Expectation violation dynamically engages a bottom-up encoding state

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

Abstract

Expectation violation has been shown to engage adaptive memory formation, resulting in better memory for unexpected information. In two experiments we tested whether this mechanism is engaged dynamically in a goal-irrelevant manner during retrieval, and how it affects trial-by-trial recognition. Participants encoded images of objects, and then learned a contingency between a cue and category (man-made or natural) with new objects. Targets and parametrically manipulated similar foils, comprising set events, were used at retrieval. In each retrieval trial a cue appeared, which either matched or mismatched (according to the established contingency) the following object, for which participants made an old/new decision. We found that unexpected events at retrieval were associated with increased activation along the ventral visual stream, whereas expected events engaged parietal regions of the core recollection network. For targets and most similar foils, we found an interaction between current and previous expectation status on memory performance, such that expected events following unexpected ones ($U_{prev}E_{curr}$) showed a boost in performance. This behavioural effect was associated with activation in the hippocampus, SN/VTA and occipital cortex. A combination of two unexpected events ($U_{prev}U_{curr}$) resulted in the poorest memory performance and was associated with increased activation in occipital cortex. Taken together, our findings suggest expectation violation engages an encoding mechanism, supported by bottom-up processing, in a task-independent manner. Therefore, when the goal is to retrieve information, the mnemonic consequences of the shift towards an encoding state is detrimental in real-time, but beneficial for subsequent similar events.

Abstract word count: 245

Introduction

Environmental setting and contextual features shape how we perceive and remember our daily life. In order to efficiently process the inputs it receives, the brain shifts between top-down and bottom-up information streams to balance processing sensory inputs versus utilising stored information. Bottom-up processing can reflect an encoding state, prioritising transformation of perceptual inputs into memories. Discrepancies between expected and observed events have been shown to engage such an encoding state in the hippocampus (Axmacher et al., 2010; Bein, Duncan, & Davachi, 2020; Kafkas & Montaldi, 2015; Kumaran & Maguire, 2007; Meeter, Murre, & Talamini, 2004). Nevertheless, it is still unclear whether this encoding state can be engaged dynamically, in a goal-irrelevant manner (i.e. peripheral or irrelevant to the decision made in the task). Here we examine whether expectation violation, taking place at retrieval and orthogonal to the recognition task, results in a shift towards an encoding state, as indicated by a gain on bottom-up processing. We further examined how this violation and ensuing bottom-up processing affect trial-by-trial recognition of similar events.

Predictive processing frameworks postulate bottom-up processing is modulated by top-down predictions (Friston, 2005). In memory research, prediction errors, or expectation violation, have been shown to engage hippocampal processing (Duncan, Ketz, Inati, & Davachi, 2012; Frank, Montemurro, & Montaldi, 2020; Hindy, Ng, & Turk-Browne, 2016; Kafkas & Montaldi, 2015; Kumaran & Maguire, 2007; Long, Lee, & Kuhl, 2016), which together with the midbrain dopaminergic system (Kafkas & Montaldi, 2018b; Lisman & Grace, 2005; Shohamy & Wagner, 2008), support 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 (Garrido, Barnes, Kumaran, Maguire, & Dolan, 2015; Greve, Cooper, Kaula, Anderson, & Henson, 2017; Gruber, Watrous, Ekstrom, Ranganath, & Otten, 2013; Li, Cullen, Anwyl, & Rowan, 2003; Long et al., 2016). Therefore, any additional resources diverted toward encoding in these scenarios is likely to boost later memory performance, similar to attention effects (for a review see Aly & Turk-Browne, 2017). However, to demonstrate a ubiquitous expectation-driven encoding mechanism, it is critical to provide evidence of a dynamic shift towards encoding in a

task-independent manner, for example, when retrieval is necessary. Under these circumstances, a shift towards encoding still serves an adaptive purpose, but it might no longer be 'beneficial' in terms of reporting memory for the unexpected event. Evidence for such task-independent encoding effects (e.g. expectation-driven encoding at the expense of retrieval) is currently lacking.

Given the malleability of existing memories following expectation violation (G. Kim, Lewis-Peacock, Norman, & Turk-Browne, 2014), it is also worth considering what the mnemonic consequences of a dynamic 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 in line with the view that the hippocampus continuously shifts between encoding and retrieval states (Buzsáki, 2002; Colgin, 2016; Hasselmo, Bodelon, & Wyble, 2002; Ketz, Morkonda, & O'Reilly, 2013). These states can be viewed as a pendulum, and certain environmental factors can disrupt the on-going swing between states, favouring one over another, to support effective processing. Previous research has shown a strategy cue (e.g. encode or integrate) can indeed foster hippocampal trade-off between encoding and retrieval of information (Richter, Chanales, & Kuhl, 2016). Here, we test whether such shifts can also occur following an implicit, task-irrelevant expectation violation, and their continuous mnemonic outcomes in the form of correct recognition of targets and similar foils.

