nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-021-01206-5

Check for updates

Developmental differences in memory reactivation relate to encoding and inference in the human brain

Margaret L. Schlichting[1]✉, Katharine F. Guarino[2], Hannah E. Roome[3,4,5] and Alison R. Preston[3,4,5]✉

Despite the fact that children can draw on their memories to make novel inferences, it is unknown whether they do so through the same neural mechanisms as adults. We measured memory reinstatement as participants aged 7–30 years learned new, related information. While adults brought memories to mind throughout learning, adolescents did so only transiently, and children not at all. Analysis of trial-wise variability in reactivation showed that discrepant neural mechanisms—and in particular, what we interpret as suppression of interfering memories during learning in early adolescence—are nevertheless beneficial for later inference at each developmental stage. These results suggest that while adults build integrated memories well-suited to informing inference directly, children and adolescents instead must rely on separate memories to be individually referenced at the time of inference decisions.

Young adults reactivate memories when they encounter new related experiences. Such reactivation can facilitate memory integration, whereby related events experienced at different times are stored as overlapping memory traces[1–3]. Memory integration thus promotes forming links between memories that extend knowledge beyond direct observation. Memory integration in adults relies on hippocampus (HPC) and medial prefrontal cortex (PFC)[2,4–7] and has been shown to benefit behaviours such as inferential reasoning, which requires simultaneous consideration of multiple memories[3]. For example, when asked to derive a relationship that has not been directly observed but must be inferred across several prior events, adults benefit from having previously connected (or integrated) their memories at encoding[5,6,8–10]. However, inferential reasoning can also be accomplished via an alternative mechanism in which memories for the original experiences are stored separately to be later recombined[11–13] when making the inference itself. Such a retrieval-based mechanism depends on memories for the individual events but importantly does not rely on them having been integrated at encoding.

It has been suggested that young children's inference ability[14,15] arises predominantly from a retrieval-based mechanism. For example, children struggle disproportionately with reasoning given their memory performance[16,17], and they are insensitive to manipulations designed to promote integration during encoding[18]. Both of these findings are consistent with the idea that reasoning during childhood relies primarily on operations engaged during inference itself. Given that inferential reasoning is a predictor of academic success[19,20] and the real possibility that children approach it in a fundamentally different way, it is crucial that we understand the neural mechanisms supporting its improvement throughout development.

We suggest that how memories for related experiences are formed will depend on both (1) the refinement of HPC-based mechanisms that support the ability to flexibly reactivate neocortical representations of related memories during new learning[21,22] and (2) the frontal mechanisms available to mediate conflict among reactivated memories, the results of which will ultimately influence how memories are formed in HPC[23–25]. Importantly, we suggest that reactivation is a necessary but not sufficient condition for memory integration to occur, as additional mechanisms must be engaged after memory reactivation to ultimately link related memories according to their overlapping features. Our overarching hypothesis is that, developmentally speaking, reactivation and integration will emerge in succession, paralleling the maturation of HPC[26–30] and its PFC connections[31,32], respectively. As such, there will be a period—specifically, adolescence—during which individuals reactivate but do not integrate.

Both HPC and PFC along with their interconnecting pathways show a protracted developmental trajectory continuing into (at least) adolescence[26,27,31], with PFC being particularly late to mature[33]. We suggest that the emergence of the first step necessary for an adult-like integration mechanism—namely, memory reactivation during new encoding—would therefore require that HPC retrieval is flexible enough to allow for the reinstatement of related memories during a similar but not identical (that is, partially overlapping) new experience[34,35]. Such flexible retrieval may not mature until around 10 years of age[25,36], thereby preventing any top-down influence on HPC codes and leaving inference in children to be carried out entirely to the time of decision itself.

Along with the maturation of HPC retrieval mechanisms is expected to come a greater likelihood of memory reinstatement in adolescence; and yet, we suggest that such reactivation will nevertheless continue to have a different behavioural consequence than it does in adults. In particular, we suggest that reactivation of past memories in adolescence may yield memory competition that is resolved by suppression and an ultimate de-emphasis of the relationships among memories. We suggest that such a phenomenon in our task may be due to the combined influence of at least two factors. First, there is general maturation of top-down control networks during adolescence that has been linked to improvements in higher-order cognitive abilities[37–40]. Some have even observed that

[1]Department of Psychology, University of Toronto, Toronto, Ontario, Canada. [2]Department of Psychology, Loyola University Chicago, Chicago, IL, USA. [3]Center for Learning & Memory, The University of Texas at Austin, Austin, TX, USA. [4]Department of Psychology, The University of Texas at Austin, Austin, TX, USA. [5]Department of Neuroscience, The University of Texas at Austin, Austin, TX, USA. ✉e-mail: meg.schlichting@utoronto.ca; apreston@utexas.edu

adolescence may be a peak period for top-down control of memory behaviours, with adolescents showing enhanced engagement of lateral PFC relative to adults[41]. Such increased control of memory among adolescents in our task may correspond to an enhanced tendency to suppress related, interfering memories[25,32]. Consistent with this idea, past work in rodents has highlighted adolescence as a unique time during which previous memories are suppressed during new, similar experiences[42,43].

Second, there are developmental changes in the HPC[26–30] that may lead to differences in memory representation. In terms of overall memory quality, adolescence is associated with increases in precision[44,45], greater richness of episodic detail[46] and enhancements in recollective quality[47]—all potentially attributable to changes in HPC encoding[41,48–52]. Given these findings, we expect that adolescents will be nearly adult-like in their ability to remember individual experiences. However, they may not yet have the ability to flexibly link across related experiences due to the nature of HPC development: namely, that posterior HPC (pHPC) matures earlier than anterior[26–28]. Informed by past functional studies also showing greater reliance on pHPC in children and adolescents than in adults[53], we therefore suggest that adolescent memory will accordingly reflect precise pHPC representations[4,54–56] (the granularity of which has been shown to increase over this developmental window[57]) rather than integrated ones[4], which engage later-developing HPC[27] and medial PFC[31] mechanisms. Together, the increasing availability of control mechanisms enabling memory suppression in tandem with hippocampal biases towards separate storage of related memories may ultimately yield representations emphasizing the unique features of individual experiences[58,59]. Importantly, such memories can nevertheless support successful inference[1,4,8], as they may be particularly easy to access and recombine during the decision.

Here, we test these hypotheses in a functional magnetic resonance imaging (fMRI) study in typically developing children, adolescents and adults. We anticipated nonlinearity in the developmental trajectory[60], such that adolescents would rely on memory mechanisms for inference that are distinct from those used by either children or adults. We also underscore that the maturation of memory-based inference probably unfolds through gradual change in the availability of these different neural signatures rather than an abrupt transition between mechanisms with development, consistent with an 'overlapping waves' perspective more typically discussed in the context of overt strategy[61]. As such, here we characterized development continuously from childhood through early adulthood.

Results

Integration of new memories in a pair learning task. Eighty-six participants aged 7 to 30 years performed an associative inference task (Methods)[4–8,12,16,62,63]. Stimuli (faces, scenes and objects) were organized into groups of three—termed ABC triads—and presented to the participants as overlapping AB and BC pairs that repeated in alternation along with non-overlapping control pairs[6,63] (Fig. 1a). This design allowed us to use an fMRI pattern classification approach (multivoxel pattern analysis (MVPA)[64]) to decode reinstatement of the related C content type—which was either a face or a scene, depending on the triad—to test for the predicted developmental differences in flexible retrieval. We hypothesized that while adolescents and adults would reactivate related memories during new encoding, children under age 10[36] would not, despite the additional encoding opportunities afforded by repetition. We further reasoned that should related memories be successfully brought to mind, there may nevertheless be lingering developmental differences in the way that conflict between memories is resolved. In particular, we predicted that reactivation in adolescence would be uniquely associated with both an upregulation of control regions implicated in memory suppression during later repetitions of overlapping pairs and impeded performance due to added competition. In contrast, we predicted that reactivation would be behaviourally advantageous in adults, who may instead integrate.

After each study run, the participants completed self-paced inference and memory tests (Fig. 1b,c). In the inference test, the participants were asked to link A and C items that were indirectly related through their common association with B at both the general (category) and specific (item) levels. Hereafter, to limit the influence of guessing, we consider inferences as correct only if the participant made the correct selection for both the category-level and item-level judgements. Pair memory was assessed only at the specific, item level. Importantly, all participants were aware of and had practiced both memory and inference tests before beginning the first study to reduce the influence of age-related differences in strategic approach to the task.

The participants were highly accurate on both the inference (mean, 83.08%; range, 20.83–100%; 95% confidence interval (CI), (78.73, 87.43)) and pair memory (94.67%, 56.94–100%, (92.68, 96.66)) tests, with a developmental trajectory characterized by rapid improvements at younger ages followed by plateau at ceiling in adolescence (Fig. 2). We then interrogated whether memory varied across direct pair types (AB, BC and non-overlapping) to quantify differential encoding of overlapping versus non-overlapping pairs (Fig. 2a). There were significant effects of both age ($\chi^2(3)=34.11$, $P<0.001$) and trial type ($\chi^2(2)=7.09$, $P=0.03$; the interaction was not significant, $\chi^2(6)=10.13$, $P=0.12$), with trial type effects primarily driven by worse performance for BC pairs. Notably, this difference may be a function of stimulus type rather than overlap per se, as only BC pairs contain an object with a face or scene as opposed to two objects. Age was also related to response times (RTs) on correct trials, with adolescents and adults being faster than children (Fig. 2b; age effect: $\chi^2(3)=55.68$, $P<0.001$). There was no significant main effect of condition ($\chi^2(2)=2.00$, $P=0.37$), but there was a significant age-by-condition interaction ($\chi^2(6)=13.80$, $P=0.03$) such that children showed the smallest RT difference among conditions. While speculative, one possibility for the relatively smaller difference in RT across direct pair types for children is that they do not encode the overlapping AB and BC pairs in a way that reflects their shared relationships; rather, children may treat overlapping the same as non-overlapping pairs and encode them in pattern-separated memories[58,65].

We next compared developmental improvements in memory (collapsed across direct pair type) with those in inference. Accuracy (Fig. 2c) was higher on the pair memory test than on the inference test ($\chi^2(1)=34.28$, $P<0.001$), with the magnitude of this difference decreasing with development (age-by-test-type interaction: $\chi^2(3)=10.30$, $P=0.02$; there was also a main effect of age, $\chi^2(3)=31.35$, $P<0.001$). This result replicates our previous findings in a different sample[16] and highlights that while the task was within the abilities of all ages—that is, performance was well above chance for all trial types across the entire age range (Fig. 2a,c, confidence bands)—younger participants disproportionately struggled with inference. These behavioural findings suggest developmental differences in how participants approach inferential reasoning; however, further neural mechanistic insight into the specific source of the age-related differences requires the fMRI approach that we turn to next.

Identifying developmental differences in memory reactivation. One important clue as to when memories are being combined to make an AC inference—that is, during encoding in preparation for an upcoming decision or, conversely, later during the decision itself—might stem from how reactivation unfolds across repeated experiences with the same, overlapping associations. We hypothesized that reactivation changes over repetition would be particularly diagnostic for understanding how memory mechanisms in adolescents differ from those in both children and adults. On the basis of prior work[6], we hypothesized that reactivation of overlap-

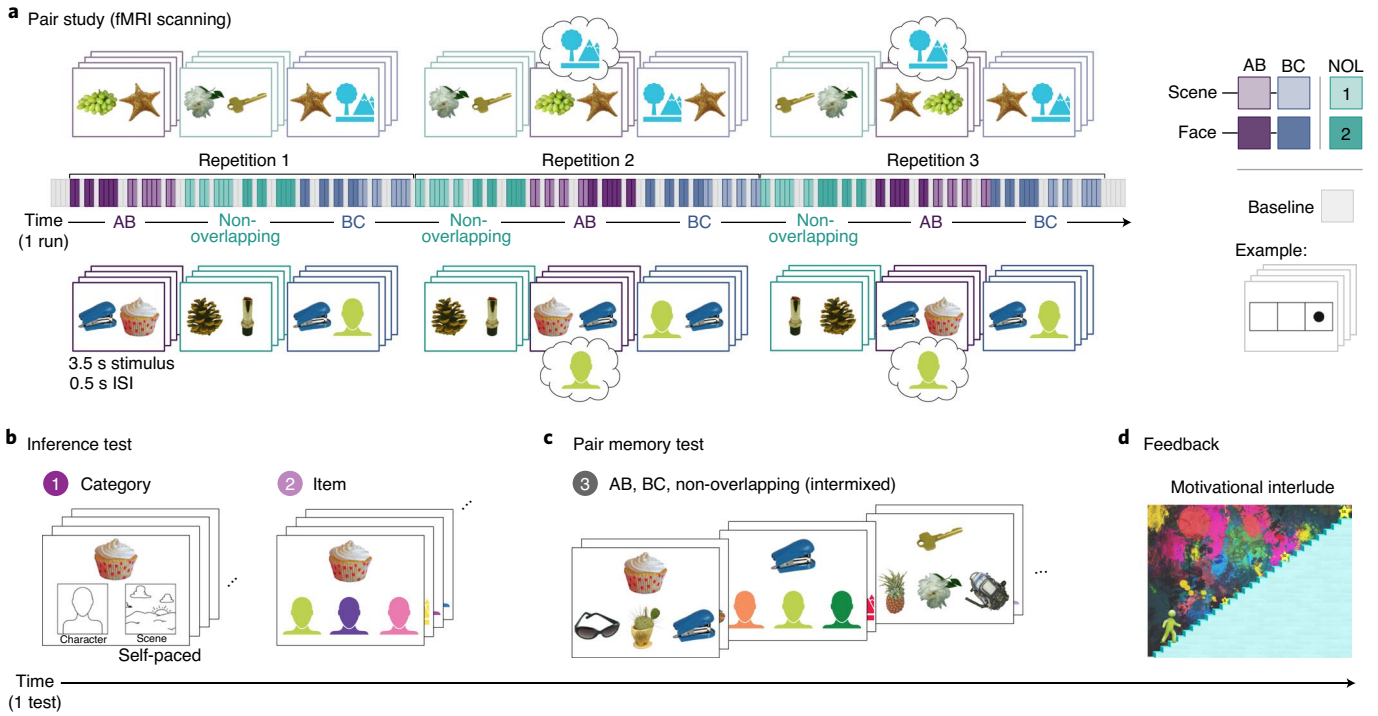

Fig. 1 | Experimental task. a, Each of four study–test cycles began with participants studying pairs for a later memory test. The timeline depicts every stimulus presentation (the pairs are in colour; the baseline is in grey) for one run. Overlapping (AB and BC) and non-overlapping (labelled NOL in the figure) pairs were blocked by type and jittered within each block to enable analysis at both the block and trial levels. AB pairs were our main trials of interest, with non-overlapping pairs serving as a content-matched baseline. Half of the AB pairs were each associated with a familiar scene (top) or face (bottom). The faces and scenes in the real experiment were images from popular movies and TV shows; they are replaced with uniquely coloured silhouettes in the figure for copyright reasons. The fill colours of the scene and face silhouettes indicate different identities. Moreover, the objects have been replaced with similar photos from the Bank of Standardized Stimuli[133,134] in all panels. **b**, After studying, the participants made self-paced inference judgements in which they first indicated the category (character) and then the identity (lime-green face, representing a particular character identity) of the C item indirectly related to the probe (cupcake). Foils (incorrect options) were other items from the same run that had occurred in the same position, condition and run (that is, C items were foiled by Cs that were members of different triads; different character identities are represented by different fill colours across the three options). **c**, Participants then completed a self-paced three-alternative forced-choice memory test for all studied pairs from the preceding study run. As in the inference test, the foils in the pair memory test were other items from the same run that had occurred in the same position, condition and run (again represented by different fill colours for the face and scene silhouettes). Images in **a**–**c** reproduced with permission from refs. [133,134]. **d**, After the tests, the participants received feedback about their memory performance before moving on to the next study with a new set of pairs. Specifically, the participants saw their previously selected avatar climbing a staircase, with the distance moved proportional to their memory performance. The participants' avatars continued to climb the staircase over four study–test cycles to earn bonus pay.

ping memories would increase across repetition in adults, promoting integration and inference (Fig. 3c, coral dots). Adolescents, in contrast, may reactivate overlapping memories during initial repetitions but then resolve this competition by accentuating the differences between overlapping memories in subsequent encounters with related pairs (Fig. 3c, magenta dots)[59]. This adolescent pattern would further differ from that of children, who we predicted would show no significant reactivation at all (Fig. 3c, purple dots). Such a result would be evidence of a developmental pattern in which adolescence is more than a stage between childhood and adulthood—rather, participants in this group may engage a fundamentally unique, adolescent-specific mechanism due to their particular neuromaturational state[60].

