Neural Mechanisms of Creative Problem Solving: From Representational Change to Memory Formation

Maxi Becker[1], Tobias Sommer[3], Roberto Cabeza[1,2]

[1]Humboldt University Berlin,

Department of Psychology, Berlin, Germany

[2]Duke University,

Center for Cognitive Neuroscience, Durham, NC, USA 27708

[3]University Medical Center Hamburg-Eppendorf,

Institute of Systems Neuroscience, Hamburg, Germany

Corresponding author: Maxi Becker

Email address: maxi_becker@gmx.net

Acknowledgements: This work was funded by the Einstein Foundation Berlin (EPP-2017-423) and MB was additionally funded by the Sonophilia Foundation. We would like to express our deepest gratitude to Luise Oehler for her substantial assistance in data assessment and John D. Haynes, Simon W. Davis, Cortney Howard and Loris Naspi for useful discussions on experimental design and data analysis.

Declaration of Interests: All authors report no financial interests or conflicts of interest.

Author Contributions: M.B. & R.C. conceived of the study. M.B. acquired, analysed the data and wrote the manuscript. M.B., T.S. & R.C. contributed to revising the manuscript.

Summary

Despite the need for innovative solutions to contemporary challenges, the neural mechanisms driving creative problem-solving still remain largely unknown. We focused on the powerful creative process of insight, wherein rapid knowledge reorganization and integration—termed representational change—yield solutions that evoke suddenness, positive emotion, and enduring memory. We posit that this process manifests as abrupt shifts in activation patterns within brain regions housing solution-relevant information, including the visual cortex for visual problems, alongside regions linked to feelings of emotion, suddenness and subsequent memory. Our findings substantiate these hypotheses, revealing sudden representational changes in visual cortex, conjoined with activations in the amygdala and hippocampus—forming an interconnected network. Importantly, the representational change and hippocampal effects serve as predictive markers of subsequent memory. This study provides the foremost clear evidence of an integrated insight mechanism and its lasting memory influence. Implications extend to education, emphasising the significance of discovery-based learning approaches for boosting knowledge retention.

Introduction

Creativity and memory have a synergistic relationship. Finding creative solutions depends on having relevant preexistent knowledge, and, conversely, novel solutions are typically well-remembered, promoting new knowledge of the world and future adaptation. *Insight* is a fundamental process in creative problem-solving that occurs when a problem is solved via a novel approach by reorganising preexistent knowledge structures (Danek et al., 2020; Dietrich & Kanso, 2010). Insight problem solving, which is often studied in the context of scientific discoveries (Kounios & Beeman, 2015), is associated with boosted subsequent long-term memory, sometimes after a single experience, unlike other forms of learning that require multiple repetitions (Ash et al., 2012; Danek et al., 2013; Kizilirmak, Galvao Gomes da Silva, et al., 2016; Ludmer et al., 2011). The study of insight can shed light on mechanisms of rapid learning and memory formation, and has practical implications in educational settings, as insight-driven learning experiences can be effective in improving knowledge retention (Gruber et al., 2014; Ludmer et al., 2011).

This raises the question of how the insight process works and why it is often associated with better memory? The insight process has cognitive and evaluative components. The main cognitive component is associated with *representational change* (RC) whereby internal (conceptual or perceptual) representations of the solution are reorganised and integrated into a coherent whole (Ohlsson, 1984; Schilling, 2005). This process is assumed to involve a change in the pattern of currently activated cell assemblies coding the problem in domain-specific cortex (Becker et al., 2024; Knoblich et al., 1999). For instance, in visual problems requiring object recognition, RC may involve breaking up perceptual chunks into their constituent elements to make them available for a novel recombination (Tang et al., 2015). Multivariate functional magnetic resonance imaging (fMRI) can identify representation-related activation patterns in regions such as visual cortex (Kriegeskorte et al., 2008; Norman et al., 2006), which can be used to detect RC.

The evaluative component (or *AHA! experience*), which is characterized by positive emotions and the conviction that the solution is correct and arrived suddenly (Danek & Wiley, 2017; Kizilirmak & Becker, 2022), has been associated with activations in

the amygdala and the hippocampus (Becker, Sommer, et al., 2020; Jung-Beeman et al., 2004; Kizilirmak, Thuerich, et al., 2016; Shen et al., 2017; Tik et al., 2018). Amygdala activation may reflect the positive emotions (Janak & Tye, 2015), and the hippocampus activation, the sudden arrival and/or surprising content of the solution. It is well-known that novelty (Axmacher et al., 2010; Nyberg, 2005) and surprise (Kumaran & Maguire, 2006, 2007; Schiffer et al., 2012) activate the hippocampus, particularly its anterior formation. Importantly, at the time of insight, cognitive and evaluative elements are unlikely to function in isolation but rather in a closely interconnected manner, facilitating the efficient processing of solution-relevant information (Mednick, 1962; Schilling, 2005). Finally, the neural mechanisms of this *insight-related better memory* remain largely unknown (Kizilirmak et al., 2019). However, both insight components likely contribute to this effect, as has been demonstrated in behavioural studies (Ash & Wiley, 2008; Auble et al., 1979; Danek et al., 2013; Danek & Wiley, 2020; Kizilirmak, Galvao Gomes da Silva, et al., 2016). There is evidence that the quality of visual cortex representations during the perception of visual stimuli predicts later memory (Davis et al., 2021), and the fundamental contributions of hippocampus in declarative memory encoding has been continuously demonstrated (for reviews, see (Eichenbaum et al., 2007; Moscovitch et al., 2016).

In sum, insight during problem solving likely reflects sudden multivariate pattern changes associated with enhanced solution-relevant representations (RC). This sudden conceptual change (RC) is efficiently integrated into the solution process leading to awareness of the solution accompanied by an emotional and suddenness/surprise response in hippocampus and amygdala resulting in enhanced encoding of the solution and better subsequent memory.

To investigate this neural mechanism of insight and its impact on subsequent memory, the current study scanned participants with fMRI while performing a *visual insight task*, in which they tried to identify (solve) hard-to-recognize, high-contrast photos of real-world objects (*Mooney images*, see Fig.1-A/B; (Imamoglu et al., 2013; Kizilirmak, Thuerich, et al., 2016; Ludmer et al., 2011). Due to ambiguous figure/ground separation, black/white contours in these images are often grouped into meaningless shapes, hindering successful object identification. Perceptually rechunking or regrouping these contours can lead to sudden object identification,

eliciting insight (Kizilirmak, Thuerich, et al., 2016; Ludmer et al., 2011). Following Mooney object identification, participants provided ratings of the perceived suddenness, emotion, and certainty about the solution — which were used to index insight – and then selected the semantic category of the hidden object. Five days later, they performed a *recognition task* with confidence measures, and then generated the name of the hidden object.