Engagement of an encoding state should also be demonstrated in enhanced perceptual processing, as encoding relies on sensory inputs. Previous research has shown increased activity along the ventral visual stream in response to novelty/expectation violation (Hawco & Lepage, 2014; Kafkas & Montaldi, 2015; Stoppel et al., 2009). We have also shown that expectation-modulated memory is sensitive to perceptual similarity between existing 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 similar set events (target and similar foils), is necessary. This allows us to examine how expectation violation affects trial-by-trial memory decisions and representations (Duncan, Sadanand, & Davachi, 2012). For example, if the first unexpected target was missed due to a shift towards encoding, the following expected similar foil might benefit from the re-encoding of the target, and more likely to be correctly rejected.

Therefore, we can better tap the interplay between memory search and shift to encoding due to expectation violation, and how this relationship may affect subsequent processing (and recognition) of similar events.

In two experiments, we examined a) whether expectation violation dynamically engages an encoding state, b) whether this response is dependent on the perceptual similarity between inputs, and c) its mnemonic consequences. To do so, we used a paradigm in which expectation was manipulated at retrieval, implicitly and independently of the requirements of the recognition task. Following encoding of images of objects, 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) 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 sequence was violated. This design allowed us to answer two questions; first whether expectation violation engages a bottom-up encoding mechanism, even when it's not goal-relevant. Second, what the mnemonic consequences of this violation are in real time as well as for subsequent similar objects (interactions between previous and current set events' expectation status). We reasoned that an expectation-driven encoding could harm online retrieval processes, but aid recognition of subsequent similar events.

Experiment 1

Methods

Participants. 30 participants (4 males) between the ages 18-22 (M = 19, SD = 1.04) gave informed consent and took part in the experiment. 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) and recognition performance below chance (1). Data from 26 participants are reported below.

Materials. 78 images of natural (39) and man-made (39) objects were selected from the Similar Objects-Lures Image Database (SOLID; Frank, Gray, & Montaldi, 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 by Kafkas and Montaldi, 2018a). 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 part involved a 5-minute filled delay task, during which participants solved arithmetic problems.

In the third part, 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 part 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, meaning an 'old' response could be given for multiple set events (i.e. for target and foils).

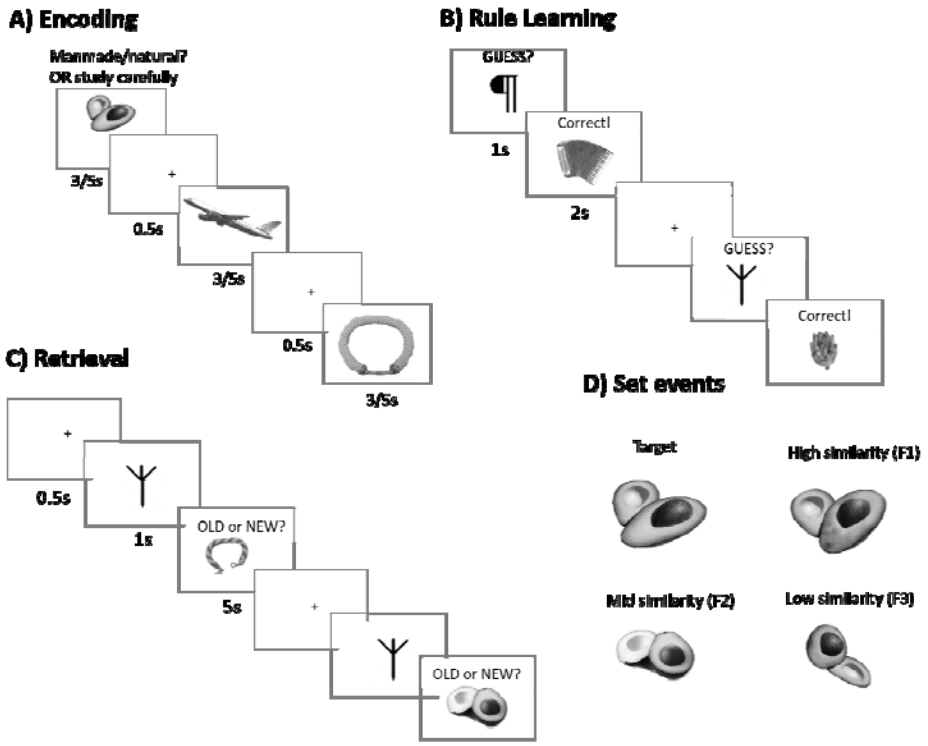

Figure 1 | Experimental design. A) In Experiment 1, during the first round of encoding, participants responded whether the object was man-made or natural. In the second encoding round, participants were asked to study the object carefully. In Experiment 2, each object appeared three times consecutively; in addition to the two presentation form 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. In Experiment 1 participants had 1 second to make a decision, and 3 seconds in Experiment 2 C) In the final retrieval task, the same cues were presented

before each set event, old objects (targets) and new similar foils (F1, F2, F3; only 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. Participants were instructed to ignore the symbols and respond whether the object is old or new**.** D) Example of set events.

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 (1-4) 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. Models were computed using the lme4 package (Bates, Mächler, Bolker, & Walker, 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 best-performing model (H_0: $\beta = 0$), we used an omnibus χ^2 Wald test (West, Welch, & Galecki, 2014), as implemented in the car package (Fox & Weisberg, 2018). Extraction and plotting of the effects reported below was conducted using the effects (Fox, 2003), emmeans (Searle, Speed, & Milliken, 1980) and ggplot2 (Wickham, 2009) packages. To examine the dynamic interaction between item similarity and expectation status, for each set event we devised three models of interest, modelling each event separately as a function of the preceding set event. 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.