We sought to address these hypotheses by first training an MVPA classifier to identify patterns of activation in anatomically defined ventral temporal cortex (VTC) associated with face versus scene viewing (Fig. 3a,b). Classifier cross-validation performance was well above chance (one-sample *t*-test versus 0.5; mean = 89.42%, $t(79) = 177.75$, $P < 0.001$, Cohen's $d = 19.87$, 95% CI = (88.4, 90.4)), demonstrating our ability to discriminate between face and scene viewing on the basis of VTC activation patterns. Perhaps more importantly, age did not explain significant variance in classifier accuracy (model comparison using the Akaike information criterion (AIC): $AIC_{base} = 1{,}581.5$, $AIC_{age} = 1{,}500.2$, $F(3,76) = 1.37$, $P = 0.26$), such that we were similarly able to decode perception of face versus scene stimuli across the age range.

Related memories are reactivated during encoding. We next applied our trained MVPA classifier to fMRI patterns from the pair study task to decode the contents of memory (Fig. 3c). For each fMRI study pattern, the classifier returned continuous values reflecting the probability that it was associated with face processing or scene processing. Importantly, the participants were always viewing two objects during this task; however, the related content was either a face or a scene depending on the condition. For each participant, we generated a reactivation index for which values significantly above zero represent reliable reactivation of the related (more than the unrelated) content type during AB study. Repetition one serves as a baseline, as AB study occurs prior to encountering any overlapping (BC) face or scene content.

Across the group, irrespective of age, there was statistically significant reactivation on the second (one-sample *t*-test versus

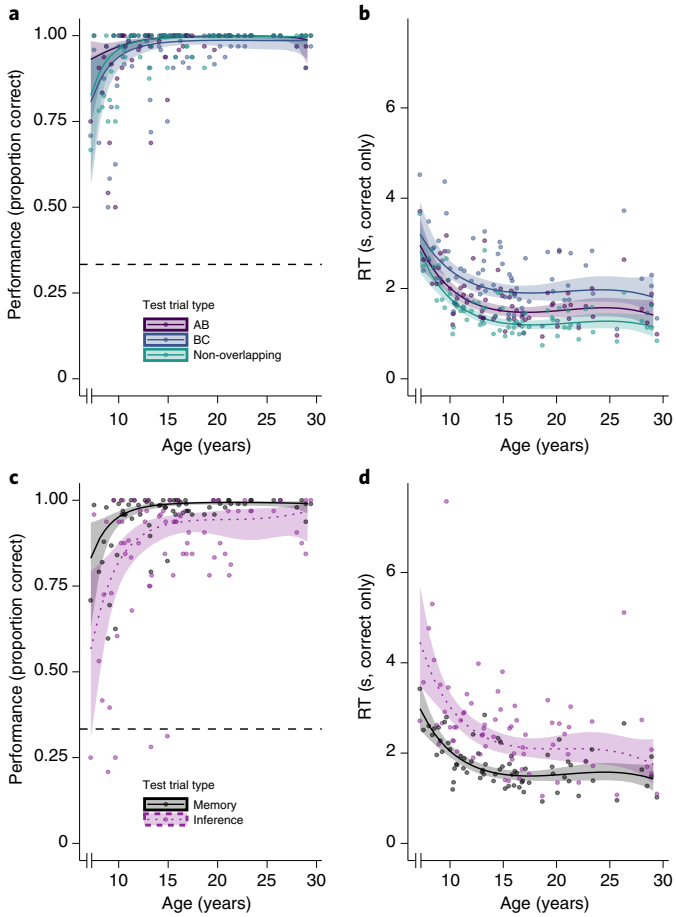

Fig. 2 | Task performance. a,**b**, Performance (accuracy) (**a**) and RT for correct trials (**b**) in the pair memory test (types AB, BC and non-overlapping) as functions of age. There were significant main effects of both age ($\chi^2(3) = 34.11$, $P < 0.001$) and trial type ($\chi^2(2) = 7.09$, $P = 0.03$) on accuracy, but no significant interaction ($\chi^2(6) = 10.13$, $P = 0.12$; 7,944 trials). For RTs, there was a significant effect of age ($\chi^2(3) = 55.68$, $P < 0.001$) and an age-by-condition interaction ($\chi^2(6) = 13.80$, $P = 0.03$); there was no significant main effect of condition ($\chi^2(2) = 2.00$, $P = 0.37$; 7,535 trials). **c**, Memory and inference performance (accuracy) as a function of age. There were significant main effects of test type (memory or inference; $\chi^2(1) = 34.28$, $P < 0.001$) and age ($\chi^2(3) = 31.35$, $P < 0.001$) as well as an age-by-test-type interaction ($\chi^2(3) = 10.30$, $P = 0.02$; 10,592 trials). **d**, RT as a function of age. There were significant main effects of age and trial type, such that RT decreased over development ($\chi^2(3) = 54.69$, $P < 0.001$) and memory was faster than inference ($\chi^2(1) = 16.65$, $P < 0.001$; there was no significant interaction, $P = 0.27$; 9,749 trials). For all panels, we used (generalized) linear mixed effects models to assess whether age, test trial type or their interaction was associated with accuracy and RT on individual trials. In all charts, the lines and bands depict model predictions and 95% CIs derived from the better-fitting models including age; the dots depict individual participant means (for accuracy) and medians (for RT) by condition. For all panels, $N = 86$ participants.

0; mean = 0.13, $t(83) = 4.54$, $P < 0.001$, Cohen's $d = 0.50$, 95% CI = (0.07, 0.18)) and third (mean = 0.08, $t(83) = 2.63$, $P = 0.01$, $d = 0.29$, 95% CI = (0.02, 0.13)) repetitions but not the first (mean = −0.002, $t(83) = -0.08$, $P = 0.94$, $d = 0.009$, 95% CI = (−0.05, 0.05)). Reactivation indices were also significantly greater on both the second ($t(83) = 3.55$, $P < 0.001$, $d = 0.39$, 95% CI = (0.06,0.20)) and third ($t(83) = 2.15$, $P = 0.03$, $d = 0.24$, 95% CI = (0.01,0.15)) repetitions than on the first. These results suggest that on average, the participants showed neural evidence of reactivating the associated content type after it had been introduced. We next turn to assessing developmental differences in this signature.

Development of reactivation reveals neural mechanistic shift. We hypothesized that the transition into adulthood would be accompanied by an increased tendency to form integrated memories that link related experiences during encoding. A mature integration mechanism would predict that reactivation, once it occurs, would be maintained or elevated across repetitions[6], and that such reactivation would be beneficial for subsequent inferential reasoning. Conversely, an active differentiation mechanism—whereby the commonalities across memories are detected yet de-emphasized during later encoding opportunities[4,59,66]—might yield initial reactivation that diminishes on subsequent repetitions, while in parallel, lateral PFC control systems ramp up to aid in interference resolution. Of note, such representations emphasizing the unique aspects of related memories may be used to make successful inferences while also supporting a host of other detail-oriented memory behaviours. We thus expected the direction of change in reactivation over repetitions—that is, whether reactivation increased (integration) or decreased (differentiation) from repetitions two to three—to vary with age. Importantly, brain imaging is required to achieve a direct quantification of such processes (potentially occurring outside of awareness) without influencing participants' strategies. Consistent with our hypothesis, we found that age explained additional variability in reactivation scores above and beyond repetition alone ($AIC_{base} = 24.24$, $AIC_{age} = 23.31$, $\chi^2(9) = 18.93$, $P = 0.03$). There was a significant age-by-repetition interaction ($\chi^2(6) = 16.55$, $P = 0.011$), demonstrating that how reactivation unfolded across repeated learning experiences was related to development (Fig. 4a; the results were similar when excluding statistical outliers ('Outlier exclusions'); interaction: $\chi^2(6) = 16.22$, $P = 0.013$).

Inspecting the resultant model fit curves showed that in late childhood and adolescence, there was a reliable decrease in reactivation from repetitions two to three. In fact, reactivation on repetition three did not exceed chance levels until mid-adolescence (16.09 years old), while repetition two reactivation emerged earlier (10.11 years old)—consistent with a differentiation scheme in this age range. The adolescent pattern contrasted with that in young adults ages 20 and older, who demonstrated the predicted integration signature, in which reactivation was above chance across both repetitions two and three. Consistent with prior observations of limited retrieval flexibility in children[36], we saw no statistically significant evidence of reactivation before age 10; moreover, children of this age showed significantly less reactivation than adolescents. Model predictions can also be visualized at four age points in Fig. 4b.

Together, these results show that there are fundamental shifts in the neural mechanism engaged during encoding of related memories. Children may not take advantage of commonalities across memories at all (as associations are never co-activated in the brain), while adolescents actively differentiate these experiences; in contrast, adults may tend to build up integrated representations that span related events.

Reactivation variability across memories relates to behaviour. We found developmental differences in the tendency to reactivate related memories during new learning. However, we also know that there exists variability at the specific pair level, such that integration and differentiation strategies can be used for distinct memories within a single individual[4,67]. We next leveraged this within-person variability to ask whether one type of encoding mechanism might be behaviourally advantageous for a given age—and critically, whether which mechanism is most beneficial might shift over development. In particular, under a differentiation mechanism, reactivation that is originally high and then drops could reflect initial memory

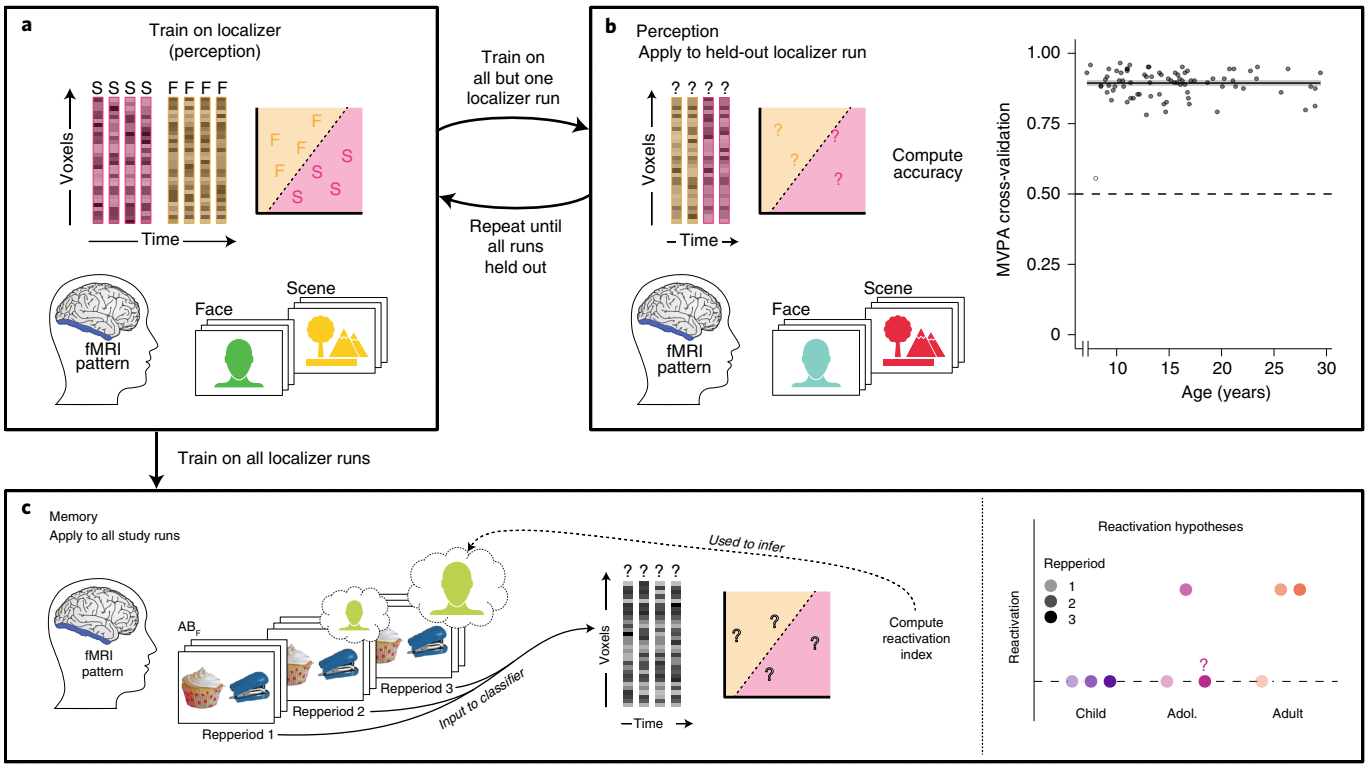

Fig. 3 | Perception and memory reactivation decoding analyses. a, For both perception and memory analyses, an MVPA classifier was trained on patterns of activation from a visual localizer task that occurred after, and used separate stimuli from, the main memory experiment. As in Fig. 1, the stimuli were replaced with uniquely coloured silhouettes for copyright reasons. fMRI patterns were extracted from VTC, and the classifier was trained to discriminate scene (pink) from face (orange) viewing. The boundary discriminating between the categories is depicted as a line separating face and scene viewing trials in a two-dimensional space. **b**, Perception decoding approach and results. Using cross-validation, classifiers were trained on $n-1$ localizer runs as in **a** and applied to the held-out nth run, where n is the number of localizer runs included for a given participant. Classification performance (accuracy; y axis) was high and did not significantly differ with age (x axis; model comparison: $AIC_{base}=1{,}581.5$, $AIC_{age}=1{,}500.2$, $F(3,76)=1.37$, $P=0.26$; see Methods and Supplementary Information for the details). Note that one outlier (age = 8.04 yr) was identified as showing accuracy that was not reliably above chance and was >4 s.d. below the mean (open circle); because such low performance on the training dataset precludes our ability to interpret the results of any application to a different task, the data from this participant were excluded from all subsequent analyses. $N=81$ participants are shown in the figure. **c**, Memory decoding approach. Left, the classifier trained on all localizer runs was applied to the fMRI study task patterns. We summarized the classifier evidence (probabilities) across AB trials by computing a single reactivation index per participant and repetition, which was defined as face minus scene evidence for AB_F trials plus scene minus face evidence for AB_S trials (that is, the interaction term). A reactivation index reliably above zero indicates that classifier evidence depends on trial type. Right, predictions for reactivation as a function of repetition for children (purple; no significant reactivation for any repetition), adolescents (magenta; reactivation on repetition two followed by potential drop-off on repetition three) and adults (coral; significant reactivation on repetitions two and three). The objects shown in the figure are from the Bank of Standardized Stimuli[133,134]. Images reproduced with permission from refs. [133,134].

co-activation followed by later suppression and strengthening of the individual AB and BC associations; such a memory scheme could support successful inference at retrieval. In contrast, inference through integration would suggest that pairs showing reactivation enhancements over time should be most likely to be correct.