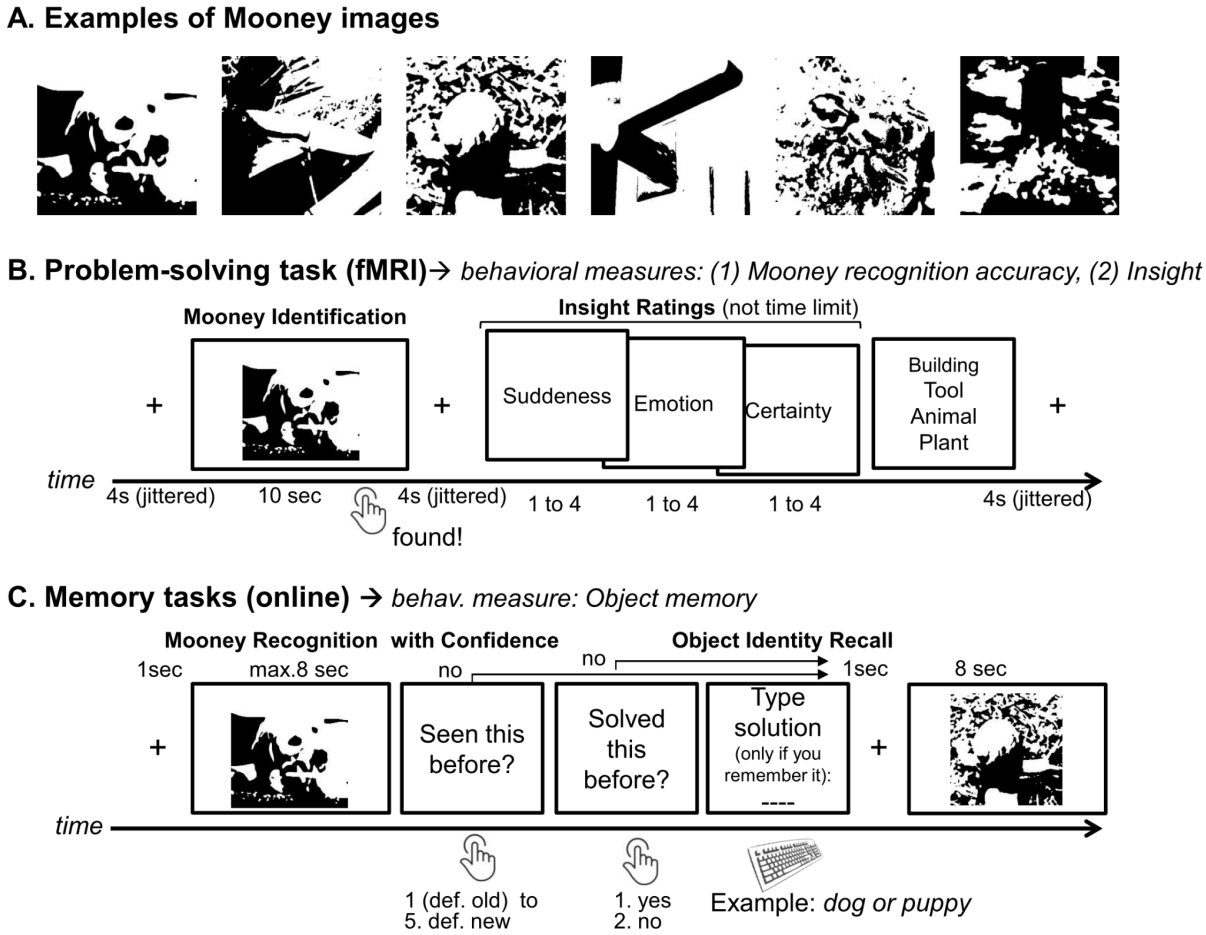

***Figure 1**. Experimental design*

Note. **Panel A** shows examples of Mooney images (from left to right: dog, knife, snail, doorknob, spider, tombstone). **Panel B** describes the Mooney paradigm presented in the scanner. **Panel C** illustrates the subsequent memory test seven days after the solving phase in the MRI.

To evaluate this neural mechanism of insight and its impact on memory, we tested four specific fMRI hypotheses. First, *ventral occipito-temporal cortex (VOTC) exhibits visual insight-related RC.* We focused on six VOTC regions-of-interest (ROIs) associated with visual object recognition (posterior and anterior fusiform gyrus–pFusG, aFusG; inferior lateral occipital complex–iLOC; anterior, mid, and

posterior inferior temporal gyrus–aITG, mITG, pITG). RC was measured during Mooney object identification by comparing multivariate activity patterns before and after the solution was found, and determined if the pattern change was greater for high than low insight trials.

Second, *the amygdala and hippocampus show insight-related activity*. We examined if at the time of the solution, activity in these regions was greater for high than low insight trials.

Third, *solution-relevant information is more efficiently integrated in a "solution network" consisting of VOTC, amygdala, and hippocampus during insight*. Given that the entire VOTC visual hierarchy contributes to object identification (DiCarlo et al., 2012) and they are all closely interconnected, we employed the whole VOTC in this analysis. Efficient information integration was quantified via functional connectivity of the *solution network* as well as different graph measures assumed to be higher high versus low insight trials.

Finally, we hypothesised that *both insight components (cognitive and evaluative) are associated with better subsequent memory*. In particular, we expected that RC in VOTC and hippocampal activity predict *insight*-related better memory. Insight-related memory was identified by behavioural and neural differences between problems solved with high- compared to low-insight that were later remembered vs. forgotten.

Results

Behavioural results

In line with previous research, insight during the Problem Solving Task (Fig.1-B) was quantified using the three 4-point insight ratings following Mooney object identification (Danek & Wiley, 2017, 2020), namely *positive emotion* (M=2.8, SD=.41), *suddenness* (M=3.0, SD=.39) and *certainty* (M=3.14, SD=.34). However, due to their significant factor loadings onto a single latent factor (λ=.62-.57; ps<.001, see Supplements Fig.S2; (Danek & Wiley, 2017; Webb et al., 2016)), the three ratings were combined to one insight measure. To ensure a similar amount of trials per condition (high and low insight), a median split of the sum of the three ratings

(max: 12 points) was used to distinguish between high (*HI-I)* or low (*LO-I*) insight trials. All analyses were repeated with a continuous *insight sum* measure (sum of suddenness, emotion and certainty ratings). The results did not change significantly, except when specifically mentioned.

On average, participants indicated a solution for 68.5% (*SD*=16.4%) of all presented Mooney images. The solution, i.e. the subsequently selected category of the object, was correct in 51.8% (*SD*=12.8%) of these cases. Of all correctly identified Mooney images, 65.2% (*SD*=6.6%) were solved with *HI-I* and 34.7% with *LO-I* trials. Consistent with previous results, this *insight-accuracy* effect was significant ($Chi^2(1)=98.68$, $p<.001$, odds ratio [*LO-I*]=.32) (Becker, Wiedemann, et al., 2020; Danek & Salvi, 2020). Of all incorrectly identified Mooney images, 40.5% (*SD*=21%) were solved with *HI-I* and 59.5% (*SD*=21%) with *LO-I*.

The median response time for correct solutions was 3.5sec (*SD*=0.8sec) where participants were faster during *HI-I* (2.75sec, *SD*=0.81sec) than *LO-I* solutions (5.08sec, *SD*=1.12sec). This difference in response time between both conditions was significant ($Chi^2(1)=574.5$, $p<.001$, $ß=.35$). The median response time for incorrect *HI-I* solutions was 3.4sec (*SD*=1.2sec) and 6.1sec (*SD*=1.9sec) for *LO-I* solutions (see Fig.S3, Supplements). Response time and accuracy has been shown to covary with the difficulty of Mooney images (Becker et al., 2021). Because we observed a systematic difference in response times and accuracy between the *HI-I* and *LO-I* condition but were not interested in differences in difficulty, response time and accuracy were included as covariates of no interest into all subsequent analyses if appropriate.

Subsequent memory of each Mooney object identification was measured five days later. Due to the problem solving context, we were primarily interested in memory for the solution (i.e., object identity recall) but in order to recall the Mooney images, participants had to correctly recognize the image and to remember having solved it first (see Fig.1-C). Hence, subsequent memory for object identity was defined as correct recognition *and* memory of having solved it *and* memory for the solution, i.e. the name of the depicted object. Participants correctly recognized 68.7% (*SD*=15.4%) of the Mooney images (false alarm rate 19.3%, *SD*=14.1%). For 74.0% (*SD*=13.7%) of the correctly recognized images, they remembered correctly having solved it and for 61.6% (*SD*=14.7%) also the solution, i.e. the Mooney object. Median

response time was 2.50 sec (*SD*=1.05) for all old Mooney images. Images correctly identified via *HI-I* in the scanner were recognized in 1.85sec (*SD*=.80) and via *LO-I* in 2.38sec (*SD*=.88). Median response time was 3.73sec (*SD*=2.0) for all new Mooney images.

Trials solved via high vs low insight are remembered better

To investigate the behavioral insight-related better memory, we estimated the effect of insight on subsequent memory using a general linear mixed (binomial) model including random subject and item intercepts and controlling for accuracy and the respective run order (1-4). This insight factor (*not solved*, *HI-I, LO-I*) significantly predicted variance in subsequent memory ($Chi^2(2)=628.53$, $p<.001$; see Fig.2-B). Posthoc analyses revealed that trials solved either with *HI-I* (69%; 95CI[59%, 77%], $z=21.73$, $p<.001$) or *LO-I* (34%; 95CI[25%, 45%], $z=12.82$, $p<.001$) were more likely to be remembered than trials that were not previously solved in the scanner (7%; 95CI[4%,11%]).

Critically, *HI-I* trials were significantly more likely to be remembered compared to *LO-I* trials ($z=13.06$, $p<.001$), confirming insight-related better memory. This effect remained significant when additionally controlling for response time ($Chi^2(1)=67.45$, $p<.001$, odds ratio: *HI-I*=2.09).

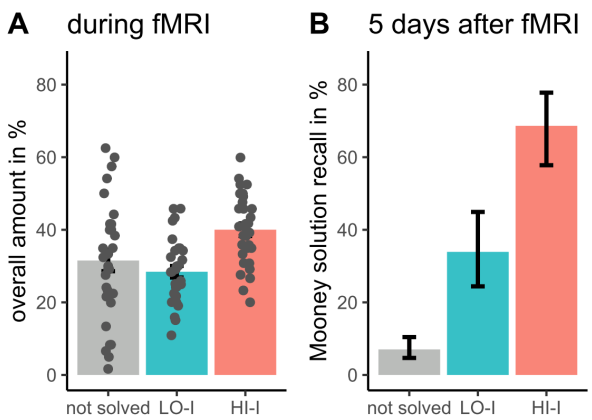

Figure 2. Behavioral Insight Memory Effect.