Results

Predicting hits

For targets following F1 set events, we found significant main effects of the previous 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 ($\beta = -0.735$, $X^2(1) = 4.14$, $p = 0.042$). 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).

Predicting correct rejections of F1 (CR1)

For F1 events following targets, we found a trend towards an interaction between the two events' expectation status (β = -0.58, $X^2(1) = 3.59$, p = 0.058), with simple contrasts showing more CR1 for expected F1 events following unexpected targets, than unexpected F1 (z = 2.29, p = 0.022). 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).

Predicting correct rejections of F2 (CR2) and F3 (CR3)

F2 events were not affected by previous targets from the same set (all p's > 0.21). We found more CR2 for unexpected F2, following F1 and F3 events, compared to expected F2 (for F1 β = 0.35, $X^2(1) = 4.65$, p = 0.031; for F3 β = 0.32, $X^2(1) = 4.34$, p = 0.037). For F2 following F3 events we found more CR2 for unexpected F2 events (β = 0.323, $X^2(1) = 4.34$, p = 0.037). CR3 were not modulated by any preceding set events (all p's > 0.263).

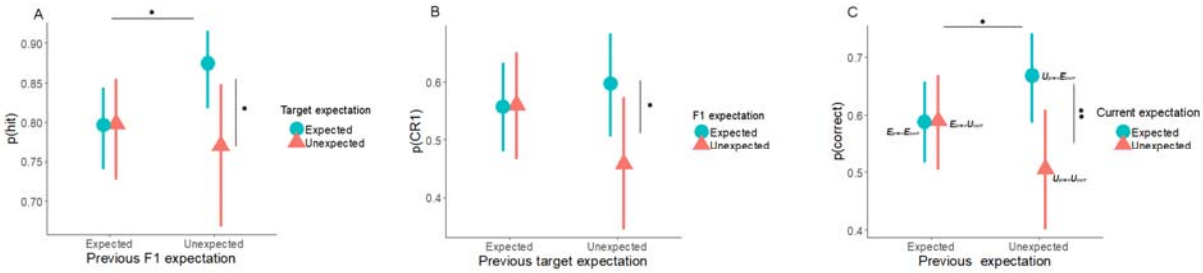

Figure 2 | Experiment 1 Results. A) Predicting hits. More hits were observed for expected targets following unexpected F1 foils, compared to unexpected targets following unexpected F1, and compared to expected targets following expected F1. **B) Predicting CR1.** More CR1 for expected F1 following unexpected targets, compared to unexpected F1 following unexpected targets. **C) 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 $E_{prev}E_{curr}$.

Concatenated targets and F1

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). 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 the previous one (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 ($\beta = 0.341$, $X^2(1) = 4.9$, $p = 0.027$), as well as a significant interaction between expectation status of the current and previous events ($\beta = -0.682$, $X^2(1) = 6.5$, $p = 0.01$). 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 ones, compared to following expected events ($E_{prev}E_{curr} < U_{prev}E_{curr}$; $z = -2.21$, $p = 0.027$).

Taken together, these results suggest expectation-violation dynamically shifts processing away from retrieval and towards encoding, and the mnemonic consequences of this shift are reflected on the subsequent set event, as a function of perceptual similarity. When the previous and current events are both unexpected ($U_{prev}U_{curr}$), we observed poor memory performance for targets and F1 foils. On the other hand, when an unexpected event is followed by an expected one ($U_{prev}E_{curr}$) a boost in performance was observed. To examine whether these effects engage the circuit involved in adaptive memory formation (hippocampus and midbrain; Kafkas & Montaldi, 2018; Shohamy & Adcock, 2010), as well as to test the notion 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.

Experiment 2

Methods

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 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 the 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. 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.

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 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, Connolly, & Haxby, 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 & Ranganath, 2019). As one participant's last trial at encoding could not be estimated, only 77 sets were analysed for this participant. Data were mean-centred and pairwise Pearson's correlations between all trials were calculated and Fisher z-transformed. 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.

<u>Behavioural Results</u>

Examining responses to targets and F1 events, we found a main effect of the previous set event's expectation status ($\beta = 0.309$, $X^2(1) = 4.13$, $p = 0.042$), 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 ($U_{prev}E_{curr} > U_{prev}U_{curr}$; $z = 2.68$ $p = 0.007$). Therefore, these results provide a direct replication of the results from Experiment 1. 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 ($\beta = 0.488$, $X^2(1) = 6.4$, $p = 0.011$), with more CR2 following unexpected targets. All other effects were not significant (all p's > 0.193). For F2 following F1 events, there was a trend towards a main effect of the previous F1 expectation ($\beta = 0.367$, $X^2(1) = 3.14$, $p = 0.076$), with more CR2 following unexpected F1 events.