We hypothesized that, if there are developmental differences in how overlapping memories are represented and used to support inference, they might become apparent when looking at how reactivation is related to performance within individuals, on a trial-by-trial basis. We asked whether variability in the degree to which reactivation changed from repetitions two to three for a particular pair was associated with the probability of making a correct inference. Consistent with our hypothesis, we found a significant interaction between age and reactivation change score in inference accuracy (Fig. 4c; interaction: $\chi^2(3)=8.13$, $P=0.043$; effect of age: $\chi^2(3)=20.60$, $P<0.001$; model comparison: $AIC_{base}=1{,}773.1$, $AIC_{age}=1{,}761.8$, $\chi^2(6)=23.36$, $P<0.001$). The nature of the interaction was such that among young adults, increasing reactivation from repetitions two to three was associated with a higher probability of making a correct response compared with when reactivation declined. In contrast, younger participants showed the opposite pattern: correct inferences were more likely on those trials showing reactivation decreases. The results were similar, albeit no longer exhibiting a statistically significant interaction, when outliers were removed (interaction: $\chi^2(3)=7.68$, $P=0.053$; effect of age: $\chi^2(3)=17.36$, $P<0.001$; model comparison: $AIC_{base}=1{,}709.2$, $AIC_{age}=1{,}700.8$, $\chi^2(6)=20.43$, $P=0.0023$).

The ability to benefit from reactivation during encoding (that is, show a positive slope on the reactivation change–accuracy relationship) thus seems to emerge in early adolescence, sometime between 10 and 15 years of age. Of note, this age range is approximately the same as the one over which, on average, the participants showed reactivation initially on repetition two that declined on repetition three. This finding suggests that adolescents are engaging in a mechanism that is fundamentally different from adults—yet, it is one that does confer behavioural advantage.

Reactivation impacts frontoparietal and hippocampal activation. We found that changes in the level of ventral visual stream (that is, VTC) reinstatement of previously stored memories over repetitions

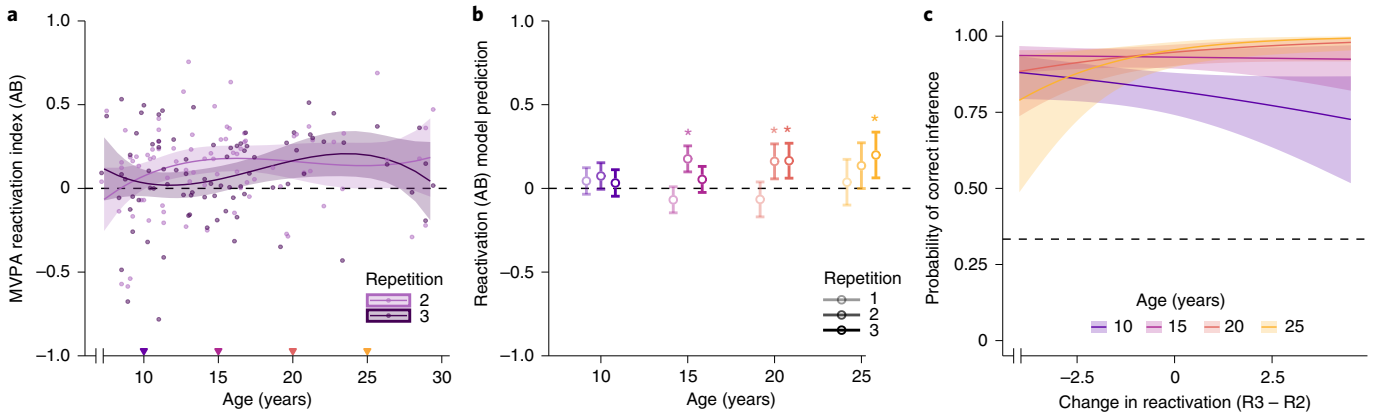

Fig. 4 | Memory reactivation decoding results. a, Blockwise decoding results showing reactivation (*y* axis) as a function of age plotted continuously (*x* axis). Age significantly improved the model fit beyond the base model including only repetition ($AIC_{base} = 24.24$, $AIC_{age} = 23.31$, $\chi^2(9) = 18.93$, $P = 0.03$); in the better-fitting model with age, there was a significant age-by-repetition interaction ($\chi^2(6) = 16.55$, $P = 0.011$). Repetition one serves as a baseline, as AB study occurs prior to encountering any overlapping (BC) face or scene content, and is not plotted here for the sake of simplicity. The coloured inverted triangles along the *x* axis indicate age points at which model predictions are shown in the subsequent panels. Adults maintain reactivation of related content across encoding repetitions, whereas adolescents show reactivation only on repetition two (light purple line; not significant on repetition three, shown in dark purple). Children show no significant evidence of reactivation. **b**, Model predictions from **a** visualized at four age points (10, 15, 20 and 25 years) across all three repetitions (light to dark). The asterisks denote age points and repetitions for which the model predictions are significantly above 0, indicating reliable reactivation according to the better-fitting model. The plots in **a** and **b**, represent 252 observations across $N = 84$ participants. **c**, Applying our classifier to individual trials rather than blocks yielded reactivation scores associated with each repetition of each specific pair. We found evidence for developmental differences (ages are shown by line colour and correspond to **b**) in the direction of the within-participant relationship between reactivation change from repetitions two to three (*x* axis) and subsequent inference performance (*y* axis; interaction: $\chi^2(3) = 8.13$, $P = 0.043$). Specifically, while adults (coral and orange) were more likely to get an inference decision correct when reactivation increased from repetitions two to three (>0 on the *x* axis), children (10 yr, purple) showed the opposite pattern—reactivation decreases (<0 on the *x* axis) were associated with a greater probability of correct inference at younger ages. There was also a main effect of age, such that inference accuracy was greater for older than for younger participants ($\chi^2(3) = 20.60$, $P < 0.001$). The figure displays model predictions at specific, user-defined age points; however, within the model, age was treated continuously. The plot in **c** represents 2,528 observations across $N = 84$ participants.

predicts performance on a trial-by-trial basis—but that critically, the nature of this relationship changes over development. We next asked whether reactivation in VTC during repetition two was associated with subsequent (repetition three) neural engagement. In other words, where in the brain is VTC reactivation associated with later activity levels? We reasoned that reactivation might be resolved differently in the brain depending on one's developmental stage. In particular, reactivation in the face of an inability to integrate should drive increased engagement of brain regions involved with memory suppression and interference resolution, such as inferior frontal gyrus (IFG). In tandem, one might expect initial reactivation to be associated with decreased later engagement of regions reflecting memory reinstatement, such as parietal cortex[24,68], which would be further consistent with a suppression mechanism.

We asked this question using a voxelwise general linear model (GLM) in which reactivation for a particular trial on repetition two (mean-centred within participants) was included as a parametric regressor in predicting fMRI activity on repetition three. (Importantly, we restricted our consideration of the relationship between reactivation and engagement to those measures observed on different repetitions. Beyond the theoretical reasons explained above, this choice also ensures independence of our measures and thus reduces the likelihood of spurious relationships reflecting more general neural fluctuations.) Clusters therefore represent regions for which there is a significant correspondence between the degree of reactivation on the preceding (second) repetition and activity during the final (third) repetition. The only region to show a significant effect across the group was left pHPC. This region showed a reliably negative relationship—that is, more reactivation on repetition two was associated with less pHPC activation on repetition three (intercept: $F(1,83) = 14.92$, $P < 0.001$). However, this effect was not significantly related to age (Fig. 5a; $P = 0.39$), consistent with observations that pHPC matures early, showing signs of being structurally developed in early childhood[27].

We additionally tested for regions in which the relationship changed with development (positively or negatively) by including age as a parametric regressor in the group-level statistical models. This analysis revealed two significant regions: bilateral parietal cortex (Fig. 5b) and bilateral IFG (Fig. 5c). In parietal cortex, a negative reactivation–engagement relationship in children and young adolescents was attenuated to no relationship in adults, consistent with the notion that only younger participants will suppress the internally generated content on subsequent encounters (left hemisphere (LH) effect of age: $F(3,80) = 5.07$, $P = 0.003$; right hemisphere (RH) effect of age: $F(3,80) = 5.35$, $P = 0.002$)[25]. Such an interpretation relates to the role of parietal cortex in reinstating high-fidelity memory representations in a way that is behaviourally relevant and influenced by top-down goals[24,68]. Reduced memory reinstatement at later points during study might be particularly advantageous for those memories that were initially reactivated most strongly, reflected in the fact that they are associated with relatively less parietal engagement on repetition three. IFG (bilaterally) showed the opposite pattern: relationships were positive in children and young adolescents but negative in adults (LH effect of age: $F(3,80) = 6.82$, $P < 0.001$; RH effect of age: $F(3,80) = 4.59$, $P = 0.005$). Such a result is in line with the interpretation that younger participants upregulate regions associated with interference resolution in response to high initial reactivation—perhaps aiding with active disambiguation of related memories[4]. The results were similar after removing outliers (parietal—LH effect of age: $F(3,78) = 3.44$, $P = 0.021$; RH effect of age: $F(3,78) = 4.11$, $P = 0.009$; IFG—LH effect of age: $F(3,78) = 5.39$, $P = 0.002$; RH effect of age: $F(3,79) = 3.97$, $P = 0.011$).

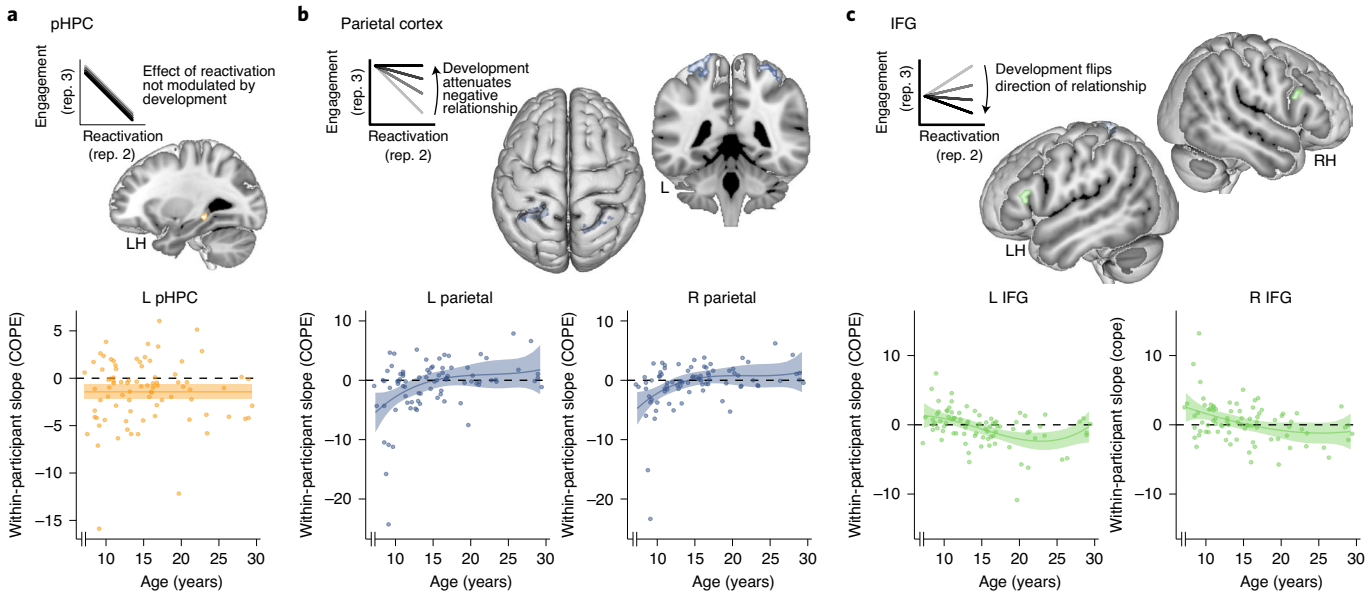

Fig. 5 | fMRI activation varies as a function of reactivation on preceding study repetition. a, pHPC showed a significant negative relationship between reactivation on repetition two and engagement on repetition three, such that greater within-participant evidence for reactivation in VTC was associated with less engagement of pHPC on the subsequent encoding experience (cluster size: 16 voxels, significant within HPC anatomical region of interest (ROI); $F(1,83) = 14.92$, $P < 0.001$). The nature of this relationship did not significantly differ over development ($P = 0.39$). **b**, In parietal cortex, there was a negative relationship between initial reactivation and subsequent engagement that was unique to the child and young adolescent ages (LH effect of age: $F(3,80) = 5.07$, $P = 0.003$; RH effect of age: $F(3,80) = 5.35$, $P = 0.002$). In other words, the negative relationship present in the children was attenuated over development to the point of being absent in adults. The effects were similar in the LH (147 voxels) and the RH (142 voxels; both L and R clusters are significant at whole-brain, grey-matter level). **c**, IFG also showed developmental effects (LH effect of age: $F(3,80) = 6.82$, $P < 0.001$; RH effect of age: $F(3,80) = 4.59$, $P = 0.005$), with a positive reactivation–engagement relationship in children and younger adolescents (especially in the RH; 35 voxels, significant within IFG anatomical ROI) giving way to a negative relationship in older adolescents and adults (especially in the LH; 26 voxels, significant within IFG anatomical ROI). For all panels, the regions were selected for showing either a main effect of reactivation (**a**) or a reactivation–engagement relationship that differed with age (**b**,**c**). For all panels, $N = 84$ participants. COPE, contrast of parameter estimate.

Discussion

We show that developmental differences in memory mechanisms influence how individuals of different ages make inferences about related episodes. Notably, we found that early adolescence was a unique period marked by initial reinstatement of memories during learning followed by later suppression—a signature consistent with differentiation of overlapping memories at this point in development. In contrast, adults showed enhanced reactivation consistent with integration at encoding, while children showed no significant evidence of reactivating at all and may store memories separately. These different memory mechanisms conferred age-specific behavioural advantages for inference: while suppressing reactivation benefitted those at the younger ages, enhancement was associated with correct inferences among adults. Interestingly, these differences emerged despite all participants being fully aware of the task structure and upcoming inference, thereby reducing the possibility that age-related differences in detecting overlap would be driving our effects. However, one limitation is that we did not assess the influence of overt strategy in this task; as such, whether younger learners can engage an integration mechanism when explicitly instructed to do so—or whether, as we would predict, their neural system acts as a fundamental limitation on this ability—remains an open question.