Note. *HI-I* = solved trials with high insight; *LO-I* = solved trials with low insight; **Panel A**. overall proportion of solved trials; **Panel B.** Behavioural insight memory effect; estimates are marginal mean percentages to correctly recall the Mooney image in the respective condition. Error bars represent 95% confidence intervals.

Hypothesis 1: VOTC exhibits visual insight-related RC

During insight-related RC, the content representation of the Mooney image changes suddenly from meaningless to meaningful when the object is identified. Our first hypothesis was that VOTC regions (aFusG, pFusG, iLOC, aITG, mITG, pITG) show this insight-related RC, as reflected in changes of distributed activity patterns from pre to post solution. In particular, we operationalized this change by two consecutive representational similarity analyses (RSA; (Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013). First, by correlating the activity patterns pre (0.5 sec after Mooney image onset) and post solution (button press), and by determining if the correlation decreased more strongly (i.e., greater change) for *HI-I* than *LO-I* solutions (see Fig.3-A). Second, in areas showing greater change for *HI-I* than *LO-I* solutions as a necessary condition for insight-related RC we tested whether this change reflects a stronger conceptual reorganisation of the identified Mooney object from meaningless to meaningful. For this purpose, we created two model-based representational similarity matrices (RSM) that were derived from two conceptually highly distinct neural networks, AlexNet and Word2Vec (Krizhevsky et al., 2017; Mikolov et al., 2013). AlexNet was trained using large image databases, Word2Vec using large text corpi. The correlation of the distributed activity patterns elicited by the Mooney images and these model-based RSMs reflects the neural representational strength of the Mooney stimulus. We contrasted the representational strength pre and post solution and examined if it increased more from pre to post solution for *HI-I* than *LO-I* (see Fig.3-B1:4 for further details).

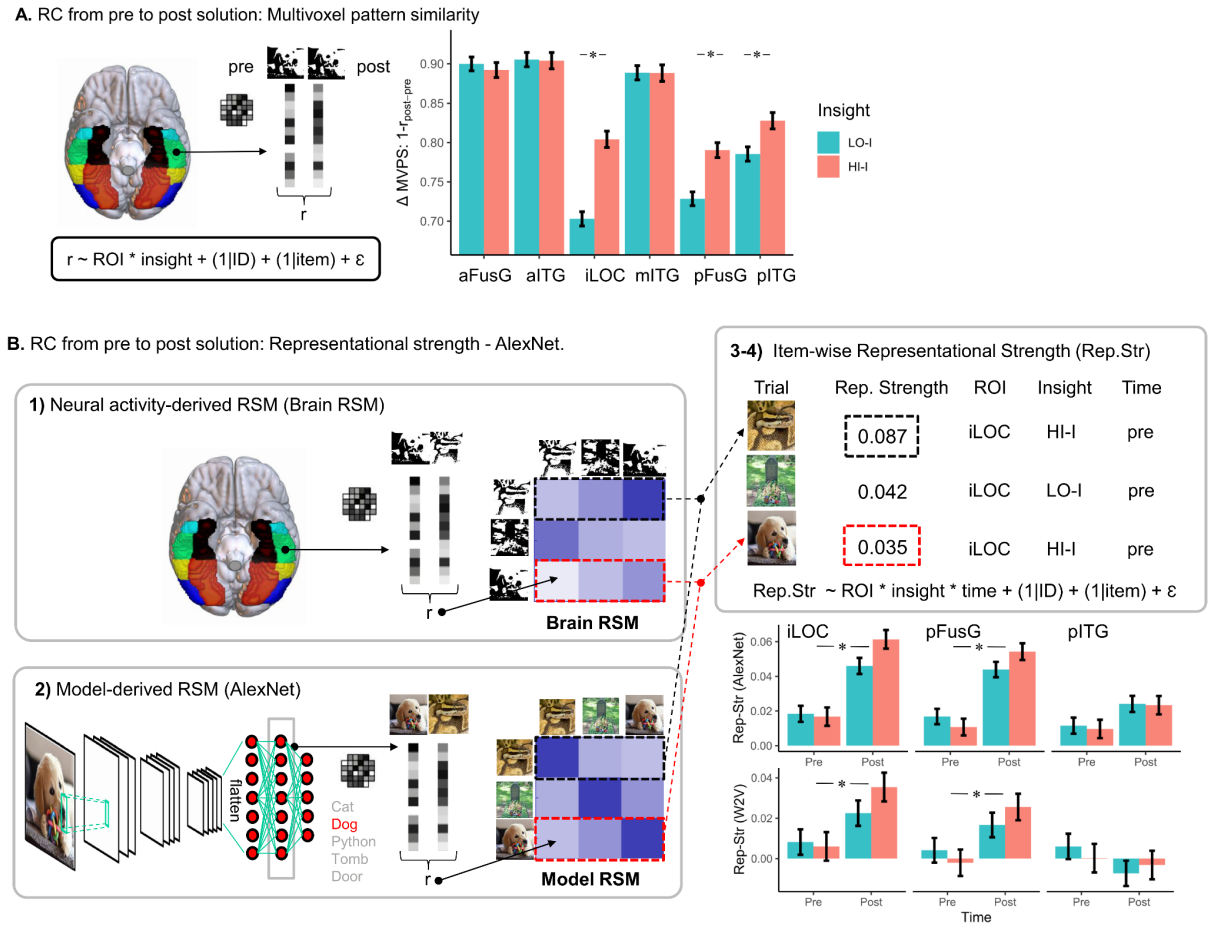

Figure 3. VOTC regions show Representational Change (RC)

Note. Values = estimated marginal means ± SEM; asterisk = statistical significance at p<.05; *HI-I* = solved trials with high insight. **Panel A. RC from pre to post solution: Multivoxel pattern similarity.** Multivoxel patterns per ROI for each time point (pre- and post solution) are extracted and subsequently correlated. Pre= 0.5sec after stimulus presentation; post = during solution button press. Those Pre-Post Solution Similarity values (r) are subsequently estimated in a linear mixed model as a function of insight and ROI. Bar plots show change (Δ) in Multivoxel Pattern Similarity (MVPS) analysis. Change in MVPS = 1 minus the correlation between the post and pre solution multivoxel pattern in the respective ROI. **Panel B. RC from pre to post solution: Representational strength - AlexNet.** This RSA method employs four steps. (**1**) A neural activity-derived RSM (Brain RSM, size 120x120) is generated for each region-of-interest (ROI) and each time point (pre- post solution, see Panel A) where each cell is representing a multivoxel pattern similarity value for each Mooney image pair. (**2**) A conceptual *Model RSM* (here using AlexNet, size 120x120) is generated where each cell is representing a similarity value for each Mooney object pair. (**3**) For each brain region and each time point, the row of each stimulus (~120) in the *Model RSM* and in the *Brain RSM* are correlated yielding a representational strength measure (i.e. brain-model fit) per region and time point. (**4**) The representational strength is used as a dependent variable in linear mixed models to investigate which ROIs exhibit an insight-related increase in representational strength from pre- to post solution (time). Barplots show Representational Strength [Rep-Str] = "second order correlation" between multivoxel patterns in respective ROI at pre and post response time point and conceptual *Model RSMs* (AlexNet and Word2Vec [W2V]).

RC from pre to post solution: Multivoxel pattern similarity. To identify ROIs that showed the hypothesised RC related changes in the correlation between pre- and post solution activity patterns, we estimated the effect of insight in the above named VOTC regions on this correlation and their interaction using a linear mixed model including random subject and item intercepts and controlling for response time, accuracy and the run order. There was a significant main effect for insight (*HI-I* vs. *LO-I)* ($Chi^2(1)=115.66$, $p<.001$, $ß=.10$) over all regions and an interaction between ROI and insight indicating differences between ROI regarding changes in pre- and post solution activity patterns ($Chi^2(5)=267.27$, $p<.001$) (Fig.3-A). Posthoc analyses revealed that bilateral iLOC ($z=-14.61$, *p-Bonferroni*<.001), bilateral pFusG ($z=-12.27$, *p-Bonferroni*<.001) as well as bilateral pITG (z=-6.02, *p-Bonferroni*<.001) significantly reduced similarity in activity patterns from pre to post solution during *HI-I* compared to *LO-I*. None of the other brain areas showed a significant insight condition effect in the hypothesised direction.

RC from pre to post solution: Representational strength - AlexNet. We restricted the representational strength analyses to areas (iLOC, pFusG and pITG) that showed the hypothesised greater change in pre- to post solution activity patterns. To identify areas where the representational strength increased more strongly from pre to post solution for *HI-I* than *LO-I* (see Fig.3-B), we estimated a mixed model with the factors time (pre vs. post), insight (*HI-I vs. LO-I*) and ROI as well as their interaction. We additionally included a random subject and item intercept and controlled for response time, accuracy and the respective run order (1-4).