<u>fMRI results</u>

Current by previous expectation interactions. 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 a previous 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), 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). There was also 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) for current unexpected events following previous expected ones, compared to current unexpected events, which followed an unexpected event ($E_{prev}U_{curr} > U_{prev}U_{curr}$). Critically, in both contrasts the examined set events had the same expectation status, and differed only on the expectation status of the previous event.

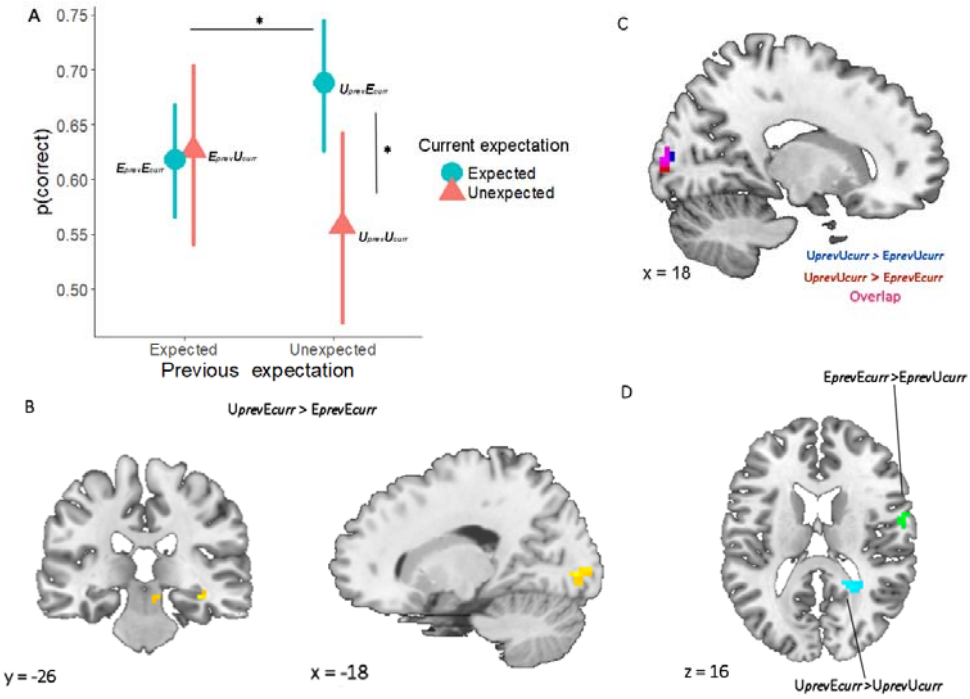

Figure 3 | Behavioural and neural expectation interactions for targets and F1 foils. A) Behavioural results. Replicating the results from Experiment 1, a current by previous expectation status interaction was observed, with $U_{prev}E_{curr}$ showing a boost in memory performance**. B) UprevEcurr > EprevEcurr 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).

Main effects of expectation. To explore whether expected and unexpected events, across event types, responses, and 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). 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), was observed for expected > unexpected events.

Expectation by memory interactions. 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.

Expectation by presentation order interactions. 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).

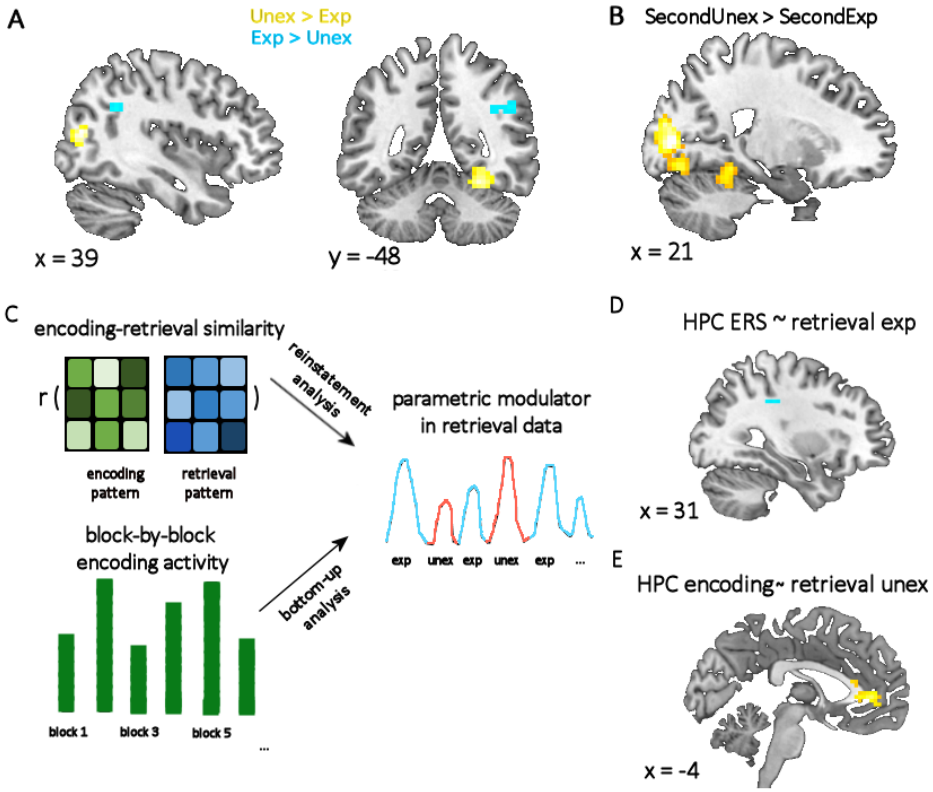

Figure 4 | Overall neural expectation effects. A) Main effect of expectation. Unexpected events, compared to expected ones, engaged more activation in right middle occipital cortex and right fusiform. Conversely, excepted events, compared to unexpected ones, engaged activation in right angular and supramarginal 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 supramarginal gyrus and right 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.