That children under 10 years failed to reactivate memories during learning is consistent with prior work suggesting immaturity of HPC retrieval mechanisms before this age[36,69]. Importantly, this lack of memory reactivation was observed in the context of our ability to decode perception at all ages, during both the localizer (Fig. 3b) and BC encoding trials in the main memory task (Supplementary Fig. 5a). While this finding may seem incompatible with past work showing that children can benefit from learning new information that relates to their prior knowledge[70,71], we suggest that this may be explained by certain features of our task that we designed to tap HPC mechanisms. A number of studies have shown subtle changes in HPC structure that continue into adulthood[16,26,28,48,72] and parallel behavioural gains in associative, detailed and recollective memory behaviours[49,73,74]—that is, those that depend on HPC[75,76]. We suggest that reactivation in our task requires a level of retrieval flexibility probably not present in children[36], who might be less apt to bring to mind a related memory (here, BC associations) when confronted with a similar but not identical new experience (AB). In particular, the related experiences in our task are by design partially overlapping, meaning they (for example, A$\underline{B}$) provide only a partial match to the to-be-retrieved trace ($\underline{B}$C). Additional features of our task such as the relatively limited amount of encoding experience and large number of arbitrary pairs relative to related paradigms[14] might have further decreased the likelihood that children would reactivate related memories while encoding overlapping events. Future work will be needed to understand how this mechanism scales up to explain how more well-established, complex knowledge structures formed over extended experience may scaffold new learning in children[77,78].

This lack of reactivation in children would mean that two memories for overlapping experiences are formed in the same way as those for two non-overlapping experiences[58], because related memories are never co-activated. Such a mechanism is consistent with our behavioural results, in which RTs did not differ for AB versus non-overlapping pairs, suggesting neither facilitation nor interference as a result of overlap. It would thus follow that the separately encoded but related memories for AB and BC associations would be stored and then separately accessed and recombined when faced

with the AC inference decision; this hypothesis might suggest that retrieval-phase rather than encoding-phase neural signatures in the youngest children would be most related to inference success. One limitation of this work is that we are not able to test this hypothesis directly because we did not acquire fMRI data during the inference test; therefore, it remains an interesting question for future study.

Adolescents showed evidence of initial reactivation, consistent with a level of HPC retrieval flexibility surpassing that of children. However, it is important to note that we measured reinstatement at the category level, reading out patterns in VTC as the product of HPC operations; we did not quantify the reinstatement of particular memories in HPC directly, which would require a different experimental design. With this limitation in mind, our results nevertheless converge with recent evidence showing adult-like HPC retrieval signatures in 13- and 14-year-olds[79]. More broadly, we saw evidence in adolescents for a unique neural encoding mechanism that differed from those of both children and adults. Of note, our results suggest that the adolescent period is a distinct stage[60] of memory development—not simply an intermediate step between childhood and adulthood as has been suggested in prior behavioural reports[16,17]. Combining our controlled behavioural task with an fMRI decoding approach, we have been able to characterize memory reactivation over development to reveal this insight into the adolescent brain.

Our task included overlapping pairs that allowed us to ask how encoding and retrieval interact to influence memory formation in development. In adolescents, initial reactivation was followed by a notable drop back to baseline during a subsequent learning experience. Further reasoning that high levels of reactivation would elicit competition among memories and differential engagement of control regions (particularly among adolescent learners), we found that greater reactivation was associated with increased IFG engagement in children and adolescents, accompanied by decreased recruitment of parietal cortex. These findings align well with IFG-guided suppression of memories activated in parietal cortex in this age range[25] and are broadly consistent with prior work highlighting developmental differences in controlled aspects of memory in general[80,81] and IFG in particular[41,82] that track age-related memory improvements. Our findings go beyond prior work to provide a key mechanistic example of how controlled encoding operations might contribute not only to the quality of memories stored but also to their contents and organization.

We suggest that in adolescents, reinstatement followed by subsequent study will weaken the connections between memories, as has been suggested previously[59], leading to memories for related experiences becoming more distinct from one another than two unrelated experiences[4,58,66] across repetitions[4,66,83,84]. We propose that differentiated representations are beneficial to inferential reasoning in adolescents and yet simultaneously require that they are engaging a fundamentally different mechanism from adults—namely, one in which they recombine memories at retrieval[85]. Such a proposal is in line with previous work on the development of reasoning, which highlights that ongoing maturation of controlled retrieval processes (selecting individual task-relevant memories) supported by IFG[86] and frontoparietal connections[87] underlies the performance gains observed into adolescence. Here, we extend these ideas by jointly incorporating memory and reasoning components in our task, highlighting that in addition to these retrieval differences, there is important developmental change in memory organization due to ongoing maturation of complex encoding mechanisms. Our approach thus links memory with reasoning literatures to show how traditionally conceptualized memory mechanisms guide how knowledge is organized—and therefore ultimately constrain how we might use knowledge to make flexible decisions.

The ability to make decisions that span multiple memories is a critical component of behavioural flexibility. In children, this ability is related to academic achievement[19,20], underscoring that the importance of understanding developmental change in this mechanism goes well beyond the lab. Here, we provide neural support for previous suggestions that children do not store memories with respect to their shared content, and we extend this framework into an understudied period of memory development to uncover an adolescent-specific neural phenomenon. Our results directly linking memory operations to the later ability to reason about those memories represent an important step towards bridging these literatures. More directly, these data suggest that child, adolescent and adult learners may rely on different mechanisms to achieve maximal behavioural flexibility—an idea that might be tested in future research and educational settings.

Methods

Participants. All experimental procedures were approved by the Institutional Review Board at the University of Texas at Austin. One hundred and twenty-five volunteers ranging in age from 6 to 30 years (actual range, 6.41–29.33) made up the cross-sectional sample who participated in a behavioural screening session ('Experiment overview'). Adult participants provided informed consent, and permission was obtained from one or more parents or guardians of minor participants (that is, individuals under the age of 18 years). Minors additionally provided informal assent. The participants were compensated US$10 per hour for the first session (mock scanner) and US$25 per hour for the second session (MRI); they also had the opportunity to earn an additional US$5–US$15 in bonus pay based on performance during the MRI session ('Memory task', 'Motivational interlude' section).

Of this initial group of 125 volunteers, 97 returned to the lab for the MRI session. Reasons for exclusion prior to the MRI session were: opted out or otherwise unable to schedule the scan session ($N=6$ minors and 8 adults (18 years or older)); had a Child Behavior Checklist Total Problems Score ($N=5$ minors) or Symptom Checklist 90-Revised Global Severity Index ($N=4$ adults) in the clinical range; left-handedness ($N=1$ minor); had contraindication(s) to MRI ($N=2$ adults); and diagnosed with a psychiatric condition or learning disability ($N=2$ minors). No participants scored below our inclusion threshold for IQ (>2 s.d. below the mean of FSIQ-2).

Of those 97 participants who were scanned, 11 were excluded from all further analyses for the following reasons: did not provide at least two fMRI runs of the encoding task due to terminating the session early ($N=3$ minors) or excessive motion, defined as fewer than two encoding runs with less than one-third of the time points exceeding our framewise motion threshold (see below; $N=6$ minors); incidental finding ($N=1$ minor); and technical difficulties with data acquisition ($N=1$ minor).

The final sample reported here includes 86 individuals whose ages on the date of MRI ranged from 7.16 to 29.42 years. Minors made up most of our sample ($N=65$ participants), and efforts were made to achieve approximate sex balance among both minors (35 females) and adults ($N=21$ total participants; 11 females). Our target sample size was 21 in each of four age bands: children 7–10 years, younger adolescents 11–14.5 years, older adolescents 14.5–17 years and adults 18–30 years. Power analyses using data[4] from a similar task showed that $N=21$ participants would yield 80% power to detect the behavioural effect of a within-participant manipulation of integration at the group level (Cohen's $d=0.56$); as such, we recruited participants until we had a minimum of 21 per band that could be included in our primary reactivation analysis. Our sample size also aligns with prior developmental work on a similar topic[16,17]. Two participants were excluded from the reactivation analysis specifically, and as such, our overall sample size was slightly larger than this minimum at 86 participants. Note that these age bands were arbitrarily defined and used only to ensure even sampling across the age range, with greater representation among the narrower bands for the developing groups (7–17) relative to the adult group; however, all analyses reported here treat age as a continuous variable.

In addition to participant-level exclusions, we excluded study runs that were (1) high motion, defined as more than one-third of the volumes with fast motion ('Motion-related participant-level and run-level exclusions'), (2) incomplete or (3) associated with poor subsequent memory, defined as test performance for the direct pairs (AB, BC and non-overlapping) not reliably above chance (binomial test; minimum 13 correct trials of 24 total). Most participants contributed all four study runs (79.07% of participants; mean, 3.76 runs; range, 2–4; 95% CI, (3.65, 3.86)) and all three localizer runs (89.41% of participants; mean, 2.85 runs; range, 1–3; 95% CI, (2.74, 2.95)). Participant information, including how many runs (out of 7 total) were contributed by each person, are displayed in Supplementary Fig. 1a. The time between sessions ranged from 0 to 76 days (mean = 15.78, median = 13.5; Supplementary Fig. 1b). For all analyses, a participant's reported age is their age in years and months (converted to decimals) on the MRI session date.

Experiment overview. Data collection and analysis were not performed blind to the conditions of the experiments. The experiment unfolded across two sessions that usually took place on separate days (Supplementary Fig. 1b). The primary

purpose of the first session was to determine whether the participant would continue to the MRI session on the basis of eligibility and interest.

During the first session, all participants (with their parents, if minors) were first exposed to the mock MRI scanner. Audio recordings of scanner noises were played over speakers while the participants lay supine in the mock scanner bore. The participants or their parents also provided information on demographics, socio-economic status and pubertal stage (Petersen Development Scale[88]; participant-completed, only for ages 8–17 yr; these measures were for exploratory purposes only and are not considered further). The participants were screened for the presence of psychiatric symptoms using the Child Behavior Checklist[89] (parent-completed) for minors or Symptom Checklist 90-Revised[90] for adults. The participants also completed the Wechsler Abbreviated Scale of Intelligence, Second Edition[91] as a measure of IQ.

In addition, the participants completed a stimulus rating task, enabling us to custom-select the faces and scenes that were familiar for each participant ('Memory task', 'Stimuli' section). Finally, the participants practiced both the memory and repeat detection tasks that would be performed in the MRI scanner on day 2. The practice tasks included different stimuli from the main experiment.

Memory task. *Stimuli.* The memory task stimuli were familiar faces and scenes (20 per category) from popular children's movies, as well as 160 common objects. The faces, scenes and objects were organized into 32 ABC triads (that is, groups of three stimuli: A, B and C) and 32 non-overlapping pairs. Of the 32 ABC triads, 16 were object–object–face ($A_OB_OC_F$), and 16 were object–object–scene ($A_OB_OC_S$), for which the face and scene stimuli were always in the C item position. All 32 non-overlapping pairs comprised two objects. Four triads and pairs were created for practice stimuli (four per condition). During study ('Pair study phase'), the triads were presented to the participants as 'overlapping' AB and BC pairs, which are related by virtue of the identical B item (always an object) in an associative inference task[4–8,12,16,62,63,92,93]. The participants later inferred the relationship between A and C. This task is similar to others used to measure integration more frequently in the developmental literature, in which even younger children learn overlapping facts[9,14,18,20,94–101] or inequalities[102–105] to derive knowledge. We chose the associative inference task because participants can learn many arbitrary pairings of stimuli, thereby affording more trials; also, the straightforward nature of the content allows us to detect the retrieval of a single, held-out item (C) during learning.

Custom selection of scenes and faces. Our goal was to quantify the degree to which reactivation of a previously associated (C) content type (face or scene) is reflected in the neural patterns engaged during AB encoding, when visual presentation is held constant (always two objects). We anticipated that reactivation might be more likely for highly familiar stimuli[106–110]; thus, we attempted to both maximize the familiarity of the C items for each person and equate it across conditions (scene versus face triads) and ages. To that end, faces and scenes were selected custom for each participant from a larger set according to their responses on a separate familiarity rating task completed during the behavioural screening session (day 1).

In the familiarity rating task, the participants were shown up to 225 images (126 faces and 99 scenes) in a random order one at a time on a computer screen. For each image, the participants indicated how familiar they were with the picture using the following options: not at all (coded as 1), a little bit (2) or very (3). The participants made their responses verbally, and the experimenter input their choice into the computer. For pictures rated as very familiar, the participants were also asked to name the character or describe the scene. The experimenter scored these responses during the task as either correct or incorrect and input their accuracy (1 or 0) into the computer. From these ratings, stimuli were selected to maximize familiarity for each person, with the additional constraint that only one image from a given movie or show could be selected for a particular participant. For example, while participants might view multiple characters and scenes from the movie *Frozen* during the stimulus rating phase, only one image from *Frozen* would appear in the final task. Familiarity ratings were automatically calculated during the task such that the task ended as soon as a participant achieved maximum familiarity for a full stimulus set (that is, 20 faces and 20 scenes each from a unique movie were all rated a 3). Thus, the majority of participants did not make familiarity ratings for all 225 stimuli in our set. The average familiarity ratings for faces and scenes selected for the memory task are shown in Supplementary Fig. 2.

Selection of objects. A single set of 160 objects was used for all participants. We made this choice for two reasons. First, from a logistical perspective, having participants rate familiarity for such a large set of objects would have made our behavioural (day 1) session prohibitively long. Second, more critically, as our goal was to measure face and scene (not object) reactivation during encoding, differences in familiarity among the object stimuli would not bias our results towards any particular outcome. Thus, in place of custom selection, we chose 160 objects that would probably be familiar to participants spanning our age range, taking into account published normative data on age of acquisition[111] for the objects' names. In particular, we reasoned that if an object name was learned early in life, a photograph of that object would probably also be familiar to a child around the same age (or younger). The objects selected for our final set had ages of acquisition ranging from 2.5 to 14.67 (mean = 5.53, median = 5.42) years. The assignment of objects to conditions was determined randomly for each participant.

Pair study phase. There were four study–test cycles that each contained a unique set of pairs. During the scanned pair study phase (Fig. 1a), the participants saw AB, BC and non-overlapping pairs on the screen (3.5 s stimulus presentation, 0.5 s inter-stimulus interval (ISI)) and were encouraged to imagine the two items interacting to aid their memory. No response was required during the pair study trials. There were a total of four triads per condition plus eight control pairs per run, yielding a total of eight AB (object–object; four related to a face, four related to a scene), eight BC (four object–face, four object–scene) and eight non-overlapping (object–object) pairs.

The pairs were blocked by type (AB_F, AB_S, BC_F, BC_S, non-overlapping$_1$ and non-overlapping$_2$, for which the non-overlapping pairs were arbitrarily split into two 'conditions' to match the triad block structure). The four pair-encoding trials within each block were jittered by interspersing them with a variable number of baseline task trials (1.5 s stimulus, 0.5 s ISI; range, 0–2; mean, 1 baseline trial between pair-encoding trials), during which the participants indicated with a button press the location at which a dot appeared on the screen (left, middle or right box). This jitter with baseline trials meant that the delay between pair-encoding trials (that is, from the offset of one pair to the onset of the next pair) ranged from 0.5 s (for zero intervening baseline task trials) to 4.5 s (for two intervening baseline task trials). The block durations were held constant at 24 s, and there was no additional interval between blocks. This mixed fMRI design enabled us to both extract single-trial estimates and analyse our data as a traditional blocked design.