The model with the two-way condition*time interaction was significant ($Chi^2(1)=21.54$, $p<.001$, $ß=.06$) indicating an overall insight effect over all three regions. Furthermore, the model with the three-way insight condition*time*ROI interaction significantly explained variance in representational strength ($Chi^2(6)=68.31$, $p<.001$). Posthoc analyses revealed a significant insight condition*time interaction effect for bilateral pFusG ($Chi^2(1)=17.23$, *p-Bonferroni*<.05, $ß=.08$) and bilateral iLOC ($Chi^2(1)=9.49$, *p-Bonferroni*<.05, $ß=.08$) while pITG (p>.81) did not reach significance (see Fig.3-B).

RC from pre to post solution: Representational strength - Word2Vec. To ensure that the interaction results obtained via the AlexNet model are robust and generalize to other conceptual models, we also measured RC using a purely verbal conceptual model – Word2Vec (Mikolov et al., 2013). As displayed in Fig.3-B, the results largely replicated those found with AlexNet. The two-way insight*time interaction ($Chi^2(1)=21.47$, $p<.001$, $ß=.07$) was significant indicating an insight effect over all three regions. However, the three-way insight condition*time*ROI interaction was also significant ($Chi^2(6)=73.91$, $p<.001$). Posthoc analyses revealed a significant condition*time interaction effect for bilateral pFusG ($Chi^2(1)=12.50$, *p-Bonferroni*<.05, $ß=.08$) and bilateral iLOC ($Chi^2(1)=6.46$, *p-Bonferroni*<.05, $ß=.08$) while pITG (p>.08) did not reach significance.

In sum, the combination of distributed activity patterns changes and increase in representational strength of the Mooney objects from pre to post solution confirm RC in VOTC, particularly in pFusG and iLOC. These findings support our first hypothesis that VOTC-RC is enhanced by insight. Although these results provide a strong candidate for a neural mechanism of perceptual/conceptual reorganisation (RC), i.e. the cognitive component of insight, they cannot explain the evaluative component of insight, namely the positive emotion, certainty, and suddenness associated with the insight experience. In line with previous research, we hypothesised that this evaluative component is mediated by the amygdala and hippocampus.

Hypothesis 2: Amygdala and hippocampus show insight-related activity

To examine our second hypothesis that the evaluative component of insight is mediated by the amygdala and anterior/posterior hippocampus, we compared amygdala and hippocampal activity during Mooney object identification in *HI-I* vs. *LO-I* trials. Note, we used univariate activity because both brain regions are unlikely to represent visual information like VOTC. Univariate activity during solution, i.e. onset of the button press, for *HI-I* and *LO-I* trials was estimated using GLMs as implemented in SPM (Ashburner, 2012). To test whether amygdala and hippocampal activities were predicted by insight (*HI-I*, *LO-I*), we computed a linear mixed model as a function of the insight condition including a random subject intercept and controlling for the run order (1-4) and the respective ROI. To control for differences in

response time and accuracy as a proxy for difficulty between *HI-I* and *LO-I* trials, we repeated this analysis using single trial GLMs and included trial-specific response time and accuracy as covariates of no interest which did not change the significance of the results (see Supplementary Material).

As illustrated by Fig.4-A, amygdalar and hippocampal activity significantly distinguish between *HI-I* and *LO-I* trials during solution ($Chi^2(1)=29.80$, $p<.001$) but there are also differences between the ROIs ($Chi^2(1)=22.49$, $p<.001$). Posthoc analyses revealed that amygdala activity ($t(707)=4.026$, $p<.001$, ß=.28) and anterior hippocampus ($t(707)=3.873$, $p<.001$, ß=.35) was significantly higher for *HI-I* compared to *LO-I* trials while posterior hippocampus only showed a trend towards significance in the hypothesised direction for this condition ($p=.099$). To investigate potential normalisation artefacts in the small amygdala and hippocampal ROIs, we also conducted an ROI analysis in native space (see Supplementary Material) and the results stayed the same except that the positive trend for *HI-I>LO-I* trials in posterior hippocampus reached significance.

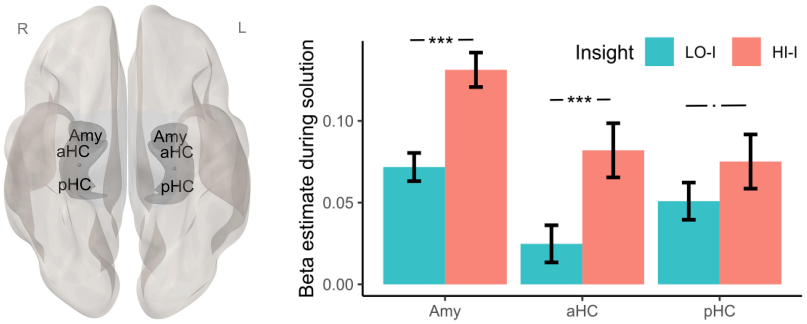

Figure 4. Amygdala and hippocampal activity during solution are associated with insight-related better memory.

Note. Values = estimated marginal means ± SEM. *** represents p-value <.0001; - .- represents p-value <.10. Amygdala (Amy), anterior (aHC) and posterior (pHC) hippocampal activity during solution divided by condition (*HI-I* [high insight], *LO-I* [low insight]).

Activity in amygdala and hippocampus during solution is associated with insight on a trial by trial level when controlling for response time and accuracy

To rule out the influence of response time and accuracy differences on the observed effects in amygdala and hippocampus activity, we performed a control analysis. Using covariates to remove unwanted variance is not feasible in the SPM GLM

framework. Therefore, we conducted the analyses with the data from the single-trial analyses containing a beta estimate of univariate amygdala, anterior and posterior hippocampal activity (ROI) for the solution button press at each trial. The model included insight (HI-I, LO-I) as dependent variable and response time, accuracy and run order (1-4) as covariates of no interest as well as a random subject and item intercepts.

There was a main effect for condition ($Chi^2(1)$= 69.68, p<.001, ß=.11) implying that brain activity was higher for *HI-I* (M=1.06, 95%CI[0.86,1.26]) than for *LO-I* (M=0.64, 95%CI[0.44 – 0.84]) insight trials during solution over all three ROIs. There was also a condition*ROI interaction effect ($Chi^2(2)$= 23.66, p<.001) implying differences between the ROIs in this condition. Post Hoc analyses revealed that amygdala activity (z=7.61, p<.001) was significantly higher for *HI-I* (M=1.54, 95%CI[1.33 – 1.75]) than for *LO-I* (M=0.97, 95%CI[0.751.19]) insight trials during solution. Anterior hippocampus (z=6.96, p<.001) was also higher for *HI-I* (M=1.05, 95%CI[0.84,1.26]) than for *LO-I* (M=0.53, 95%CI[0.31,0.75]) trials during solution. Posterior hippocampus only showed a trend towards significance for insight (z=1.77, p=.076) in the same direction.

Hypothesis 3: The VOTC, amygdala, and hippocampus network becomes more integrated during insight

Next, we examined the assumption that the sudden conceptual update (RC) is efficiently integrated into the solution process. For this, we created a *solution network* consisting of all bilateral VOTC ROIs (aITG, mITG, pITG, aFusG, pFusG, iLOC) relevant for object identification (DiCarlo et al., 2012). The *solution network* (18 ROIs in total) further included amygdala and anterior as well as posterior hippocampus because they are likely involved in the evaluation of the solution content. Efficient information integration was quantified via increased functional connectivity within this network during the solution process, i.e. stimulus presentation (ten seconds block while participants tried to identify the Mooney images). Efficient information integration was additionally quantified via graph measures representing functional segregation/integration and network centrality (Rubinov & Sporns, 2010) based on the functional connectivity values from our predefined *solution network* during stimulus presentation. We assumed increased network centrality (degree),

network functional integration (global efficiency, [reduced] average path length) as well as functional segregation (clustering coefficient, local efficiency) to represent efficient information integration.

Increased functional connectivity of the solution network during insight

Functional connectivity (FC) was computed between each pair of the ROIs in the *solution network* resulting in a sum FC score for each ROI (*Mass* = sum of t-squared statistics over all connections per ROI, Nieto-Castanon, 2020). The resulting connectivity values were predicted by the insight factor (*HI-I>LO-I*) in CONN. Note, due to hemispheric asymmetries in FC (Perez et al., 2023; Schöne et al., 2022), we split the ROIs in the left and right hemisphere but FDR-controlled for multiple comparison. The *solution network* showed significantly increased functional connectivity (*Mass*=382.25, *p-FDR*=.007) for *HI-I>LO-I* (see Fig. 5-A/B). The VOTC-RC ROIs, pFusG and iLOC, were particularly strongly connected to other ROIs in the network including amygdala and hippocampus. Note, the statistics with all individual connections in this network are listed in the Supplements, Table S1.