Bottom-up parametric encoding analysis. Using hippocampal encoding activity as a parametric modulator (Figure 4e), we found increased hippocampal encoding correlated with increased involvement of anterior cingulate cortex (x = 1, y = 39, z = 3, k = 173; Figure 4e) for unexpected events. We did not observe any activation for expected events at retrieval that was modulated as a function of hippocampal encoding activity.

Reinstatement analysis. Increased hippocampal encoding-retrieval similarity (Figure 4d) was associated with increased activity in middle temporal gyrus (x = -42, y = -36,

z = -3, k = 11) and supramarginal gyrus (x = 30, y = -30, z = 36, k = 9) during retrieval of expected events. We did not observe any activation for unexpected events at retrieval that correlated with hippocampal ERS.

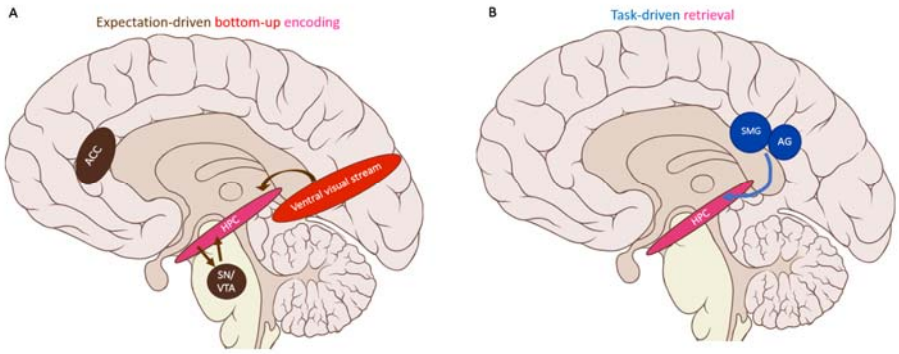

Figure 5 | Illustration of the brain networks involved in processing unexpected and expected events. A) Expectation-driven, task-independent encoding. Expectation violation engages bottom-up processing along the ventral visual stream, regardless of memory performance. The later mnemonic consequences of this shift towards encoding involve the hippocampus and midbrain dopaminergic regions, underlying beneficial memory performance. The ACC possibly interacts with the dopaminergic system to instigate encoding. B) **Task-driven retrieval.** In the absence of expectation violation (expected events), engagement of parietal regions of the core recollection network support reinstatement and memory performance.

Discussion

In two experiments, we used a contextual expectation manipulation during retrieval to examine how expectation violation dynamically modulates information processing and subsequent hippocampal-dependent memory. We found unexpected events were associated with increased involvement of regions along the ventral visual stream and correlated with hippocampal encoding activity. Furthermore, the combination of two unexpected similar events ($U_{prev}U_{curr}$) resulted in the poorest memory performance, suggesting expectation violation can hamper memory performance by preferentially engaging a bottom-up encoding state. Interestingly, we also found a later beneficial effect of contextual surprise, such that presentation of an unexpected event did not support its recognition, but it did boost 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 first direct evidence for dynamic engagement of a

bottom-up encoding mechanism, at the expense of retrieval, and its mnemonic consequences (Figure 5).

Expectation violation is often associated with improved memory performance, attributed to adaptive memory formation (Kumaran & Maguire, 2007; Lisman & Grace, 2005; Shohamy & Wagner, 2008). Our results lend support to the latter part of this notion, and extend it to account for task-independent effects. In line with the idea that increased weight is given to bottom-up inputs upon encountering a prediction error (Kafkas & Montaldi, 2018b; 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 preferentially respond to stronger levels of unexpected novelty (Kafkas & Montaldi, 2014), resulting in 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 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 $U_{prev}U_{curr}$).

Whilst occipital involvement was observed for both $U_{prev}E_{curr}$ and $U_{prev}U_{curr}$ events (i.e. irrespective of memory performance), $U_{prev}E_{curr}$ events were also associated 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 processing of perceptual inputs for memory formation. In terms of shifting the weight towards bottom-up inputs, the hippocampus is well placed at the apex of the sensory hierarchy and well connected with midbrain dopaminergic regions (Lisman & Grace, 2005), to be in the driver's seat. This could also explain why we did not observe hippocampal or SN/VTA activation for any other unexpected contrasts; their co-activation is likely associated with a boost in memory performance (Kafkas & Montaldi, 2015). Taken together, these findings suggest that expectation-driven increased weight on bottom-up inputs is task-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, Barnes,

Kumaran, Maguire, & Dolan, 2015; Greve, Cooper, Kaula, Anderson, & Henson, 2017; Gruber, Watrous, Ekstrom, Ranganath, & Otten, 2013; Li, Cullen, Anwyl, & Rowan, 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 (Duncan, Sadanand, et al., 2012; G. Kim et al., 2014). Therefore, the consequences of processing unexpected events may be contrasting in the short- and the long-term. In the short-term, while another task is performed, unexpected events can be disruptive, while in the long-term they can be beneficial, when memory for similar events is tested (as in the present study) or when memory for the unexpected stimulus is tested after a delay (Frank et al., 2020; Kafkas & Montaldi, 2015).