Each pair was presented three times across the run. This repetition gave the participants multiple opportunities to learn each pair, thus ensuring adequate memory, and it allowed for the possibility that neural signatures of differentiation[4,58,59,66,112] or integration would evolve—or only appear—across repeated experiences[6,63]. Repetitions were distributed across thirds of the run such that every pair was presented once before being shown a second time, and twice before being shown a third. The order of pair-encoding blocks was further constrained such that (1) two blocks of the same general type (AB, BC or non-overlapping) always occurred back-to-back, with the specific order shuffled across repetitions and runs within participants; (2) BC blocks occurred last within the repetition; and (3) AB and non-overlapping blocks occurred first within a repetition equally often for each participant. This final constraint was implemented to ensure that AB and non-overlapping blocks did not differ in their average delay from BC blocks, when faces and scenes were presented, as systematicity in this regard could have influenced our comparison of AB versus non-overlapping blocks.

Test phase. After each pair study phase, the participants completed a self-paced inference and memory test for the immediately preceding pairs (Fig. 2 depicts performance). The test was not scanned. We first tested the participants on their ability to make inference judgements for all eight ABC triads prior to testing any of the direct associations. This ordering was chosen to prevent further direct pair learning during the test that might influence inference behaviour. The participants first completed a category-level, two-alternative forced-choice judgement for all triads (Fig. 1b, left), in which they were presented with the A item (always an object) and asked to indicate whether the C item indirectly related through association with a common B was a face (character) or a scene. After completing all category-level inference trials, the participants then identified the specific face or scene indirect (C) associate for every A object in a three-alternative forced-choice test (Fig. 1b, right). Again, the A item (object) served as the cue, and C items served as the options. We included the category-level inference judgement to assess whether participants could recall some information about the indirectly related item when the correct answer was not currently present. Hereafter, we consider correct inference trials to be those for which participants got both the category-level and item-level judgements correct. This strategy has the benefit of reducing the likelihood that a correctly guessed item-level inference test trial will be treated as correct.

Following the inference test, we tested the participants on their memory for the directly encoded pairs (AB, BC and non-overlapping; Fig. 1c) in the same manner as the item-level inference test. A items were cues for all AB test trials, and B items were cues for all BC test trials.

For all item-level inference and memory test trials, foils (incorrect options) were always other studied items of the same condition, position (A, B or C for overlapping pairs) and study run, to prevent the participants from using information other than the specific associative relationships to make their decisions. Note that because the foils were same-condition and same-position, the foils were always matched in stimulus type (face, object or scene) to the correct answer.

Motivational interlude. After completing each test, the participants viewed an animation of their avatar (chosen at the beginning of the experiment) climbing a staircase (Fig. 1d). The distance the character moved was proportional to the participants' accuracy on the direct pairs (AB, BC and non-overlapping) in the immediately preceding test. The staircase had three goal levels (represented by stars), and the participants were informed before beginning the experiment that they would receive a bonus payment in the amount of the highest star goal they had reached: US$5, US$10 or US$15. Our intention was to motivate the participants and keep them engaged with the task; as such, the threshold to reach

the first goal was set low enough that all participants received some amount of bonus payment. The participants needed to achieve accuracies of 30%, 59% and 88% over the course of the whole experiment to earn a US$5, US$10 or US$15 bonus, respectively. After viewing the animation, the participants continued on to complete another study phase with new pairs until they completed all four cycles.

Visual localizer (repeat detection) task. The participants also performed a separate 1-back repeat detection task with faces, scenes, objects and scrambled objects. The fMRI data acquired during this task were used to train our MVPA classifier to decode viewing of different stimulus types. The visual localizer task always took place after all four study–test cycles of the memory task were completed and thus did not interfere with the memory task data.

Stimuli. The repeat detection task stimuli were 72 familiar faces, 72 scenes, 72 intact common objects and 72 scrambled common objects. The face and scene stimuli were drawn from the same set as those used in the main memory task, but for a given participant were different from those selected for the memory experiment.

Task design. The participants viewed the stimuli on the screen one at a time for 1.5 s with a 0.5 s ISI. The participants indicated with a button press when a stimulus was identical to (that is, an exact repeat of) the immediately preceding picture. The stimuli were blocked by type (six presentations per block, for a total block duration of 12 s), and there was exactly one repeat per block. There were four blocks of each stimulus type per run, as well as five baseline blocks of the same duration. During the baseline blocks, the participants performed the same baseline task as during encoding, in which they indicated the location of a dot in an array of three boxes (1.5 s stimulus, 0.5 s ISI). The participants completed up to three runs of the visual localizer task. Behavioural responses were collected purely to ensure that the participants were paying attention to the stimuli (Supplementary Fig. 3) and are otherwise not considered in our analyses.

MR data acquisition. Imaging data were acquired on a 3.0T Siemens Skyra MRI system. Functional data were collected in 75 oblique axial slices using an EPI sequence, oriented approximately 20° off the AC–PC axis (TR, 2,000 ms; TE, 30 ms; flip angle, 73°; 128 × 128 × 75 matrix; 1.7 mm isotropic voxels; multiband acceleration factor, 3; GRAPPA factor, 2). Between one and three field maps were collected (TR, 589 ms; TE, 5 ms/7.46 ms; flip angle, 5°; matrix size, 128 × 128 × 60; 1.5 × 1.5 × 2 mm voxels) for each participant to correct for magnetic field distortions. Field maps were planned (1) before the first study run, (2) before the first visual localizer run and (3) any time a participant came out of the scanner for a break. Four participants had only one field map acquired due to technical difficulty and/or operator error. Two or three oblique coronal T2-weighted structural images were acquired perpendicular to the main axis of the HPC and in approximately the same orientation as one another (TR, 13,150 ms; TE, 82 ms; 384 × 60 × 384 matrix; 0.4 × 0.4 mm in-plane resolution; 1.5 mm through-plane resolution; 60 slices; no gap); these images were not incorporated into the analysis for the present manuscript. A T1-weighted three-dimensional MPRAGE volume (256 × 256 × 192 matrix, 1 mm isotropic voxels) was also collected for automated segmentation using Freesurfer[113] and spatial normalization to the MNI template brain using Advanced Normalization Tools (ANTS)[114].

fMRI preprocessing. The data were preprocessed and analysed using FMRI Expert Analysis Tool (FEAT) Version 6.00, part of FMRIB's Software Library (FSL) Version 5.0.9 (http://www.fmrib.ox.ac.uk/fsl), and ANTS[114]. Motion correction was applied to each functional run using MCFLIRT, and then non-brain structures were removed using BET, both part of FSL. All functional runs were then registered to the middle functional 'reference' run (in most cases, the third study run) by applying affine transformations calculated in ANTS. Anatomical images (mean coronal, MPRAGE) were then registered to the functional reference run after field-map-based unwarping of the functional data (implemented in FEAT as part of GLM analysis; see below) as follows. Each participant's MPRAGE was directly registered to their functional data using ANTS affine transformations. Non-brain structures were removed from the anatomical images using a mask derived from Freesurfer output. The result of the registration process was that all data (functional and structural, including Freesurfer parcellations) were coregistered in each participant's native functional space. All analyses were carried out in this native space except group-level GLMs.

Pre-statistics processing. In preparation for both univariate (GLM) and multivariate (MVPA) analyses, the following pre-statistics processing was applied: field-map-based EPI unwarping using PRELUDE+FUGUE, spatial smoothing using a Gaussian kernel with a full width at half maximum of 4 mm, grand-mean intensity normalization of the entire four-dimensional dataset by a single multiplicative factor, and highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with $\sigma = 50$ s). For most participants, the first field map was used to unwarp all study scans, and the second was used to unwarp all visual localizer scans. However, in many cases, participants took breaks between scans during which they were taken out of and then put back into the scanner to yield several mini scanning sessions. For these participants, the field map from the same mini-session was selected because it would most closely match the functional run in question in terms of the participant's physical positioning in the scanner. In other words, we always chose the field map that would best reflect magnetic field inhomogeneities for the particular head position in a given functional run. Separate field maps were collected for all but four participants ('MR data acquisition') for the visual localizer and study runs to ensure a similar quality of correction for both phases of the experiment, which would serve as the training and test data for the MVPA classifier, respectively.

Motion-related participant-level and run-level exclusions. Realignment parameters from MCFLIRT were used to compute framewise displacement (FD) for each fMRI volume. For each participant and run, we then defined the number of 'bad' volumes as those exceeding an FD threshold of 0.5 mm, plus one volume before and two after each high-motion volume. We used these numbers to exclude runs for which more than one-third of the total run time was corrupted by motion. As noted in the 'Participants' section, we required that participants have at least two study runs that met this criterion to be included in any analyses, and at least one visual localizer run to be included in the multivariate analyses; however, the majority of participants contributed all runs for both tasks (Supplementary Fig. 1a).

ROI definition. Content-sensitive ventral visual stream regions were defined anatomically for each participant by summing entorhinal cortex, fusiform gyrus, inferior temporal cortex and parahippocampal gyrus regions identified by Freesurfer. The resulting VTC region was used to mask functional data for MVPA.

We also defined regions in MNI template space for small-volume correction of univariate analyses. Medial PFC was delineated by hand on the 1 mm MNI template, restricted to those regions in the "medial prefrontal network" described in previous work[115]. We used the Harvard–Oxford atlas to define both IFG and HPC ROIs.

Univariate fMRI analysis. *Estimation of condition-level activation.* The task data were interrogated for regions that showed differential engagement during encoding of overlapping (AB) as compared with non-overlapping object–object pairs. The data were modelled using a GLM implemented in FEAT Version 6.00. Because we anticipated large gains in memory[47,52,53,77,78,116–120] and were interested specifically in developmental differences in the mechanisms engaged during successful overlapping versus non-overlapping encoding, we limited our analysis to those trials that were later remembered (that is, correct on the corresponding direct pair test). Individual trials (pair presentations and baseline trials) were modelled as 3.5 s events and convolved with the canonical (double-gamma) haemodynamic response function (HRF). The trials were split according to condition (AB_F, AB_S, BC_F and BC_S; non-overlapping trials were split into two groups in a parallel fashion as the overlapping pairs) and repetition (one, two or three), yielding a total of 18 regressors of interest. Subsequently, incorrect trials were collapsed into a single regressor of no interest. The baseline task was also modelled in a separate regressor. Temporal derivatives were included for all task regressors. Motion parameters calculated during the motion correction step and their temporal derivatives were added as additional confound regressors. FD and DVARS, two measures of framewise data quality, were also added to the model as regressors of no interest[93,121]. Temporal filtering was then applied to the model.

After modelling functional data within each run, we combined the resulting statistics images across study runs for each participant using fixed effects. As the data were already coregistered across runs, no additional registration or spatial normalization was necessary. Overlap-sensitive regions were defined as those that responded more on repetitions two and three (that is, after overlap had been introduced) for AB versus non-overlapping encoding (AB > non-overlapping) and vice versa (non-overlapping > AB), irrespective of the associated C item's content type (that is, collapsed across AB_F and AB_S trials). We reasoned that any region that differentially responded to these AB and non-overlapping pairs must be involved in detecting or resolving overlap, and we were thus interested in both directions of this contrast.

Contrast images for each participant were then warped to the 2 mm isotropic MNI template using ANTS and combined across participants using permutation tests (one-sample *t*-test; 1,000 iterations) implemented in FSL's randomise[122]. As we wanted our later assessment of age-related differences in sensitivity to overlap ('Assessing effects of age') to be independent of region definition, age was not incorporated into the group analysis.

The resulting group statistical maps were thresholded at a voxelwise $P < 0.005$ and submitted to cluster correction as follows. Smoothness was estimated using the residuals (warped to MNI template space) from every study run for each participant using AFNI's 3dFWHMx utility. We used the spatial AutoCorrelation Function estimation method (-acf flag), which no longer assumes a Gaussian noise distribution and generally results in a larger (more conservative) estimate of smoothness relative to prior releases of this tool, thus reducing the likelihood of a type I error[123]. We then used these run-level values to compute the average smoothness parameters across all encoding runs within participants, and then finally across participants to yield a group-level mean smoothness estimate. This entire analysis was done separately for each ROI (grey matter, HPC, IFG and medial PFC) within which cluster correction was performed. The minimum cluster extents at a significance threshold of $P < 0.05$ were determined for each ROI using

3dClustSim. The minimum cluster sizes were determined to be 10 voxels for HPC, 17 voxels for IFG, 27 voxels for medial PFC and 71 voxels for grey matter. All clusters exceeding these criteria within the three a priori anatomical regions[2,4–8,12,93] and/or at the whole-brain grey matter level are reported here.

Assessing effects of age. All overlap-sensitive regions were identified for showing a main effect of overlapping versus non-overlapping pairs, irrespective of age. To determine whether there were in fact effects of age present within these clusters, we extracted contrast estimates (COPEs) for each participant and condition (AB and non-overlapping). We used linear models ('Statistical analyses') to assess whether the activation difference observed between encoding of overlapping and non-overlapping pairs was modulated by age. Note that because the functional regions were identified for showing either AB > non-overlapping or the reverse, the effects of condition are trivial; we were specifically interested in whether there were significant effects of age and/or interactions between age and condition. We used a model comparison approach to ask whether age explained any additional variability in activation beyond condition. Predictions from the best-fitting model and statistics for significant regions are provided in Supplementary Fig. 5 and Supplementary Table 1, respectively.

Estimation of trial-level neural patterns. In addition to condition-level univariate analyses, we extracted neural patterns for individual trials. These patterns were used as inputs to our trialwise classification analysis (see below). Trial-level neural patterns were generated under the assumptions of the GLM using a modified LS-S approach[124]. Statistics images associated with each encoding trial were estimated for each repetition and participant using custom Python routines. Pair presentations were modelled as 3.5 s events and convolved with the canonical (double gamma) HRF. Motion parameters calculated during the motion correction step and their temporal derivatives were added as additional confound regressors. As in the other univariate models, FD and DVARS were also added to the model as regressors of no interest[94,122]. Temporal filtering was applied to the model. This process resulted in one statistic image for each of the 32 AB pairs, 32 BC pairs and 32 non-overlapping pairs for each of three repetitions, for a total of up to 288 images per participant (those participants contributing fewer runs had correspondingly fewer images).

Multivariate fMRI classification analysis. *Blockwise reactivation analysis.* Our main classification analysis was carried out on the preprocessed time-series data in the native space of each participant. We first asked whether our classifier could discriminate between face and scene viewing during visual stimulus presentation (visual localizer task). We then applied our trained classifier to assess reactivation of related face or scene memories (C item content type) during study of overlapping (AB) object–object pairs.