To quantify the connection strength between VOTC-RC ROIs and amygdala as well as hippocampus, we set up another *subnetwork* consisting only of left and right pFusG, iLOC, amygdala and anterior/posterior hippocampus (12 ROIs in total). For the contrast *HI-I>LO-I*, this subnetwork showed stronger functional connectivity during insight (*Mass*=91.36, *p-FDR*=.046) (see Fig. 5-C), indicating that VOTC-RC ROIs and amygdala and anterior hippocampus are also directly connected during insight. Note, the posterior hippocampus did not show any significant connections with VOTC-RC areas. The statistics with all individual connections in this subnetwork are listed in the Supplements, Table S2.

Increased functional segregation and integration of the solution network during insight

Insight-related increased functional connectivity in the *solution network* was consistent with results from different graph theory measures.

Network centrality. Degree ($t(30)=3.01$, $p=.002$, $ß=.42$) was increased for the contrast *HI-I>LO-I* indicating an insight-related stronger overall connectedness of ROIs in the *solution network*. Individual ROIs that showed more insight-related degree within the *solution network* were bilateral pFusG (right: ($t(30)=3.59$,

p-FDR=.008, *ß*=.49; left: *t*(30)=3.28, *p-FDR*=.008, *ß*=.55), bilateral iLOC (right: *t*(30)=2.79, *p-FDR*=.020, *ß*=.65; left: *t*(30)=3.44, *p-FDR*=.008, *ß*=.71) as well as bilateral anterior hippocampus (right: *t*(30)=2.14, *p-FDR*=.045, *ß*=.72; left: *t*(30)=2.68, *p-FDR*=.022, *ß*=.67) and bilateral amygdala (right: *t*(30)=2.28, *p-FDR*=.038, *ß*=.60; left: *t*(30)=2.32, *p-FDR*=.038, *ß*=.68).

Functional segregation. The *solution network* also exhibited increased local efficiency (*t*(28)=1.95, *p*=.030, *ß*=.03) as well as an increased clustering coefficient (t(28)= 2.44, p=.010, *ß*=.04) for *HI-I>LO-I*. No individual ROI survived multiple comparison correction for either of these two graph measures. However, right pFusG showed increased local efficiency (*t*(26)=1.89 *p-uncor*=.034, *ß*=0.03) and increased clustering coefficient (*t*(26)=1.90, *p-uncor*=.034, *ß*=0.03) when exploring uncorrected p-values. Those results imply that for *HI-I>LO-I* processing is more specialised as it occurs within more densely interconnected ROIs of the *solution network* (see Rubinov & Sporns, 2010).

Functional integration. Furthermore, the *solution network* exhibited an insight-related higher global efficiency (*t*(30)=1.81, *p*=.040, *ß*=.02) and a reduced average pathlength (*t*(30)=-2.55, *p*=.008, *ß*=-.06). For global efficiency, no individual ROI survived the threshold. For average pathlength, bilateral pFusG (left: *t*(30)=-2.62, *p-FDR*=.023, *ß*=-.09; right: *t*(29)=-2.37, *p-FDR*=.031, *ß*=-.05) and iLOC (left: *t*(30)=-3.29, *p-FDR*=.010, *ß*=-.09; right: *t*(30)=-3.14, *p-FDR*=.010, *ß*=-.08) showed significantly less average pathlength for *HI-I>LO-I.* Those results imply that information is rapidly integrated from different brain regions of the *solution network*. Specifically, those VOTC regions identified to be involved in RC (pFusG & iLOC) exhibit hub properties within the *solution network* including increased functional integration.

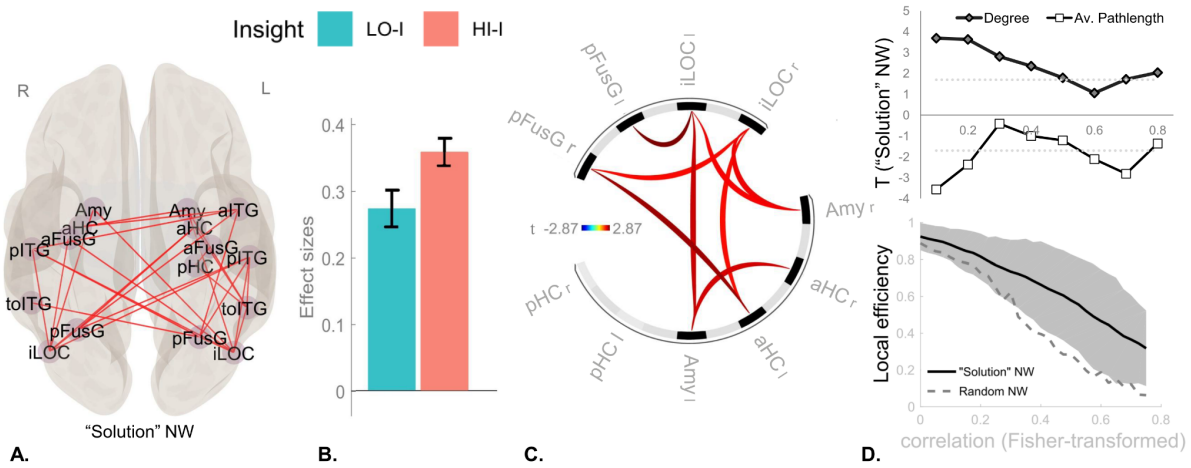

Figure 5. Insight-related increased information integration for *solution network* (NW) including VOTC-RC, amygdala & hippocampus during stimulus presentation.

Note. VOTC-RC = Ventral occipital temporal cortex regions that elicit representational change, all network measures stem from contrast: HI-I>LO-I. HI-I= trials solved with high accompanied insight; LO-I= trials solved with low accompanied insight**. Panel A.** Insight-related increased functional connectivity (*p-FDR*<.05) for *solution network* consisting of VOTC regions including amygdala and hippocampus. **Panel B.** Effect size for connection strength in *solution network* split by condition (*HI-I,LO-I)***. Panel C.** Direct functional connections (*p-FDR*<.05) for a subnetwork considering only connections between left and right pFusG, iLOC, amygdala and anterior and posterior hippocampus. **Panel D.** Network measures for *solution network* at different correlation thresholds indicating efficient information integration. *Upper*: T-values indicating differences in Degree or Average (Av.) Pathlength for *HI-I>LO-I* contrast in *solution network*. Dashed lines represent significance threshold for p=.05 (one-sided). *Lower:* Local efficiency in *solution network* compared to random network.

Hypothesis 4: VOTC-RC and hippocampal activity predict insight-related better subsequent memory

Having linked RC in VOTC and amygdala as well as hippocampal activity to the insight process, we examined whether the underlying neural correlates of both insight components also predicted better subsequent memory.

Insight-related RC predicts better subsequent memory.

First, we tested whether insight-related better memory is associated with our two consecutive measures of RC, i.e. 1) the changes of multivoxel activity patterns (i.e., decrease in multivoxel pattern similarity) and 2) an increase in representational strength of the object from pre to post solution in those regions that elicit RC (pFusG, iLOC). To test insight-related better memory for 1), we calculated a two-way interaction between insight (*HI-I* vs *LO-I*) and memory (remembered vs forgotten)

using a mixed effects model. To test insight-related better memory for 2), we calculated a three-way interaction between insight, memory and time (pre, post solution) using a mixed effects model. Both models included random subject and item intercepts controlling for run order, accuracy and solution time.

1) RC from pre to post solution: Multivoxel pattern similarity. There was an overall memory effect ($Chi^2(1)=10.44$, $p=.001$, $ß=.06$) controlling for insight as well as an insight-memory interaction effect ($Chi^2(1)=6.38$, $p=.011$, $ß=.05$) for pre to post solution activity pattern similarity for both RC ROIs (see Fig.5-A). The significant three-way interaction between insight, memory and ROI ($Chi^2(3)=50.44$, $p<.001$) indicated differences between both ROIs in insight-related better memory. Post Hoc analyses revealed that pFusG showed a main effect for memory ($Chi^2(1)=6.37$, $p=.012$) when controlling for insight but only a trend towards significance for an insight-memory interaction ($Chi^2(1)=3.61$, $p=.057$, $ß=.05$) in the hypothesised direction. In contrast, iLOC showed not only a memory effect ($Chi^2(1)=5.36$, $p=.021$) but also an insight-memory interaction effect ($Chi^2(1)=3.84$, $p=.049$). Note, repeating the analyses with the continuous *insight sum* measure yielded a significant insight-memory interaction ($Chi^2(1)=22.20$, $p<.001$, $ß=.09$) for pFusG while the results for iLOC did not change significantly.