Most of the previous research into shifts between encoding and retrieval states was done in rodents (Hasselmo & Stern, 2014), or in humans where explicit instructions were utilised (Richter et al., 2016). Recently, Bein and colleagues (2020) showed 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. Bein et al. (2020) postulated the bias they observed towards one hippocampal state or another could be modulated by dopaminergic inputs (Kafkas & Montaldi, 2018b; Lisman & 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 (Bein et al., 2020). The mnemonic consequences of this shift are manifested in more accurate recognition of later highly similar expected events. Further support for engagement of 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 (Haque, Inati, Levey, & Zaghloul, 2020). We propose the increased gain on bottom-up inputs, together with the hippocampal-midbrain activations, are markers of encoding state triggered by expectation violation, even in the absence of explicit reward or instructions.

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 (but for subsequent expected events), it is worth considering why increased encoding of the U_{prev} event results in better memory performance for the subsequent E_{curr} event. One possibility is that the initial shift towards encoding results in a sharper representation of U_{prev}. This then representation stands out in memory, when the subsequent expected similar event is presented (Haque et al., 2020). Therefore, the current expected event could be compared to originally encoded objects, and the unexpected one from earlier at retrieval. The combined delta similarity between these representations can inform and aid the current recognition decision for the current expected event. An alternative interpretation for this finding is a prospective memory enhancement (Dunsmoor, Murty, Davachi, & Phelps, 2015; Oyarzún, Packard, de Diego-Balaguer, & Fuentemilla, 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 in 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 offline consolidation).

Critically, the interactions between current and previous events' expectation were observed only for targets and F1 (i.e. the highest degree of similarity to the encoded object). Moreover, these effects were unaffected 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; Haque et al., 2020). In such situations, the ability to process and compare fine details in current inputs and stored representations underpins correct recognition decisions (Yassa & Stark, 2011). Therefore, engagement of perceptual processing by expectation violation serves an adaptive purpose (Hawco & 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 relevant and therefore redundant (Frank et al., 2020).

Further support for the role of perceptual load in engaging encoding state can be seen in the occipital and fusiform effects for unexpected events presented second within the set, compared to their expected counterparts. Recognition decisions for these events encompass interference from the first set event, and may therefore lead to increased engagement of perceptual processing in order to better compare the current sensory input to the stored representations. Whilst the encoding effects of unexpected events observed along the ventral visual stream can be explained by the bottom-up account (Kafkas & Montaldi, 2018b; H. Kim, 2011), we also found hippocampal encoding activity correlated with ACC activation for unexpected events at retrieval. ACC has been implicated in cognitive control and error monitoring (Carter & van Veen, 2007). These functions are directly relevant to the unexpected events utilised in the current task, but they do not necessarily explain the correlation to hippocampal encoding reported above. More recent findings suggest that ACC interacts with the dopaminergic system to instigate encoding of rewarding information (Aberg, Kramer, & Schwartz, 2020; Elston & Bilkey, 2017). Our results are in line with this notion, suggesting ACC is not only involved in error detection, but also in subsequent encoding of the 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 a task-relevant retrieval, or an active engagement of a retrieval state, irrespective of task demands. Although the observed activation in regions of the core recollection network (Hayama, Vilberg, & Rugg, 2012) in response to expected events (in the univariate and in the ERS analyses) is indicative of reinstatement, it does not differentiate between the two alternatives. To elucidate whether these effects relate to task-specific retrieval, or a dynamic engagement of retrieval (irrespective of the task at hand), future studies could orthogonalise expectation and memorial state, therefore allowing both encoding and retrieval states to be engaged and measured independently. Alternatively, one could include a baseline (expectation-less) measure for retrieval performance, to be compared to retrieval of expected events. 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 elucidate how the hippocampus shifts between memory states (Bein et al., 2020; Colgin, 2016; Kafkas & Montaldi, 2018b). Finally, it would be of interest to characterise the temporal dynamics of the expectation violation and bottom-up processing. This approach could shed light on the role ventral visual processing stream plays in modulating memory circuits, and vice versa.

In conclusion, we found novel evidence for a ubiquitous encoding mechanism, triggered by expectation violation in a goal-irrelevant matter. Overall, unexpected events were associated with activation along the ventral visual stream, whereas expected events were associated with activations in parietal regions of the core recollection network. A combination 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 combination of two unexpected events was associated with poor recognition, and activation in occipital cortex. These results indicate that the expectation-driven shift towards an encoding state, supported by bottom-up processing, is detrimental to memory performance in real time, but beneficial for correct recognition of subsequent similar events.