Decoding visually presented content. We assessed whether the classifier could discriminate between the viewing of faces and scenes on the basis of VTC activation patterns from the visual localizer (repeat detection) task using a within-participant cross-validation approach. Specifically, we trained a pattern classifier (sparse multinomial logistic regression implemented in PyMVPA; $\lambda = 0.1$, the package default) to differentiate face from scene viewing on the basis of activation patterns from a subset of localizer runs (Fig. 3a). We trained on all six volumes (12 s) of data in each face and scene block, with volume labels shifted by 6 s to account for haemodynamic lag. The classifier was then tested on patterns from the held-out run (one 'fold'). This approach was repeated until all runs had been held out once. Cross-validation was performed on detrended and z-scored data within anatomically defined VTC; no further feature selection was performed. Accuracy was computed by comparing the classifier-predicted to actual stimulus type (face or scene) for each fMRI volume; an average accuracy was then calculated across all volumes and all folds to yield a single decoding accuracy score per participant (Fig. 3b). Five participants (four children ages 8–9 and one adult age 19) were excluded from this cross-validation analysis because they contributed fewer than two fMRI runs of the localizer task, making it impossible to perform cross-validation across runs on these participants (total $N = 81$).

Decoding internally generated content (reactivation). Our next analysis was designed to determine whether there are developmental differences in the tendency to reactivate related memories during encoding. To assess this, we trained our classifier (sparse multinomial logistic regression, $\lambda = 0.1$ as above) on all localizer task runs that met our inclusion criteria for each participant. The classifier was then applied to all study task volumes (Fig. 3c). As above, this analysis was performed on detrended and z-scored data within VTC, with no other feature selection applied. We then computed a 'reactivation index' over the classifier evidence (probabilities) that summarized, for each participant, the degree to which their neural patterns reflected reinstatement of the related more than the unrelated type of content. The reactivation index was defined as face minus scene evidence for AB_F trials plus scene minus face evidence for AB_S trials (that is, the interaction term). Reactivation indices above zero thus indicated that classifier evidence was dependent on trial type. As control analyses, we also computed the same score for BC trials ('perception index') and non-overlapping trials ('control index'), which should yield decoding of the perceived BC stimulus type and chance-level decoding of nonsense non-overlapping input, respectively. Two participants (eight- to nine-year-old children) were excluded from this analysis. One child had no localizer task data, and therefore this within-participant analysis was impossible; the other was an outlier in cross-validation accuracy for decoding of perception in VTC and was thus excluded (Fig. 3b, open circle; total $N = 84$). Note that for this analysis only, we decided to exclude this outlier participant, for whom classification performance was not reliably above chance. The reason for this exclusion was thus not so much that this person was an outlier per se, but rather that their results of applying a classifier that cannot discriminate among conditions in the training set will be uninterpretable when applied to a separate task (here, pair encoding). This group of 84 participants were included for all subsequent reactivation-related analyses that follow.

Trialwise reactivation analysis. Having established developmental differences in reactivation at the block level, we next quantified reactivation on a trial-by-trial basis. This analysis was performed to ask whether there are developmental differences in the behavioural and neural consequences of reactivation within participants. We used the same classifier trained to discriminate face from scene viewing as we did for the blockwise decoding analysis. However, instead of applying the classifier to each volume in the study runs, we applied it to the trial-level neural patterns (statistics images that reflected each repetition of each pair). This yielded an estimate of the degree to which each specific pair reflected reinstatement of its related type of content for each repetition. We then computed the log odds of the classifier output corresponding to the condition of interest as our trialwise measure of reactivation. For example, $\log[\text{prob}_F/(1-\text{prob}_F)]$ would reflect face evidence for face-related trials. This transformation has been used in previous work[125] to correct for non-normality in the raw classifier output, which we also observed here. As our goal was to assess developmental differences in the neural mechanisms involved during successful memory formation rather than differences in memory ability per se, we restricted our analyses to correctly remembered trials only.

Relating initial reactivation to subsequent engagement. To assess whether initial reactivation modulates subsequent engagement, we asked whether the trialwise reactivation measures described above ('Trialwise reactivation analysis') from repetition two were related to activation on repetition three. Trialwise reactivation scores were mean-centred within participants and included as a trial-by-trial parametric modulator for all repetition-three AB trials. These GLMs were otherwise identical to the main models, except the addition of this parametric regressor. As in the previous analysis, only trials for which the corresponding direct memory test was correct were included. Note that because our reactivation measures were derived from repetition two to assess activation differences during the subsequent repetition three, the measures are coming from different time points and represent independent data, and the only relationship is through the pair (content) itself.

Statistics images were then combined across runs within participants as above for the main models. At the group level, we were interested in effects that were consistent across the group as well as those that varied with age. We thus ran one main effects model disregarding age (mirroring our general approach in the main analyses) and a separate model that included mean-centred age as a parametric regressor. Both analyses were run using FSL's randomise, as above. Cluster correction was performed using the same method as for the main models. Within identified clusters, we extracted the participant-level contrast estimates (COPEs in FEAT) associated with the parametric regressor for visualization of the effects (Fig. 5).

Statistical analyses. *Model specifications.* As this study is a cross-sectional developmental study, age was an across-subjects factor; all other measures were repeated within subjects. Statistical analyses (except those carried out in FSL) were performed using R[126]. We primarily used (generalized) linear mixed effects models implemented in the lme4 package[127] to model individual trials, except when we had only one observation per participant (in which case we used linear models; stats::lm). For models assessing within-participant relationships (and optionally interactions with age), the predictors were scaled and centred within participants to remove subject-specific effects. Factors were effect coded to allow for the interpretation of lower-order terms as main effects in the presence of interactions. Repetition was typically treated as a factor so as not to require consistently increasing or decreasing reactivation across repetitions; one exception to this was for the analyses of motion (Supplementary Results, 'Effect of increased motion over repetitions is not significantly modulated by age' section), in which we did expect consistent increases across repetitions. In addition, because the developmental trajectories in question are potentially nonlinear[31,40], we opted to model age with a basis spline function. This approach uses a linear combination of basis functions, thereby allowing us to remain agnostic as to the particular shape of the relationship[128]. Basis splines have the advantage of being fit locally (that is, separately at specific parts of the age range). Such local fitting means that basis splines are less affected by values at either extreme end of our age range (that is, the youngest and oldest participants in our sample) compared with polynomials, which are fit globally[129]. In all analyses, participants were treated as random effects.

Model comparison and statistical reporting. For all analyses, we took a model comparison approach in which a model including age was compared with a base model in which age was not considered. The R package stats::anova was used to perform model comparison, and the age model was said to significantly improve on the base model at a threshold of $P < 0.05$ (two-tailed; uncorrected). We then report the statistics for the better-fitting model, either the base (in cases where adding age to the model did not significantly improve the fit) or the one additionally incorporating age (in cases where adding age did significantly improve the fit). We assessed statistical significance of each of our fixed effects including interaction terms in the better-fitting model using a Wald chi-squared test (type III SS for models including interaction terms, type II otherwise) for linear mixed effects models (lme4::lmer and lme4::glmer) and *F*-test for linear models (stats::lm; all Wald chi-squared tests were implemented in R using car::Anova)[130]. We used ggeffects::ggemmeans[131] in R to visualize the predicted responses and compute CIs at various ages (all data figures).

Outlier exclusions. We did not incorporate outlier exclusion into our primary analysis, with the single exception of removing one participant for whom we could not decode perceived stimulus type because this precluded the application of the trained classifier to the main memory task (Fig. 3b, open circle). However, to ensure that our other findings were not disproportionately influenced by outliers, we verified that the results were similar after removing statistical outliers, or data points with a standardized residual greater than 2.5 (in R, LMERConvenienceFunctions::romr.fnc with the default settings; the results excluding outliers are reported throughout the main text). For the analysis shown in Fig. 4c, all trials associated with the repetitions that were identified as outliers from the corresponding memory reactivation analysis were excluded (linear mixed effects model shown in Fig. 4a,b; a total of five repetitions were excluded, one each from five participants), as identifying outlier observations with a binomial linking function is not straightforward.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available on the Open Science Framework (https://osf.io/hg6wf/)[132].

Code availability

The custom code that supports the findings of this study is available on the Open Science Framework (https://osf.io/hg6wf/)[132].

Received: 21 July 2020; Accepted: 1 September 2021;
Published online: 15 November 2021

References

1. Preston, A. R. & Eichenbaum, H. Interplay of hippocampus and prefrontal cortex in memory. *Curr. Biol.* **23**, R764–R773 (2013).
2. Schlichting, M. L. & Preston, A. R. Memory integration: neural mechanisms and implications for behavior. *Curr. Opin. Behav. Sci.* **1**, 1–8 (2015).
3. Shohamy, D. & Wagner, A. D. Integrating memories in the human brain: hippocampal–midbrain encoding of overlapping events. *Neuron* **60**, 378–389 (2008).
4. Schlichting, M. L., Mumford, J. A. & Preston, A. R. Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nat. Commun.* **6**, 8151 (2015).
5. Schlichting, M. L. & Preston, A. R. Hippocampal–medial prefrontal circuit supports memory updating during learning and post-encoding rest. *Neurobiol. Learn. Mem.* **134**, 91–106 (2016).
6. Zeithamova, D., Dominick, A. L. & Preston, A. R. Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron* **75**, 168–179 (2012).
7. Spalding, K. N. et al. Ventromedial prefrontal cortex is necessary for normal associative inference and memory integration. *J. Neurosci.* **38**, 2501–2517 (2018).
8. Zeithamova, D. & Preston, A. R. Flexible memories: differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. *J. Neurosci.* **30**, 14676–14684 (2010).
9. Varga, N. L. & Bauer, P. J. Using event-related potentials to inform the neurocognitive processes underlying knowledge extension through memory integration. *J. Cogn. Neurosci.* **29**, 1932–1949 (2017).
10. Kuhl, B. A., Shah, A. T., DuBrow, S. & Wagner, A. D. Resistance to forgetting associated with hippocampus-mediated reactivation during new learning. *Nat. Neurosci.* **13**, 501–506 (2010).
11. Banino, A., Koster, R., Hassabis, D. & Kumaran, D. Retrieval-based model accounts for striking profile of episodic memory and generalization. *Sci. Rep.* **6**, 31330 (2016).
12. Zeithamova, D. & Preston, A. R. Temporal proximity promotes integration of overlapping events. *J. Cogn. Neurosci.* **29**, 1311–1323 (2017).
13. Kumaran, D. & McClelland, J. L. Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychol. Rev.* **119**, 573–616 (2012).
14. Bauer, P. J. & San Souci, P. Going beyond the facts: young children extend knowledge by integrating episodes. *J. Exp. Child Psychol.* **107**, 452–465 (2010).
15. Bauer, P. J., Cronin-Golomb, L. M., Porter, B. M., Jaganjac, A. & Miller, H. E. Integration of memory content in adults and children: developmental differences in task conditions and functional consequences. *J. Exp. Psychol. Gen.* https://doi.org/10.1037/xge0000996 (2020).
16. Schlichting, M. L., Guarino, K. F., Schapiro, A. C., Turk-Browne, N. B. & Preston, A. R. Hippocampal structure predicts statistical learning and associative inference abilities during development. *J. Cogn. Neurosci.* **29**, 37–51 (2017).
17. Shing, Y. L. et al. Integrating across memory episodes: developmental trends. *PLoS ONE* **14**, e0215848 (2019).
18. Bauer, P. J., Varga, N. L., King, J. E., Nolen, A. M. & White, E. A. Semantic elaboration through integration: hints both facilitate and inform the process. *J. Cogn. Dev.* **16**, 351–369 (2015).
19. Krumm, S., Ziegler, M. & Buehner, M. Reasoning and working memory as predictors of school grades. *Learn. Individ. Differ.* **18**, 248–257 (2008).
20. Varga, N. L., Esposito, A. G. & Bauer, P. J. Cognitive correlates of memory integration across development: explaining variability in an educationally relevant phenomenon. *J. Exp. Psychol. Gen.* **148**, 739–762 (2019).
21. Marr, D. Simple memory: a theory for archicortex. *Phil. Trans. R. Soc. Lond.* **262**, 23–81 (1971).
22. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
23. Badre, D. & Wagner, A. D. Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia* **45**, 2883–2901 (2007).
24. Kuhl, B. A., Johnson, M. K. & Chun, M. M. Dissociable neural mechanisms for goal-directed versus incidental memory reactivation. *J. Neurosci.* **33**, 16099–16109 (2013).
25. Paz-Alonso, P. M., Ghetti, S., Matlen, B. J., Anderson, M. C. & Bunge, S. A. Memory suppression is an active process that improves over childhood. *Front. Hum. Neurosci.* **3**, 24 (2009).
26. Gogtay, N. et al. Dynamic mapping of normal human hippocampal development. *Hippocampus* **16**, 664–672 (2006).
27. Langnes, E. et al. Anterior and posterior hippocampus macro- and microstructure across the lifespan in relation to memory—a longitudinal study. *Hippocampus* https://doi.org/10.1002/hipo.23189 (2020).
28. Demaster, D. M., Pathman, T., Lee, J. K. & Ghetti, S. Structural development of the hippocampus and episodic memory: developmental differences along the anterior/posterior axis. *Cereb. Cortex* **24**, 3036–3045 (2013).
29. Paz-Alonso, P. M., Ghetti, S., Donohue, S. E., Goodman, G. S. & Bunge, S. A. Neurodevelopmental correlates of true and false recognition. *Cereb. Cortex* **18**, 2208–2216 (2008).
30. Maril, A. et al. Developmental fMRI study of episodic verbal memory encoding in children. *Neurology* **75**, 2110–2116 (2010).
31. Calabro, F. J., Murty, V. P., Jalbrzikowski, M., Tervo-Clemmens, B. & Luna, B. Development of hippocampal–prefrontal cortex interactions through adolescence. *Cereb. Cortex* **30**, 1548–1558 (2019).
32. Menon, V., Boyett-Anderson, J. M. & Reiss, A. L. Maturation of medial temporal lobe response and connectivity during memory encoding. *Cogn. Brain Res.* **25**, 379–385 (2005).
33. Sowell, E. R. et al. Mapping cortical change across the human life span. *Nat. Neurosci.* **6**, 309–315 (2003).
34. Ackerman, B. P. Retrieval variability: the inefficient use of retrieval cues by young children. *J. Exp. Child Psychol.* **33**, 413–428 (1982).
35. Ackerman, B. P. Children's use of context and category cues to retrieve episodic information from memory. *J. Exp. Child Psychol.* **40**, 420–438 (1985).
36. DeMaster, D., Coughlin, C. & Ghetti, S. Retrieval flexibility and reinstatement in the developing hippocampus. *Hippocampus* https://doi.org/10.1002/hipo.22538 (2015).
37. Uddin, L. Q., Supekar, K. S., Ryali, S. & Menon, V. Dynamic reconfiguration of structural and functional connectivity across core neurocognitive brain networks with development. *J. Neurosci.* **31**, 18578–18589 (2011).
38. Qin, S. et al. Hippocampal–neocortical functional reorganization underlies children's cognitive development. *Nat. Neurosci.* **17**, 1263–1269 (2014).
39. Larsen, B. & Luna, B. Adolescence as a neurobiological critical period for the development of higher-order cognition. *Neurosci. Biobehav. Rev.* **94**, 179–195 (2018).
40. Murty, V., Calabro, F. & Luna, B. The role of experience in adolescent cognitive development: integration of executive, memory, and mesolimbic systems. *Neurosci. Biobehav. Rev.* https://doi.org/10.1016/j.neubiorev.2016.07.034 (2016).