2a) RC from pre to post solution: Representational strength - AlexNet. Using the AlexNet model to compute representational strength, there was no overall memory effect over both time points controlling for insight ($p>.46$) but there was a significant time*memory interaction ($Chi^2(1)=104.70$, $p<.001$, $ß=.24$) controlling for insight and a general insight*memory interaction ($Chi^2(1)=5.47$, $p=.019$, $ß=.04$) controlling for time. However, the three-way insight*memory*time interaction only showed a trend towards significance ($Chi^2(1)=2.94$, $p=.086$, $ß=.06$) (see Fig.5-B). Note, when repeating those analyses with the more sensitive (continuous) insight sum measure, the specific three-way interaction between insight*memory*time reached significance ($Chi^2(1)=4.21$, $p=.040$, $ß=.05$). Furthermore, the four-way interaction with ROI was not significant ($p>.65$) suggesting no evidence for a difference in this insight*memory*time relationship between the ROIs.

2b) RC from pre to post solution: Representational strength - Word2Vec. Those results were mostly replicated with the Word2Vec model (see Fig.5-D). There was an overall memory effect over both time points ($Chi^2(1)=10.67$, $p=.001$, $ß=.09$)

controlling for insight and a significant time*memory interaction ($Chi^2(1)=19.20$, $p<.001$, $ß=.10$) controlling for insight but no general insight*memory interaction ($Chi^2(1)=1.71$, $p=.19$, $ß=.02$) controlling for time. Importantly, the three-way insight*memory*time interaction showed a significant effect ($Chi^2(1)=6.76$, $p=.009$, $ß=.09$) (see Fig.5-B). Furthermore, the four-way interaction with ROI was not significant ($p>.47$) suggesting no evidence for a difference in this insight*memory*time relationship between the ROIs. Note, when repeating those analyses with the more sensitive (continuous) insight sum measure, none of these results changed significantly.

Insight-related anterior hippocampal activity predicts better subsequent memory.

We tested whether insight-related better memory is associated with amygdala and hippocampal activity. For this, we used the data from the single-trial analyses containing a beta estimate of univariate amygdala, anterior and posterior hippocampal activity (ROI) for the solution button press at each trial. We calculated an interaction between insight (*HI-I, LO-I*) and memory (*Remembered, Forgotten*) and examined if activity varied as a function of this interaction and the ROI controlling for response time, accuracy, run order (1-4) and random subject and item intercepts.

There was an a trend towards significance for an effect of memory ($Chi^2(1)=3.20$, $p=.07$, $ß=.04$) controlling for insight and a significant insight*memory interaction ($Chi^2(1)=5.17$, $p=.023$, $ß=.06$) for all three regions. However, the three-way interaction between insight*memory*ROI ($Chi^2(6)=34.65$, $p<.001$) indicates differences between the ROIs regarding this interaction. Post Hoc analyses revealed that amygdala showed only a trend towards significance for a main effect of memory ($Chi^2(1)=3.42$, $p=.064$, ß=.06) controlling for insight and no evidence for an insight*memory interaction effect ($p>.92$). Anterior hippocampus showed a trend towards significance for a main effect of memory ($Chi^2(1)= 3.21$, $p=.073$, ß=.06) when controlling for insight but an insight*memory interaction ($Chi^2(1)=4.79$, $p=.029$, ß=.09). Posterior hippocampus showed no memory effect (p>.83) controlling for insight except for an insight*memory interaction ($Chi^2(1)=4.50$, $p=.034$, ß=.09). Note

however, this interaction is driven by more activity for forgotten than remembered *LO-I* trials than more activity for remembered *HI-I* than *LO-I* trials (see Fig.5-B).

Note, for reasons of completeness, we explored whether FC and graph theoretical measures also predicted insight-related better memory. However, these results provided no consistent evidence (all analyses are detailed in the Supplement).

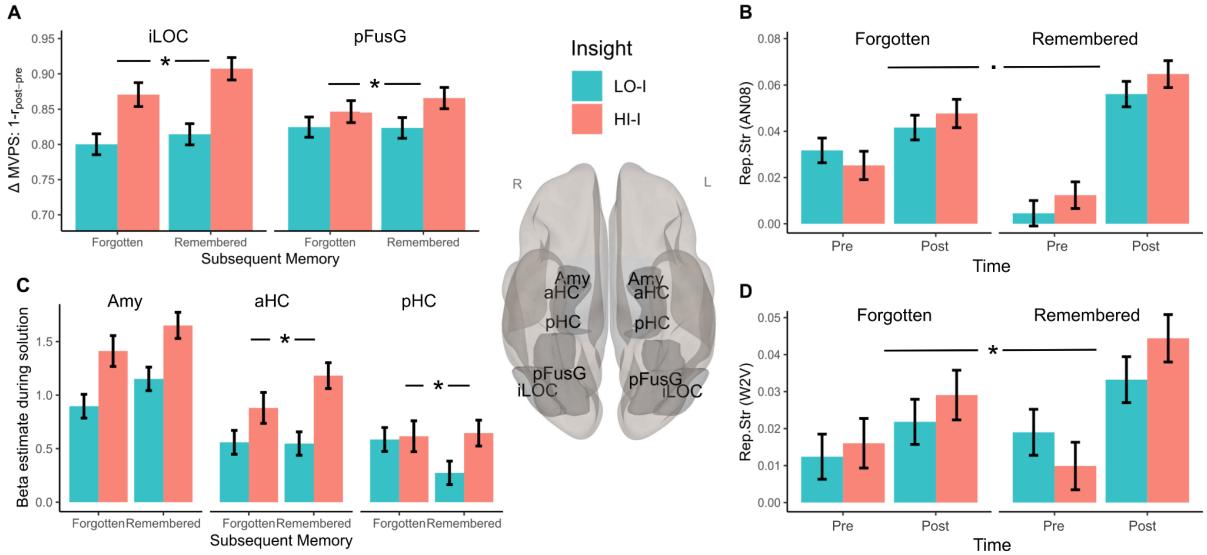

Figure 5. Representational change in VOTC and hippocampal activity during solution are associated with insight-related better subsequent memory.

Note. Values represent estimated marginal means ± (between subject) SEM. Amy = Amygdala; aHC= anterior hippocampus; pHC = posterior hippocampus; pFusG= posterior Fusiform Gyrus; iLOC= inferior Lateral Occipital Lobe. **Panel A:** Asterisk represents Insight*Memory interaction at p<.05. Change (Δ) in Multivoxel Pattern Similarity (MVPS). Change in MVPS = 1 minus the correlation between the post and pre solution multivoxel pattern in the respective ROI. **Panel B:** Dot represents trend towards significance for Insight*Memory*Time interaction at p<.10. Results display combined representational strength of solution object for iLOC & pFusG measured via a conceptual model created out of the penultimate layer of AlexNet. **Panel C:** Asterisk represents Insight*Memory interaction at p<.05. Bar represents mean activity at post solution divided by insight memory conditions. **Panel D:** Asterisk represents Insight*Memory*Time interaction at p<.05. Results display combined representational strength of solution object for iLOC & pFusG measured via a conceptual Word2Vec [W2V] model.

Discussion

Despite two decades of research, little is known about the neural mechanisms of insight, particularly regarding the underlying representational change and its subsequent impact on subsequent memory. To investigate these issues, we scanned participants with fMRI while they identified Mooney images, and tested their memory for the solutions one week later. Identifying Mooney images requires rechunking and integrating their elements into a coherent object (RC, cognitive component of insight), which often elicits positive emotions and a feeling of suddenness and certainty (evaluative component of insight). We found that insight enhanced (1) RC in ventral occipito-temporal cortex (VOTC), (2) activity in amygdala and hippocampus (3) efficient information integration within a network of solution-processing brain regions, and (4) subsequent memory involving VOTC RC and hippocampal activity. Taken together, these findings represent the first direct evidence of an integrated neural mechanism for insight and its impact on memory. The four findings are discussed in separate sections below.

Cognitive component of insight: Representational change in VOTC

This study provided the first multivariate evidence of insight-related RC. Using a visual Mooney identification task, we found insight-related RC in middle VOTC regions (pFusG and iLOC), which are areas that process complex visual features (Bussey et al., 2005; Clarke & Tyler, 2014). From pre to post solution, the conceptual representation of the target objects increased in pFusG and iLOC, suggesting a change from representations coding for meaningless black and white Mooney image blobs to representations coding for meaningful objects. We did not find RC in anterior VOTC (aFusG and aITG), which processes more abstract concepts and semantics (Bussey et al., 2005; Clarke & Tyler, 2014), which could reflect the typically lower signal-to-noise ratios (SNR) in the MRI of this region (Embleton et al., 2010) or the visual nature of the Mooney identification task. Future research could examine if insights tasks that require more abstract semantic processing (e.g., regrouping conceptual features in verbal insight tasks) yields RC in anterior VOTC regions.