References

Aberg, K. C., Kramer, E. E., & Schwartz, S. (2020). Interplay between midbrain and dorsal anterior cingulate regions arbitrates lingering reward effects on memory encoding. *Nature Communications*, *11*(1), 1829. https://doi.org/10.1038/s41467-020-15542-z

Aly, M., & Turk-Browne, N. B. (2017). How Hippocampal Memory Shapes, and Is Shaped by, Attention. In *The Hippocampus from Cells to Systems* (pp. 369–403). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50406-3_12

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*(1), 95–113. https://doi.org/10.1016/j.neuroimage.2007.07.007

Axmacher, N., Cohen, M. X., Fell, J., Haupt, S., Dümpelmann, M., Elger, C. E., … Ranganath, C. (2010). Intracranial EEG Correlates of Expectancy and Memory Formation in the Human Hippocampus and Nucleus Accumbens. *Neuron*, *65*(4), 541–549. https://doi.org/10.1016/j.neuron.2010.02.006

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bein, O., Duncan, K., & Davachi, L. (2020). Mnemonic prediction errors bias hippocampal states. *Nature Communications*, *11*(1), 3451. https://doi.org/10.1038/s41467-020-17287-1

Buzsáki, G. (2002). Theta Oscillations in the Hippocampus. *Neuron*, *33*(3), 325–340. https://doi.org/10.1016/S0896-6273(02)00586-X

Carter, C. S., & van Veen, V. (2007). Anterior cingulate cortex and conflict detection: an update of theory and data. *Cognitive, Affective & Behavioral Neuroscience*, *7*(4), 367–379. https://doi.org/10.3758/cabn.7.4.367

Colgin, L. L. (2016). Rhythms of the hippocampal network. *Nature Reviews Neuroscience*, *17*(4), 239–249. https://doi.org/10.1038/nrn.2016.21

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Dimsdale-Zucker, H. R., & Ranganath, C. (2019). Representational Similarity Analyses: A Practical Guide for Functional MRI Applications. *Handbook of*

Behavioral Neuroscience, *28*, 509–525. https://doi.org/10.1016/B978-0-12-812028-6.00027-6

Duncan, K., Ketz, N., Inati, S., & Davachi, L. (2012). Evidence for area CA1as a match/mismatch detector: A high-resolution fMRI study of the human hippocampus. *Hippocampus*, *22*(3), 389–398. https://doi.org/10.1002/hipo.20933.Evidence

Duncan, K., Sadanand, A., & Davachi, L. (2012). Memory's Penumbra: Episodic Memory Decisions Induce Lingering Mnemonic Biases. *Science*, *337*(6093), 485–487. https://doi.org/10.1126/science.1221936

Dunsmoor, J. E., Murty, V. P., Davachi, L., & Phelps, E. A. (2015). Emotional learning selectively and retroactively strengthens memories for related events. *Nature*, *520*(7547), 345–348. https://doi.org/10.1038/nature14106

Elston, T. W., & Bilkey, D. K. (2017). Anterior Cingulate Cortex Modulation of the Ventral Tegmental Area in an Effort Task. *Cell Reports*, *19*(11), 2220–2230. https://doi.org/10.1016/j.celrep.2017.05.062

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, *8*(15). https://doi.org/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. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01211-7

Frank, D., Montemurro, M. A., & Montaldi, D. (2020). Pattern Separation Underpins Expectation-Modulated Memory. *The Journal of Neuroscience*, *40*(17), 3455–3464. https://doi.org/10.1523/JNEUROSCI.2047-19.2020

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Garrido, M. I., Barnes, G. R., Kumaran, D., Maguire, E. A., & Dolan, R. J. (2015). NeuroImage Ventromedial prefrontal cortex drives hippocampal theta oscillations induced by mismatch computations. *NeuroImage*, *120*, 362–370. https://doi.org/10.1016/j.neuroimage.2015.07.016

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. N. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language*, *94*, 149–165. https://doi.org/10.1016/j.jml.2016.11.001

Gruber, M. J., Watrous, A. J., Ekstrom, A. D., Ranganath, C., & Otten, L. J. (2013). Expected reward modulates encoding-related theta activity before an event. *NeuroImage*, *64*(1), 68–74. https://doi.org/10.1016/j.neuroimage.2012.07.064

Haque, R. U., Inati, S. K., Levey, A. I., & Zaghloul, K. A. (2020). Feedforward prediction error signals during episodic memory retrieval. *Nature Communications*, *11*(1), 1–14. https://doi.org/10.1038/s41467-020-19828-0

Hasselmo, M. E., Bodelon, C., & Wyble, B. P. (2002). A Proposed Function for Hippocampal Theta Rhythm□: Separate Phases of Encoding and Retrieval Enhance Reversal of Prior Learning. *Neural Computation*, *14*, 793–817.