41. Ghetti, S., DeMaster, D. M., Yonelinas, A. P. & Bunge, S. A. Developmental differences in medial temporal lobe function during memory encoding. *J. Neurosci.* **30**, 9548–9556 (2010).
42. Pattwell, S. S., Bath, K. G., Casey, B. J., Ninan, I. & Lee, F. S. Selective early-acquired fear memories undergo temporary suppression during adolescence. *Proc. Natl Acad. Sci. USA* **108**, 1182–1187 (2011).
43. Pattwell, S. S. et al. Dynamic changes in neural circuitry during adolescence are associated with persistent attenuation of fear memories. *Nat. Commun.* **7**, 11475 (2016).
44. Holliday, R. E. & Weekes, B. S. Dissociated developmental trajectories for semantic and phonological false memories. *Memory* **14**, 624–636 (2006).
45. Brainerd, C. J. & Reyna, V. F. Explaining developmental reversals in false memory: research article. *Psychol. Sci.* **18**, 442–448 (2007).
46. Willoughby, K. A., Desrocher, M., Levine, B. & Rovet, J. F. Episodic and semantic autobiographical memory and everyday memory during late childhood and early adolescence. *Front. Psychol.* **3**, 53 (2012).
47. Billingsley, R. L., Smith, M., Lou & McAndrews, M. P. Developmental patterns in priming and familiarity in explicit recollection. *J. Exp. Child Psychol.* **82**, 251–277 (2002).
48. Daugherty, A. M., Flinn, R. & Ofen, N. Hippocampal CA3-dentate gyrus volume uniquely linked to improvement in associative memory from childhood to adulthood. *Neuroimage* **153**, 75–85 (2017).
49. Lee, J. K., Ekstrom, A. D. & Ghetti, S. Volume of hippocampal subfields and episodic memory in childhood and adolescence. *Neuroimage* **94**, 162–171 (2014).
50. Lee, J. K. et al. Changes in anterior and posterior hippocampus differentially predict item–space, item–time, and item–item memory improvement. *Dev. Cogn. Neurosci.* **41**, 100741 (2020).
51. Tamnes, C. K. et al. Regional hippocampal volumes and development predict learning and memory. *Dev. Neurosci.* **36**, 161–174 (2014).
52. Keresztes, A. et al. Hippocampal maturity promotes memory distinctiveness in childhood and adolescence. *Proc. Natl Acad. Sci. USA* **114**, 201710654 (2017).
53. Demaster, D. M. & Ghetti, S. Developmental differences in hippocampal and cortical contributions to episodic retrieval. *Cortex* **49**, 1482–1493 (2013).
54. Brunec, I. K. et al. Multiple scales of representation along the hippocampal anteroposterior axis in humans. *Curr. Biol.* **28**, 2129–2135.e6 (2018).
55. Collin, S. H. P., Milivojevic, B. & Doeller, C. F. Memory hierarchies map onto the hippocampal long axis in humans. *Nat. Neurosci.* **18**, 1562–1564 (2015).
56. Bowman, C. R. & Zeithamova, D. Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *J. Neurosci.* **38**, 2605–2614 (2018).
57. Callaghan, B. et al. Age-related increases in posterior hippocampal granularity are associated with remote detailed episodic memory in development. *J. Neurosci.* **41**, 1738–1754 (2020).
58. Duncan, K. D. & Schlichting, M. L. Hippocampal representations as a function of time, subregion, and brain state. *Neurobiol. Learn. Mem.* **153**, 40–56 (2018).
59. Hulbert, J. C. & Norman, K. A. Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice. *Cereb. Cortex* https://doi.org/10.1093/cercor/bhu284 (2014).
60. Casey, B. J. Beyond simple models of self-control to circuit-based accounts of adolescent behavior. *Annu. Rev. Psychol.* https://doi.org/10.1146/annurev-psych-010814-015156 (2014).
61. Siegler, R. S. *Emerging Minds: The Process of Change in Children's Thinking.* https://doi.org/10.1093/oso/9780195077872.003.0009 (Oxford Univ. Press, 1996).
62. Preston, A. R., Shrager, Y., Dudukovic, N. & Gabrieli, J. D. E. Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus* **14**, 148–152 (2004).
63. Zeithamova, D., Manthuruthil, C. & Preston, A. R. Repetition suppression in the medial temporal lobe and midbrain is altered by event overlap. *Hippocampus* **26**, 1464–1477 (2016).
64. Norman, K. A., Polyn, S. M., Detre, G. J. & Haxby, J. V. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* **10**, 424–430 (2006).
65. Yassa, M. A. & Stark, C. E. L. Pattern separation in the hippocampus. *Trends Neurosci.* **34**, 515–525 (2011).
66. Chanales, A. J. H., Oza, A., Favila, S. E. & Kuhl, B. A. Overlap among spatial memories triggers repulsion of hippocampal representations. *Curr. Biol.* **27**, 2307–2317.e5 (2017).
67. Molitor, R. J., Sherrill, K. R., Morton, N. W., Miller, A. A. & Preston, A. R. Memory reactivation during learning simultaneously promotes dentate gyrus/CA2,3 pattern differentiation and CA1 memory integration. *J. Neurosci.* **41**, 726–738 (2021).
68. Favila, S. E., Samide, R., Sweigart, S. C. & Kuhl, B. A. Parietal representations of stimulus features are amplified during memory retrieval and flexibly aligned with top-down goals. *J. Neurosci.* **38**, 7809–7821 (2018).
69. Sastre, M., Wendelken, C., Lee, J. K., Bunge, S. A. & Ghetti, S. Age- and performance-related differences in hippocampal contributions to episodic retrieval. *Dev. Cogn. Neurosci.* **19**, 42–50 (2016).
70. Lindberg, M. A. Is knowledge base development a necessary and sufficient condition for memory development? *J. Exp. Child Psychol.* **30**, 401–410 (1980).
71. Schneider, W., Gruber, H., Gold, A. & Opwis, K. Chess expertise and memory for chess position in children and adults. *J. Exp. Child Psychol.* **56**, 328–349 (1993).
72. Riggins, T. et al. Protracted hippocampal development is associated with age-related improvements in memory during early childhood. *Neuroimage* **174**, 127–137 (2018).
73. Ghetti, S. & Fandakova, Y. Neural development of memory and metamemory in childhood and adolescence: toward an integrative model of the development of episodic recollection. *Annu. Rev. Dev. Psychol.* **2**, 365–388 (2020).
74. Goldsberry, M. E., Kim, J. & Freeman, J. H. Developmental changes in hippocampal associative coding. *J. Neurosci.* **35**, 4238–4247 (2015).
75. Giovanello, K. S., Schnyer, D. M. & Verfaellie, M. Distinct hippocampal regions make unique contributions to relational memory. *Hippocampus* **19**, 111–117 (2009).
76. Davachi, L. Item, context and relational episodic encoding in humans. *Curr. Opin. Neurobiol.* **16**, 693–700 (2006).
77. Brod, G., Werkle-Bergner, M. & Shing, Y. L. The influence of prior knowledge on memory: a developmental cognitive neuroscience perspective. *Front. Behav. Neurosci.* **7**, 139 (2013).
78. Shing, Y. L. & Brod, G. Effects of prior knowledge on memory: implications for education. *Mind Brain Educ.* **10**, 153–161 (2016).
79. Sneider, J. T. et al. Adolescent hippocampal and prefrontal brain activation during performance of the virtual Morris water task. *Front. Hum. Neurosci.* **12**, 238 (2018).
80. Ofen, N. The development of neural correlates for memory formation. *Neurosci. Biobehav. Rev.* **36**, 1708–1717 (2012).
81. Shing, Y. L. et al. Episodic memory across the lifespan: the contributions of associative and strategic components. *Neurosci. Biobehav. Rev.* **34**, 1080–1091 (2010).
82. Tang, L., Shafer, A. T. & Ofen, N. Prefrontal cortex contributions to the development of memory formation. *Cereb. Cortex* https://doi.org/10.1093/cercor/bhx200 (2017).
83. Kim, G., Lewis-Peacock, J. A., Norman, K. A. & Turk-Browne, N. B. Pruning of memories by context-based prediction error. *Proc. Natl Acad. Sci. USA* **111**, 8997–9002 (2014).
84. Schapiro, A. C., Kustner, L. V. & Turk-Browne, N. B. Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Curr. Biol.* **22**, 1622–1627 (2012).
85. Carpenter, A. C. & Schacter, D. L. Flexible retrieval: when true inferences produce false memories. *J. Exp. Psychol. Learn. Mem. Cogn.* **43**, 335–349 (2017).
86. Whitaker, K. J., Vendetti, M. S., Wendelken, C. & Bunge, S. A. Neuroscientific insights into the development of analogical reasoning. *Dev. Sci.* **21**, e12531 (2018).
87. Wendelken, C., Ferrer, E., Whitaker, K. J. & Bunge, S. A. Fronto-parietal network reconfiguration supports the development of reasoning ability. *Cereb. Cortex* https://doi.org/10.1093/cercor/bhv050 (2015).
88. Peterson, a, Crockett, L., Richards, M. & Boxer, A. A self-report measure of pubertal status. *J. Youth Adolesc.* **17**, 117–133 (1988).
89. Achenbach, T. M. *Manual for the Child Behavior Checklist/4-18 and 1991 Profile* (Department of Psychiatry, Univ. of Vermont, 1991).
90. Derogatis, L. R. *SCL-90-R: Administration, Scoring and Procedures—Manual 1* (Clinical Psychometric Research, 1977).
91. Wechsler, D. *Wechsler Abbreviated Scale of Intelligence* (Psychological Corporation, 1999).
92. Schlichting, M. L., Zeithamova, D. & Preston, A. R. CA1 subfield contributions to memory integration and inference. *Hippocampus* **24**, 1248–1260 (2014).
93. Schlichting, M. L. & Preston, A. R. Memory reactivation during rest supports upcoming learning of related content. *Proc. Natl Acad. Sci. USA* **111**, 15845–15850 (2014).
94. Bauer, P. J., Dugan, J. A., Varga, N. L. & Riggins, T. Relations between neural structures and children's self-derivation of new knowledge through memory integration. *Dev. Cogn. Neurosci.* **36**, 100611 (2019).
95. Bauer, P. J. & Larkina, M. Realizing relevance: the influence of domain-specific information on generation of new knowledge through integration in 4- to 8-year-old children. *Child Dev.* **88**, 247–262 (2017).
96. Bauer, P. J., King, J. E., Larkina, M., Varga, N. L. & White, E. A. Characters and clues: factors affecting children's extension of knowledge through integration of separate episodes. *J. Exp. Child Psychol.* **111**, 681–694 (2012).
97. Varga, N. L., Stewart, R. A. & Bauer, P. J. Integrating across episodes: investigating the long-term accessibility of self-derived knowledge in 4-year-old children. *J. Exp. Child Psychol.* **145**, 48–63 (2016).

98. Bauer, P. J. & Jackson, F. L. Semantic elaboration: ERPs reveal rapid transition from novel to known. *J. Exp. Psychol. Learn. Mem. Cogn.* **41**, 271–282 (2016).
99. Esposito, A. G. & Bauer, P. J. Building a knowledge base: predicting self-derivation through integration in 6- to 10-year-olds. *J. Exp. Child Psychol.* **176**, 55–72 (2018).
100. Varga, N. L. & Bauer, P. J. Effects of delays on 6-year-old children's self-generation and retention of knowledge through integration. *J. Exp. Child Psychol.* **115**, 326–341 (2013).
101. Esposito, A. G. & Bauer, P. J. Going beyond the lesson: self-generating new factual knowledge in the classroom. *J. Exp. Child Psychol.* **153**, 110–125 (2017).
102. Bryant, P. & Trabasso, T. Transitive inferences and memory in young children. *Nature* **232**, 456–458 (1971).
103. Brainerd, C. & Kingma, J. Do children have to remember to reason? A fuzzy-trace theory of transitivity development. *Dev. Rev.* **4**, 311–377 (1984).
104. Chapman, M. & Lindenberger, U. Transitivity judgments, memory for premises, and models of children's reasoning. *Dev. Rev.* **12**, 124–163 (1992).
105. Reyna, V. F. & Brainerd, C. J. Fuzzy processing in transitivity development. *Ann. Oper. Res.* **23**, 37–63 (1990).
106. Staresina, B. P., Gray, J. C. & Davachi, L. Event congruency enhances episodic memory encoding through semantic elaboration and relational binding. *Cereb. Cortex* **19**, 1198–1207 (2009).
107. Liu, Z. X., Grady, C. & Moscovitch, M. Effects of prior-knowledge on brain activation and connectivity during associative memory encoding. *Cereb. Cortex* **27**, 1991–2009 (2017).
108. Reggev, N., Bein, O. & Maril, A. Distinct neural suppression and encoding effects for conceptual novelty and familiarity. *J. Cogn. Neurosci.* **28**, 1455–1470 (2016).
109. Maril, A. et al. Event congruency and episodic encoding: a developmental fMRI study. *Neuropsychologia* **49**, 3036–3045 (2011).
110. van Kesteren, M. T. R., Rijpkema, M., Ruiter, D. J. & Fernandez, G. Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. *J. Neurosci.* **30**, 15888–15894 (2010).
111. Kuperman, V., Stadthagen-Gonzalez, H. & Brysbaert, M. Age-of-acquisition ratings for 30,000 English words. *Behav. Res. Methods* https://doi.org/10.3758/s13428-012-0210-4 (2012).
112. Favila, S. E., Chanales, A. J. H. & Kuhl, B. A. Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nat. Commun.* **6**, 11066 (2016).
113. Desikan, R. S. et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980 (2006).
114. Avants, B. B. et al. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* **54**, 2033–2044 (2011).
115. Price, J. L. & Drevets, W. C. Neurocircuitry of mood disorders. *Neuropsychopharmacology* **35**, 192–216 (2009).
116. Ghetti, S. & Bunge, S. A. Neural changes underlying the development of episodic memory during middle childhood. *Dev. Cogn. Neurosci.* **2**, 381–395 (2012).
117. Townsend, E. L., Richmond, J. L., Vogel-Farley, V. K. & Thomas, K. Medial temporal lobe memory in childhood: developmental transitions. *Dev. Sci.* **13**, 738–751 (2010).
118. Keresztes, A., Ngo, C. T., Lindenberger, U., Werkle-Bergner, M. & Newcombe, N. S. Hippocampal maturation drives memory from generalization to specificity. *Trends Cogn. Sci.* **22**, 676–686 (2018).
119. Ghetti, S. & Angelini, L. The development of recollection and familiarity in childhood and adolescence: evidence from the dual-process signal detection model. *Child Dev.* **79**, 339–358 (2008).
120. Ofen, N. et al. Development of the declarative memory system in the human brain. *Nat. Neurosci.* **10**, 1198–1205 (2007).
121. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* **59**, 2142–2154 (2012).
122. Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. & Nichols, T. E. Permutation inference for the general linear model. *Neuroimage* **92**, 381–397 (2014).
123. Cox, R. W., Chen, G., Glen, D. R., Reynolds, R. C. & Taylor, P. A. FMRI clustering in AFNI: false-positive rates redux. *Brain Connect.* **7**, 152–171 (2017).
124. Mumford, J. A., Turner, B. O., Ashby, F. G. & Poldrack, R. A. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage* **59**, 2636–2643 (2012).
125. Richter, F. R., Chanales, A. J. H. & Kuhl, B. A. Predicting the integration of overlapping memories by decoding mnemonic processing states during learning. *Neuroimage* https://doi.org/10.1016/j.neuroimage.2015.08.051 (2015).
126. R Core Team R: A language and environment for statistical computing (R Foundation for Statistical Computing, 2018).
127. Bates, D., Mächler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
128. Perperoglou, A., Sauerbrei, W., Abrahamowicz, M. & Schmid, M. A review of spline function procedures in R. *BMC Med. Res. Methodol.* **19**, 46 (2019).
129. Francis, B., Elliott, A. & Weldon, M. Smoothing group-based trajectory models through B-splines. *J. Dev. Life Course Criminol.* **2**, 113–133 (2016).
130. Fox, J. & Weisberg, S. *An R Companion to Applied Regression* (Sage, 2011).
131. Lüdecke, D. ggeffects: tidy data frames of marginal effects from regression models. *J. Open Source Softw.* **3**, 772 (2018).
132. Schlichting, M. L., Guarino, K. F., Roome, H. E. & Preston, A. R. Memory reactivation modulates new encoding and impacts inference in the developing human brain. *Open Science Framework* https://doi.org/10.17605/OSF.IO/HG6WF (2021).
133. Brodeur, M. B., Dionne-Dostie, E., Montreuil, T. & Lepage, M. The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS ONE* **5**, e10773 (2010).
134. Brodeur, M. B., Guérard, K. & Bouras, M. Bank of Standardized Stimuli (BOSS) phase II: 930 new normative photos. *PLoS ONE* **9**, e106953 (2014).