Evaluative component of insight: amygdala and hippocampal activity

Our second finding was that insight boosted activity in the amygdala and hippocampus (Jung-Beeman et al., 2004; Shen et al., 2017; Tik et al., 2018). In previous insight studies, amygdala activity has been attributed to positive emotions (Kizilirmak, Thuerich, et al., 2016; Ludmer et al., 2011), aligning with the broader role of this region in emotional arousal (Rouhani et al., 2023), and hippocampal activity to sudden solution awareness (Cabeza et al., 2020; Kizilirmak & Becker, 2024), in line with its role as a novelty or “mismatch” detector (Kumaran & Maguire, 2006; Nyberg, 2005). This function has been linked to a prediction error, when there is a mismatch between a prior expectation and the outcome, which leads to the updating of these expectations and memory storage (Bein et al., 2020; Duncan et al., 2009; Sinclair et al., 2021). In our task, as a result of RC in visual cortex, the originally meaningless stimulus becomes suddenly meaningful, yielding a prediction error, which triggers hippocampal activity and the storage of the salient event.

Insight-related integration of the solution network

Our third finding was that insight boosted functional connectivity within the *solution network* (VOTC, amygdala, hippocampus). As shown by functional connectivity and graph measures, insight was associated with greater functional integration and segregation of the *solution network*. Those results suggest that during insight, solution-relevant information was more efficiently processed, as has been proposed before for the creative process (Beaty et al., 2018; Mednick, 1962) and particularly for insight (Bieth et al., 2021; Luchini et al., 2023; Schilling, 2005). This is the first direct neuroscientific evidence for these ideas in the context of insight. Importantly, during insight, both VOTC-RC regions, bilateral pFusG and iLOC, increased their network communication (increased degree and decreased average path length). This result suggests that these regions flexibly increased their local connectedness and ability to combine visual information, which could allow them to more efficiently integrate the reorganised internal representations within the ventral stream object processing pathway, including the hippocampus and the amygdala.

Note however, functional connectivity, as measured here, cannot determine the precise timing nor the direction of the effects within the *solution network*. In a feedforward direction, RC in VOTC could lead to evaluation responses in hippocampus and amygdala, whereas in a feedback direction, hippocampal and amygdala responses could guide or enhance RC in VOTC areas. For example, a Mooney image could trigger the automatic memory retrieval of a cat image, which then guides the rechunking and integration of the Mooney image (RC) in VOTC (Hindy et al., 2016). Moreover, emotions triggered by the sudden solution may trigger the amygdala to enhance solution-relevant sensory representations thanks to its ability to project to all levels of the ventral stream (Freese & Amaral, 2005; Gschwind et al., 2012; Janak & Tye, 2015).

Insight and Memory

Finally, this study provides first evidence of the neural mechanism of insight's enhancing effects on memory. Specifically, RC in VOTC and anterior hippocampal activity predicted subsequent memory effects to a greater extent for high- than low-insight trials. The amygdala showed a trend towards predicting memory but this effect did not differ as a function of insight. A couple of studies linked amygdala activity to improved subsequent memory during problem solving (Kizilirmak et al., 2019; Ludmer et al., 2011) but without considering the critical insight*memory interaction (Nieuwenhuis et al., 2011).

The VOTC-RC impact on subsequent memory is consistent with multivariate fMRI studies showing that representational strength during the encoding of everyday objects predicts later memory of these objects (Bone et al., 2020; Davis et al., 2021). Whereas in these studies subsequent memory was predicted by absolute representational strength, in our study, later memory was specifically predicted by the insight-related *change* in strength from pre- to post-solution. This is further consistent with a “prediction error”-memory mechanism, in which memories are strengthened proportional to the extent to which the outcome (e.g., suddenly identifying a meaningful object) deviates from prior expectations (e.g., solution object was unexpected) (Ergo et al., 2020; Greve et al., 2017; Rouhani et al., 2023).

The observation that hippocampal activity can predict insight-related better memory aligns with its pivotal function in converting transient percepts into durable memory traces (Moscovitch et al., 2016; Schacter, 2022). Note, consistent with previous accounts, we proposed a double role for hippocampus, as a novelty or ”mismatch” detector and as an encoder of this novel event (Axmacher et al., 2010; Kizilirmak & Becker, 2024; Sinclair et al., 2021). Hence, in the context of insight, the hippocampus may detect novel, unexpected internal events (e.g., sudden conceptual updates of solution object in VOTC) by eliciting a prediction error signal (surprise) and subsequently encodes these events, proportional to the magnitude of this conceptual update. This dual function is further supported by hippocampal activity showing an insight*memory interaction *and* a main effect for insight, as well as by increased functional connectivity between VOTC-RC regions (pFusG, iLOC) and the hippocampus. Note however, no FC evidence predicts subsequent memory, possibly due to limited power or independent contributions of both regions to insight-related memory enhancement.

Conclusion

This work enhances our mechanistic understanding of the core components of insight, a central aspect of creativity, and its connection to rapid memory formation. We presented consistent neural evidence indicating that insight in problem solving likely reflects a sudden conceptual update of solution representations efficiently integrated into a coherent whole. This update generates solution awareness alongside an emotional and suddenness/novelty response in the amygdala and hippocampus, boosting solution encoding and subsequent memory. These findings illuminate rapid memory formation mechanisms and possess educational implications, highlighting the importance of fostering insight-driven learning environments as effective avenues for problem-solving and knowledge retention.

References

Andersson, J. L., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging.

Neuroimage, *20*(2), 870–888.

Aru, J., Drüke, M., Pikamäe, J., & Larkum, M. E. (2023). Mental navigation and the neural mechanisms of insight. *Trends in Neurosciences*, *46*(2), 100–109.

Ash, I. K., Jee, B. D., & Wiley, J. (2012). Investigating insight as sudden learning. *The Journal of Problem Solving*, *4*(2).

Ash, I. K., & Wiley, J. (2008). Hindsight bias in insight and mathematical problem solving: Evidence of different reconstruction mechanisms for metacognitive versus situational judgments. *Memory & Cognition*, *36*(4), 822–837.

Ashburner, J. (2012). SPM: a history. *Neuroimage*, *62*(2), 791–800.

Auble, P. M., Franks, J. J., Soraci, S. A., Soraci, S. A., & Soraci, S. A. (1979). Effort toward comprehension: Elaboration or “aha”? *Memory & Cognition*, *7*(6), 426–434.

Avants, B., Epstein, C., Grossman, M., & Gee, J. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004

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

Bankson, B. B., Hebart, M. N., Groen, I. I., & Baker, C. I. (2018). The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *NeuroImage*, *178*, 172–182.

Beaty, R. E., Kenett, Y. N., Christensen, A. P., Rosenberg, M. D., Benedek, M., Chen, Q., Fink, A., Qiu, J., Kwapil, T. R., & Kane, M. J. (2018). Robust prediction of individual creative ability from brain functional connectivity. *Proceedings of the National Academy of Sciences*, *115*(5), 1087–1092.

Becker, M., Cabeza, R., & Kizilirmak, J. M. (2024). A cognitive neuroscience perspective on insight as a memory process: Searching for the solution. In *The Routledge*

International Handbook of Creative Cognition (1st ed., Vol. 1, p. 20). Routledge. https://osf.io/zv4dk

Becker, M., Davis, S., & Cabeza, R. (2022). Between automatic and control processes: How relationships between problem elements interact to facilitate or impede insight. *Memory & Cognition*, *50*(8), 1719–1734.

Becker, M., Kühn, S., & Sommer, T. (2021). Verbal insight revisited—Dissociable neurocognitive processes underlying solutions accompanied by an AHA! experience with and without prior restructuring. *Journal of Cognitive Psychology*, *33*(6–7), 659–684.

Becker, M., Sommer, T., & Kühn, S. (2020). Verbal insight revisited: FMRI evidence for early processing in bilateral insulae for solutions with AHA! experience shortly after trial onset. *Human Brain Mapping*, *41*(1), 30–45.

Becker, M., Wiedemann, G., & Kühn, S. (2020). Quantifying insightful problem solving: A modified compound remote associates paradigm using lexical priming to parametrically modulate different sources of task difficulty. *Psychological Research*, *84*, 528–545.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Bein, O., Duncan, K., & Davachi, L. (2020). Mnemonic prediction errors bias hippocampal states. *Nature Communications*, *11*(1), 3451.

