AD, Ranganath C, Otten LJ. 2013. Expected reward modulates encoding-related theta activity before an event. *Neuroimage* **64**:68–74. doi:10.1016/j.neuroimage.2012.07.064

Hasselmo ME, Bodelon C, Wyble BP. 2002. A Proposed Function for Hippocampal Theta Rhythm : Separate Phases of Encoding and Retrieval Enhance Reversal of Prior Learning. *Neural Comput* **14**:793–817.

Hasselmo ME, Stern CE. 2014. Theta rhythm and the encoding and retrieval of space and time. *Neuroimage* **85**:656–666. doi:10.1016/j.neuroimage.2013.06.022

Hawco C, Lepage M. 2014. Overlapping patterns of neural activity for different forms of novelty in fMRI. *Front Hum Neurosci* **8**. doi:10.3389/fnhum.2014.00699

Hayama HR, Vilberg KL, Rugg MD. 2012. Overlap between the Neural Correlates of Cued Recall and Source Memory: Evidence for a Generic Recollection Network? *J Cogn Neurosci* **24**:1127–1137. doi:10.1162/jocn_a_00202

Hindy NC, Ng FY, Turk-Browne NB. 2016. Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nat Neurosci* 1–7. doi:10.1038/nn.4284

Kafkas A. 2021. Encoding-linked pupil response is modulated by expected and unexpected novelty: Implications for memory formation and neurotransmission. *Neurobiol Learn Mem*. doi:10.1101/2021.01.04.425232

Kafkas A, Montaldi D. 2018a. How do memory systems detect and respond to novelty? *Neurosci Lett* **680**:60–68. doi:10.1016/j.neulet.2018.01.053

Kafkas A, Montaldi D. 2018b. Expectation affects learning and modulates memory experience at retrieval. *Cognition* **180**:123–134. doi:10.1016/j.cognition.2018.07.010

Kafkas A, Montaldi D. 2015. Striatal and midbrain connectivity with the hippocampus selectively boosts memory for contextual novelty. *Hippocampus* **25**:1262–1273. doi:10.1002/hipo.22434

Kafkas A, Montaldi D. 2014. Two separate, but interacting, neural systems for familiarity and novelty detection: A dual-route mechanism. *Hippocampus* **24**:516–527. doi:10.1002/hipo.22241

Ketz N, Morkonda SG, O'Reilly RC. 2013. Theta Coordinated Error-Driven Learning in the Hippocampus. *PLoS Comput Biol* **9**:e1003067. doi:10.1371/journal.pcbi.1003067

Kim G, Lewis-Peacock J a, Norman KA, Turk-Browne NB. 2014. Pruning of memories by context-based prediction error. *Proc Natl Acad Sci* **111**:8997–9002. doi:10.1073/pnas.1319438111

Kim H. 2011. Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies. *Neuroimage* **54**:2446–2461. doi:10.1016/j.neuroimage.2010.09.045

Kumaran D, Maguire EA. 2007. Which computational mechanisms operate in the hippocampus during novelty detection? *Hippocampus* **17**:735–748. doi:10.1002/hipo.20326

Li S, Cullen WK, Anwyl R, Rowan MJ. 2003. Dopamine-dependent facilitation of LTP induction in hippocampal CA1 by exposure to spatial novelty. *Nat Neurosci* **6**:526–531. doi:10.1038/nn1049

Lisman JE, Grace A a. 2005. The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory. *Neuron* **46**:703–713. doi:10.1016/j.neuron.2005.05.002

Long NM, Lee H, Kuhl B a. 2016. Hippocampal Mismatch Signals Are Modulated by the Strength of Neural Predictions and Their Similarity to Outcomes. *J Neurosci* **36**:12677–12687. doi:10.1523/JNEUROSCI.1850-16.2016

Meeter M, Murre JMJ, Talamini LM. 2004. Mode shifting between storage and recall based on novelty detection in oscillating hippocampal circuits. *Hippocampus* **14**:722–741. doi:10.1002/hipo.10214

Misaki M, Kim Y, Bandettini PA, Kriegeskorte N. 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage* **53**:103–118. doi:10.1016/j.neuroimage.2010.05.051

Murty VP, Shermohammed M, Smith D V., Carter RM, Huettel SA, Adcock RA. 2014. Resting state networks distinguish human ventral tegmental area from substantia nigra. *Neuroimage* **100**:580–589. doi:10.1016/j.neuroimage.2014.06.047

Oosterhof NN, Connolly AC, Haxby J V. 2016. CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. *Front Neuroinform* **10**:1–27. doi:10.3389/fninf.2016.00027

Oyarzún JP, Packard PA, de Diego-Balaguer R, Fuentemilla L. 2016. Motivated encoding selectively promotes memory for future inconsequential semantically-related events. *Neurobiol Learn Mem* **133**:1–6. doi:10.1016/j.nlm.2016.05.005

Peirce JW. 2007. PsychoPy-Psychophysics software in Python. *J Neurosci Methods* **162**:8–13. doi:10.1016/j.jneumeth.2006.11.017

R Development Core Team. 2008. R: A language and environment for statistical computing. http://www.r-project.org

Richter FR, Chanales AJH, Kuhl BA. 2016. Predicting the integration of overlapping memories by decoding mnemonic processing states during learning. *Neuroimage* **124**:323–335. doi:10.1016/j.neuroimage.2015.08.051

Searle SR, Speed FM, Milliken GA. 1980. Population Marginal Means in the Linear Model: An Alternative to Least Squares Means. *Am Stat* **34**:216–221. doi:10.1080/00031305.1980.10483031

Shohamy D, Adcock RA. 2010. Dopamine and adaptive memory. *Trends Cogn Sci* **14**:464–472. doi:10.1016/j.tics.2010.08.002

Shohamy D, Wagner AD. 2008. Integrating Memories in the Human Brain: Hippocampal-Midbrain Encoding of Overlapping Events. *Neuron* **60**:378–389. doi:10.1016/j.neuron.2008.09.023

Stoppel CM, Boehler CN, Strumpf H, Heinze H-J, Hopf JM, Düzel E, Schoenfeld MA. 2009. Neural correlates of exemplar novelty processing under different spatial attention conditions. *Hum Brain Mapp* **30**:3759–3771. doi:10.1002/hbm.20804

West BT, Welch KB, Galecki AT. 2014. Linear Mixed Models. Chapman and Hall/CRC. doi:10.1201/b17198

Wickham H. 2009. ggplot2. New York, NY: Springer New York. doi:10.1007/978-0-387-98141-3

Yassa MA, Stark CEL. 2011. Pattern separation in the hippocampus. *Trends Neurosci* **34**:515–525. doi:10.1016/j.tins.2011.06.006