Acknowledgements

We thank T. Tran for help with stimulus development and participant recruitment, and S. Ventura, N.-H. Hue and K. Nguyen for assistance with data collection and analysis. We also thank M. Mack, D. Zeithamova, N. Varga and K. Duncan for input on statistical analyses and helpful discussions. This work was supported by the National Institutes of Health under award numbers R01 MH100121 and R21 HD083785 (A.R.P.) and by the Canada Foundation for Innovation John R. Evans Leaders Fund (grant no. 36876; M.L.S.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

M.L.S. and A.R.P. conceptualized the study. M.L.S., K.F.G. and H.E.R. collected the data. M.L.S. and K.F.G. analysed the data. M.L.S. drafted the paper, and all authors were involved in revising and finalizing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-021-01206-5.

Correspondence and requests for materials should be addressed to Margaret L. Schlichting or Alison R. Preston.

Peer review information *Nature Human Behaviour* thanks Garvin Brod, Shaozheng Qin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature research

Corresponding author(s): Margaret Schlichting, Alison Preston

Last updated by author(s): July 21, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	MATLAB R2016a
Data analysis	FSL Version 5.0.9; MATLAB R2016a; Python Version 2.7.15; PyMVPA Version 2.6; Freesurfer Version 5.3; R Version 3.5.2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study are available on the Open Science Framework website (https://osf.io/hg6wf/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative cross-sectional experimental study of neurocognitive development.
Research sample	Children, adolescents, and adults residing in the Austin, TX area (final sample characteristics, following exclusions: ages 7.16-29.42 years; N=65 minors under age 18 including 35 females; N=21 adults 18+ including 11 females). All participants were right-handed, had normal or corrected-to-normal vision, and were free of diagnosed or suspected learning disabilities or psychiatric conditions.
Sampling strategy	Our sampling procedure yielded N=21 in each of four smaller age bands, with efforts made to have an equal number of males and females. Power analyses using data from a similar task (Cohen's d = 0.56) showed that a sample size of 21 participants would yield 80% power to detect the behavioral effect of a within-participant manipulation of integration at the group level; as such, we recruited participants in each of these age bands until we had a minimum of 21 per band that could be included in our primary reactivation analysis. Our sample size also aligns with prior developmental work on a similar topic (Schlichting et al. 2017; Shing et al. 2019).
Data collection	MRI data was collected at the University of Texas at Austin Imaging Research Center using a Siemens Skyra 3T MRI scanner. Stimulus timing and behavioral data collection was computer based. Two to three researchers, neither of whom were blind to the study hypotheses, were present during MRI data collection. Occasionally (upon request) parents of minor participants were present in the MRI control room during data collection.
Timing	May 2015-March 2017
Data exclusions	One hundred and twenty-five volunteers ranging in age from 6-30 years (actual range = 6.41-29.33) participated in a behavioral screening session prior to the intended date of MRI scanning. Reasons for exclusion prior to the MRI scanning session were: opted out or otherwise unable to schedule scan session (N=6 minors and 8 adults [18 years or older]); had a CBCL Total Problems Score (N=5 minors) or SCL-90-R Global Severity Index (GSI; N=4 adults) in the clinical range; left-handedness (N=1 minor); had contraindication(s) to MRI (N=2 adults); and diagnosed with a psychiatric condition or learning disability (N=2 minors). No participants scored below our inclusion threshold for IQ (>2 SD below the mean). Of those 97 participants who were scanned, 11 were excluded from all further analyses for the following reasons: did not provide at least 2 fMRI runs of the encoding task due to (a) terminating the session early (N=3 minors) or (b) excessive motion, defined as <2 encoding runs with <1/3 of the timepoints exceeding our framewise motion threshold (see below; N=6 minors); incidental finding (N=1 child); and technical difficulties with data acquisition (N=1 minor). The final sample reported here following exclusions outlined above includes 86 individuals whose ages on the date of MRI scanning ranged from 7.16 years to 29.42 years.
Non-participation	Three minor participants dropped out of the MRI session early (N=1 felt ill/claustrophobic; N=2 felt tired).
Randomization	Participants were not allocated into different experimental groups; participant age on the date of MRI scan dictated a given participants' point on our age range and thus cannot be randomized. Assignment of stimuli to conditions was randomized across participants; block orders were counterbalanced.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	We recruited participants from the Austin, TX area through physically and digitally posted advertisements and word-of-mouth. While efforts were made to recruit a diverse sample, participants who expressed interest were generally from socioeconomically advantaged and/or educated homes and of above-average intelligence. This self-selection bias might relate to the fact that participation in our study involved families traveling to our campus to participate on weekends or after school. If anything, these sort of characteristics of our sample of minors might underestimate the differences between children and adults, on average.
Ethics oversight	The University of Texas at Austin Institutional Review Board (IRB) approved the study protocol. In line with our approved protocol, adult participants provided informed consent, and permission was obtained from one or more parents/guardians of minor (i.e., individuals under the age of 18 years) participants. Minors additionally provided informal assent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type	Task-based fMRI; mixed design (event-related trials blocked by condition).
Design specifications	Participants completed a maximum of 288 encoding trials during fMRI scanning (24 pairs per run split evenly among AB, BC, and non-overlapping conditions; each presented three times per run). Participants who failed to complete all encoding runs, or provided MRI data of poor quality for one or more runs, had correspondingly fewer trials. Pairs were presented for 3.5 seconds with a 0.5s ISI. Encoding were jittered through the insertion of an average of 1 (range 0-2), 2-second baseline trials in between pairs. This means that the stimulus onset asynchrony ranged from 4-8 seconds (with an average of 6s. Pairs were additionally blocked by type (see Figure 1A), and blocks had a fixed duration of 24s.
Behavioral performance measures	Our primary behavioral performance measures came from the memory test after each run. We recorded responses and response times, and used a linear mixed effects regression model to assess whether performance (likelihood of making a correct response; response time for correct trials) varied significantly across conditions and/or ages. We additionally imposed the requirement that participants show above-chance memory performance (assessed using a binomial test for the direct pairs - AB, BC, XY; requires 13 correct trials of 24) for each run included in the analyses, ensuring in all data reported here participants were paying attention during the encoding task.

Acquisition

Imaging type(s)	functional and structural
Field strength	3T
Sequence & imaging parameters	Functional data were collected in 75 oblique axial slices using an EPI sequence, oriented approximately 20° off the AC-PC axis (TR = 2000 ms, TE = 30 ms, flip angle = 73; 128 x 128 x 75 matrix, 1.7 mm isotropic voxels, multiband acceleration factor = 3, GRAPPA factor = 2). Between one and three field maps were collected (TR = 589 ms, TE = 5 ms/7.46 ms, flip angle = 5 degrees; matrix size = 128 x 128 x 60; 1.5 x 1.5 x 2 mm voxels) for each participant to correct for magnetic field distortions. Fieldmaps were planned (1) before the first study run, (2) before the first visual localizer run, and (3) any time a participant came out of the scanner for a break. Four participants had only one fieldmap acquired due to technical difficulty and/or operator error. Two to three oblique coronal T2-weighted structural images were acquired perpendicular to the main axis of the hippocampus and in approximately the same orientation as one another (TR = 13150 ms, TE = 82 ms, 384 x 60 x 384 matrix, 0.4 x 0.4 mm in-plane resolution, 1.5 mm thru-plane resolution, 60 slices, no gap); these images were not incorporated into the analysis for the present manuscript. A T1-weighted 3D MPRAGE volume (256 x 256 x 192 matrix, 1 mm isotropic voxels) was also collected for automated segmentation using Freesurfer and spatial normalization to the MNI template brain using ANTS.
Area of acquisition	Whole brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	Data were preprocessed and analyzed using FEAT (FMRI Expert Analysis Tool) Version 6.00, part of FSL Version 5.0.9 (FMRIB's Software Library, http://www.fmrib.ox.ac.uk/fsl) and Advanced Normalization Tools (ANTS) 82. Motion correction was applied to each functional run using MCFLIRT and then non-brain structures were removed using BET, both part of FSL. All functional runs were then registered to the middle functional "reference" run (in most cases, the third study run) by applying affine transformations calculated in ANTS. Anatomical images (mean coronal, MPRAGE) were then registered to the functional reference run following fieldmap-based unwarping of the functional data (implemented in FEAT as part of general linear model [GLM] analysis; see below) as follows. Each participant's MPRAGE was directly registered to their functional data using ANTS affine transformations. Non-brain structures were removed from anatomical images using a mask derived from

Freesurfer output. The result of the registration process was that all data (functional and structural, including Freesurfer parcellations) was coregistered in each participant's native functional space. All analyses were carried out in this native space with the exception of group-level GLMs. In preparation for both univariate (GLM) and multivariate (MVPA) analyses, the following pre-statistics processing was applied: fieldmap-based EPI unwarping using PRELUDE+FUGUE; spatial smoothing using a Gaussian kernel of FWHM 4mm; grand-mean intensity normalization of the entire 4D dataset by a single multiplicative factor; highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with sigma=50s).

Normalization: ANTs 2.1 was used for normalization.

Normalization template: MNI152 2mm T1

Noise and artifact removal: Realignment parameters from MCFLIRT were used to compute framewise displacement (FD) for each fMRI volume. Motion parameters calculated during the motion correction step and their temporal derivatives were added as additional confound regressors. Framewise displacement (FD) and DVARS, two measures of framewise data quality, were also added to GLMs as regressors of no interest.

Volume censoring: No volume censoring was performed.

Statistical modeling & inference

Model type and settings: fMRI data was submitted to mass univariate GLMs for the purposes of the univariate analyses and to generate single-trial parameter estimates for the trial-by-trial reactivation analysis. This was performed separately for each subject and scanning run in within-subject, fixed-effects analyses. GLM FOR MULTIVARIATE: Trial-level neural patterns were generated under the assumptions of the GLM using a modified LS-S approach. Statistics images associated with each encoding trial were estimated for each repetition and participant using custom Python routines. The resulting single-trial estimates were then submitted to an MVPA that calculated reactivation across trials. Resulting MVPA reactivation scores were then submitted to (generalized) linear mixed effects regression models with subjects treated as random effects (slopes and intercepts). GLM FOR MASS UNIVARIATE: Statistics images were combined across runs within subject using fixed effects, and across subjects using mixed effects (in FSL, FLAME1).

Effect(s) tested: The central effect tested was memory reactivation (MVPA classifier evidence for stimulus A content type) varied across age (an across-participant factor) and was the degree that the dimensionality of neural representations changed over learning and how this change varied across problem complexity. Learning block and problem complexity were fully crossed and were within-participant factors.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s): Content-sensitive ventral visual stream regions were defined anatomically for each participant by summing entorhinal cortex, fusiform gyrus, inferior temporal cortex, and parahippocampal gyrus regions identified by an automated labeling algorithm (Freesurfer). The resulting ventral temporal cortex (VTC) region was used to mask functional data for MVPA. We also defined regions in MNI template space for small-volume correction of univariate analyses. Medial prefrontal cortex (MPFC) was delineated by hand on the 1mm MNI template, restricted to those regions in the "medial prefrontal network" described in previous work. We used the Harvard-Oxford atlas to define both inferior frontal gyrus (IFG) and hippocampus (HPC) ROIs.

Statistic type for inference (See Eklund et al. 2016): We used cluster-wise thresholding methods for mass univariate analyses. Group statistical maps were thresholded at a voxelwise p<0.005 and submitted to cluster correction as follows. Smoothness was estimated using the residuals (warped to MNI template space) from every study run for each participant using AFNI's 3dFWHMx utility. We used the new spatial AutoCorrelation Function estimation method (-acf flag), which no longer assumes Gaussian noise distribution and generally results in a larger (more conservative) estimate of smoothness relative to prior releases of this tool, thus reducing the likelihood of a Type I error. We then used these run-level values to compute the average smoothness parameters across all encoding runs within participant, and then finally across participants to yield a group-level mean smoothness estimate. This entire analysis was done separately for each ROI (grey matter, HPC, IFG, and MPFC) within which cluster correction was performed.

Correction: Minimum cluster extents at a significance threshold of p<0.05 were determined for each ROI using 3dClustSim (settings: p-value threshold = 0.005, corrected alpha value = 0.05, NN approach = second-nearest neighbor clustering [faces or edges], thresholding = 2-sided) Minimum cluster sizes were determined to be 10 voxels for HPC, 17 voxels for IFG, 27 voxels for MPFC, and 71 voxels for grey matter. All clusters exceeding these criteria either within the three a priori anatomical regions and/or at the whole brain grey matter level are reported here.

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis: We trained a pattern classifier (sparse multinomial logistic regression [SMLR] implemented in PyMVPA; lambda=0.1, the package default) to predict face- from scene-viewing on the basis of fMRI activation patterns within ventral temporal cortex. For the cross-validation analysis (localizer task data, perception of faces vs. scenes), the classifier was trained on a subset of localizer task runs and tested on the held-out run;

this process was repeated until all runs had been held out exactly once. We evaluated the classifier performance using accuracy, where the classifier prediction (face or scene) was compared with the actual stimulus type. For the memory reactivation analysis, the classifier was trained on all localizer runs and applied to the memory task data. We then used the classifier probabilities to compute a reactivation index for each participant (blockwise analysis) or normalized using log odds (trialwise analysis), which we then analyzed using linear mixed-effects models to assess repetition- and age-related differences (blockwise); as well as how within-subject variability in reactivation related to the probability of making a correct response.