Hasselmo, M. E., & Stern, C. E. (2014). Theta rhythm and the encoding and retrieval of space and time. *NeuroImage*, *85*, 656–666. https://doi.org/10.1016/j.neuroimage.2013.06.022

Hawco, C., & Lepage, M. (2014). Overlapping patterns of neural activity for different forms of novelty in fMRI. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00699

Hayama, H. R., Vilberg, K. L., & Rugg, M. D. (2012). Overlap between the Neural Correlates of Cued Recall and Source Memory: Evidence for a Generic Recollection Network? *Journal of Cognitive Neuroscience*, *24*(5), 1127–1137. https://doi.org/10.1162/jocn_a_00202

Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nature Neuroscience*, (April), 1–7. https://doi.org/10.1038/nn.4284

Kafkas, A., & Montaldi, D. (2014). Two separate, but interacting, neural systems for familiarity and novelty detection: A dual-route mechanism. *Hippocampus*, *24*(5), 516–527. https://doi.org/10.1002/hipo.22241

Kafkas, A., & Montaldi, D. (2015). Striatal and midbrain connectivity with the hippocampus selectively boosts memory for contextual novelty. *Hippocampus*, *25*(11), 1262–1273. https://doi.org/10.1002/hipo.22434

Kafkas, A., & Montaldi, D. (2018a). Expectation affects learning and modulates memory experience at retrieval. *Cognition*, *180*(July), 123–134. https://doi.org/10.1016/j.cognition.2018.07.010

Kafkas, A., & Montaldi, D. (2018b). How do memory systems detect and respond to novelty? *Neuroscience Letters*, *680*(January), 60–68. https://doi.org/10.1016/j.neulet.2018.01.053

Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013). Theta Coordinated Error-Driven Learning in the Hippocampus. *PLoS Computational Biology*, *9*(6), e1003067. https://doi.org/10.1371/journal.pcbi.1003067

Kim, G., Lewis-Peacock, J. a, Norman, K. A., & Turk-Browne, N. B. (2014). Pruning of memories by context-based prediction error. *Proceedings of the National Academy of Sciences*, *111*(24), 8997–9002. https://doi.org/10.1073/pnas.1319438111

Kim, H. (2011). Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies. *NeuroImage*, *54*(3), 2446–2461. https://doi.org/10.1016/j.neuroimage.2010.09.045

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

Li, S., Cullen, W. K., Anwyl, R., & Rowan, M. J. (2003). Dopamine-dependent facilitation of LTP induction in hippocampal CA1 by exposure to spatial novelty. *Nature Neuroscience*, *6*(5), 526–531. https://doi.org/10.1038/nn1049

Lisman, J. E., & Grace, A. a. (2005). The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory. *Neuron*, *46*(5), 703–713. https://doi.org/10.1016/j.neuron.2005.05.002

Long, N. M., Lee, H., & Kuhl, B. a. (2016). Hippocampal Mismatch Signals Are Modulated by the Strength of Neural Predictions and Their Similarity to Outcomes. *The Journal of Neuroscience*, *36*(50), 12677–12687. https://doi.org/10.1523/JNEUROSCI.1850-16.2016

Meeter, M., Murre, J. M. J., & Talamini, L. M. (2004). Mode shifting between storage and recall based on novelty detection in oscillating hippocampal circuits. *Hippocampus*, *14*(6), 722–741. https://doi.org/10.1002/hipo.10214

Murty, V. P., Shermohammed, M., Smith, D. V., Carter, R. M., Huettel, S. A., & Adcock, R. A. (2014). Resting state networks distinguish human ventral tegmental area from substantia nigra. *NeuroImage*, *100*, 580–589. https://doi.org/10.1016/j.neuroimage.2014.06.047

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. *Frontiers in Neuroinformatics*, *10*(July), 1–27. https://doi.org/10.3389/fninf.2016.00027

Oyarzún, J. P., Packard, P. A., de Diego-Balaguer, R., & Fuentemilla, L. (2016). Motivated encoding selectively promotes memory for future inconsequential semantically-related events. *Neurobiology of Learning and Memory*, *133*, 1–6. https://doi.org/10.1016/j.nlm.2016.05.005

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

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

Richter, F. R., Chanales, A. J. H., & Kuhl, B. A. (2016). Predicting the integration of overlapping memories by decoding mnemonic processing states during learning. *NeuroImage*, *124*, 323–335. https://doi.org/10.1016/j.neuroimage.2015.08.051

Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population Marginal Means in the Linear Model: An Alternative to Least Squares Means. *The American Statistician*, *34*(4), 216–221. https://doi.org/10.1080/00031305.1980.10483031

Shohamy, D., & Adcock, R. A. (2010). Dopamine and adaptive memory. *Trends in Cognitive Sciences*, *14*(10), 464–472. https://doi.org/10.1016/j.tics.2010.08.002

Shohamy, D., & Wagner, A. D. (2008). Integrating Memories in the Human Brain: Hippocampal-Midbrain Encoding of Overlapping Events. *Neuron*, *60*(2), 378–389. https://doi.org/10.1016/j.neuron.2008.09.023

Stoppel, C. M., Boehler, C. N., Strumpf, H., Heinze, H.-J., Hopf, J. M., Düzel, E., & Schoenfeld, M. A. (2009). Neural correlates of exemplar novelty processing under different spatial attention conditions. *Human Brain Mapping*, *30*(11), 3759–3771. https://doi.org/10.1002/hbm.20804

West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear Mixed Models*. Chapman and Hall/CRC. https://doi.org/10.1201/b17198

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

Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525. https://doi.org/10.1016/j.tins.2011.06.006