Bieth, T., Kenett, Y., Ovando-Tellez, M., Lopez-Persem, A., Lacaux, C., Oudiette, D., & Volle, E. (2021). *Dynamic changes in semantic memory structure support successful problem-solving*. https://doi.org/10.31234/osf.io/38b4w

Brett, M., Anton, J.-L., Valabregue, R., & Poline, J.-B. (2002). *Region of interest analysis using an SPM toolbox*. *16*(2), 497.

Bussey, T. J., Saksida, L. M., & Murray, E. A. (2005). The perceptual-mnemonic/feature conjunction model of perirhinal cortex function. *The Quarterly Journal of*

Experimental Psychology Section B, *58*(3–4), 269–282.

Cabeza, R., Becker, M., & Davis, S. W. (2020). Are the hippocampus and its network necessary for creativity? *Proceedings of the National Academy of Sciences*, *117*(25), 13870–13872. https://doi.org/10.1073/pnas.2008601117

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963.

Chai, X. J., Castañón, A. N., Öngür, D., & Whitfield-Gabrieli, S. (2012). Anticorrelations in resting state networks without global signal regression. *Neuroimage*, *59*(2), 1420–1428.

Clarke, A., & Tyler, L. K. (2014). Object-specific semantic coding in human perirhinal cortex. *Journal of Neuroscience*, *34*(14), 4766–4775.

Cox, R. W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173. https://doi.org/10.1006/cbmr.1996.0014

Crawley, M. J. (2012). *The R book*. John Wiley & Sons.

Danek, A. H., Fraps, T., Von Müller, A., Grothe, B., & Öllinger, M. (2013). Aha! Experiences leave a mark: Facilitated recall of insight solutions. *Psychological Research*, *77*, 659–669.

Danek, A. H., & Salvi, C. (2020). Moment of truth: Why Aha! Experiences are correct. *The Journal of Creative Behavior*, *54*(2), 484–486.

Danek, A. H., & Wiley, J. (2017). What about false insights? Deconstructing the Aha! Experience along its multiple dimensions for correct and incorrect solutions separately. *Frontiers in Psychology*, *7*, 2077.

Danek, A. H., & Wiley, J. (2020). What causes the insight memory advantage? *Cognition*, *205*, 104411.

Danek, A. H., Williams, J., & Wiley, J. (2020). Closing the gap: Connecting sudden

representational change to the subjective Aha! Experience in insightful problem solving. *Psychological Research*, *84*, 111–119.

Davis, S. W., Geib, B. R., Wing, E. A., Wang, W.-C., Hovhannisyan, M., Monge, Z. A., & Cabeza, R. (2021). Visual and semantic representations predict subsequent memory in perceptual and conceptual memory tests. *Cerebral Cortex*, *31*(2), 974–992.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., & Hyman, B. T. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968–980.

Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific Reports*, *8*(1), 10636.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Dietrich, A., & Kanso, R. (2010). A review of EEG, ERP, and neuroimaging studies of creativity and insight. *Psychological Bulletin*, *136*(5), 822.

Duncan, K., Curtis, C., & Davachi, L. (2009). Distinct memory signatures in the hippocampus: Intentional states distinguish match and mismatch enhancement signals. *Journal of Neuroscience*, *29*(1), 131–139.

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annu. Rev. Neurosci.*, *30*, 123–152.

Embleton, K. V., Haroon, H. A., Morris, D. M., Ralph, M. A. L., & Parker, G. J. (2010). Distortion correction for diffusion-weighted MRI tractography and fMRI in the temporal lobes. *Human Brain Mapping*, *31*(10), 1570–1587.

Ergo, K., De Loof, E., & Verguts, T. (2020). Reward prediction error and declarative memory. *Trends in Cognitive Sciences*, *24*(5), 388–397.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., & Snyder, M. (2019). fMRIPrep: A robust

preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116.

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, S102. https://doi.org/10.1016/S1053-8119(09)70884-5

Freese, J. L., & Amaral, D. G. (2005). The organization of projections from the amygdala to visual cortical areas TE and V1 in the macaque monkey. *The Journal of Comparative Neurology*, *486*(4), 295–317. https://doi.org/10.1002/cne.20520

Goldberg, Y., & Levy, O. (2014). word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *ArXiv Preprint ArXiv:1402.3722*.

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

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10–11), 1409–1422.

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife*, *7*, e32962.

Gruber, M. J., Gelman, B. D., & Ranganath, C. (2014). States of Curiosity Modulate Hippocampus-Dependent Learning via the Dopaminergic Circuit. *Neuron*, *84*(2), 486–496. https://doi.org/10.1016/j.neuron.2014.08.060

Gschwind, M., Pourtois, G., Schwartz, S., Van De Ville, D., & Vuilleumier, P. (2012). White-Matter Connectivity between Face-Responsive Regions in the Human Brain. *Cerebral Cortex*, *22*(7), 1564–1576. https://doi.org/10.1093/cercor/bhr226

Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nature Neuroscience*, *19*(5),

665–667.

Imamoglu, F., Koch, C., & Haynes, J.-D. (2013). MoonBase: Generating a database of two-tone Mooney images. *Journal of Vision*, *13*(9), 50–50.

Janak, P. H., & Tye, K. M. (2015). From circuits to behaviour in the amygdala. *Nature*, *517*(7534), 284–292. https://doi.org/10.1038/nature14188

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., Greenblatt, R., Reber, P. J., & Kounios, J. (2004). Neural activity when people solve verbal problems with insight. *PLoS Biology*, *2*(4), e97.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–4311.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915.

Kizilirmak, J. M., & Becker, M. (2022). *A cognitive neuroscience perspective on insight as a memory process: Encoding the solution*.

Kizilirmak, J. M., & Becker, M. (2024). A cognitive neuroscience perspective on insight as a memory process: Encoding the solution. In *The Routledge International Handbook of Creative Cognition* (p. 17). Routledge. https://doi.org/10.31234/osf.io/bevjm

Kizilirmak, J. M., Galvao Gomes da Silva, J., Imamoglu, F., & Richardson-Klavehn, A. (2016). Generation and the subjective feeling of “aha!” are independently related to learning from insight. *Psychological Research*, *80*, 1059–1074.

Kizilirmak, J. M., Schott, B. H., Thuerich, H., Sweeney-Reed, C. M., Richter, A., Folta-Schoofs, K., & Richardson-Klavehn, A. (2019). Learning of novel semantic relationships via sudden comprehension is associated with a

hippocampus-independent network. *Consciousness and Cognition*, *69*, 113–132.

Kizilirmak, J. M., Thuerich, H., Folta-Schoofs, K., Schott, B. H., & Richardson-Klavehn, A. (2016). Neural correlates of learning from induced insight: A case for reward-based episodic encoding. *Frontiers in Psychology*, *7*, 1693.

Kleiner, M., Brainard, D., & Pelli, D. (2007). *What's new in Psychtoolbox-3?*

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1534–1555. https://doi.org/10.1037/0278-7393.25.6.1534

Kounios, J., & Beeman, M. (2014). The Cognitive Neuroscience of Insight. *Annual Review of Psychology*, *65*(1), 71–93. https://doi.org/10.1146/annurev-psych-010213-115154

Kounios, J., & Beeman, M. (2015). *The Eureka factor: Creative insights and the brain*. Random House.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 4.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Kumaran, D., & Maguire, E. A. (2006). An unexpected sequence of events: Mismatch detection in the human hippocampus. *PLoS Biology*, *4*(12), e424.

Kumaran, D., & Maguire, E. A. (2007). Match–mismatch processes underlie human hippocampal responses to associative novelty. *Journal of Neuroscience*, *27*(32),

8517–8524.

Luchini, S., Kenett, Y. N., Zeitlen, D. C., Christensen, A. P., Ellis, D. M., Brewer, G. A., & Beaty, R. E. (2023). Convergent thinking and insight problem solving relate to semantic memory network structure. *Thinking Skills and Creativity*, *48*, 101277.

Ludmer, R., Dudai, Y., & Rubin, N. (2011). Uncovering camouflage: Amygdala activation predicts long-term memory of induced perceptual insight. *Neuron*, *69*(5), 1002–1014.

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Moscovitch, M., Cabeza, R., Winocur, G., & Nadel, L. (2016). Episodic memory and beyond: The hippocampus and neocortex in transformation. *Annual Review of Psychology*, *67*, 105–134.

Mumford, J. A., Poline, J.-B., & Poldrack, R. A. (2015). Orthogonalization of regressors in fMRI models. *PloS One*, *10*(4), e0126255.

Nieto-Castanon, A. (2020). *Handbook of functional connectivity Magnetic Resonance Imaging methods in CONN*. Hilbert Press. https://doi.org/10.56441/hilbertpress.2207.6598

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430.

Nyberg, L. (2005). Any novelty in hippocampal formation and memory? *Current Opinion in Neurology*, *18*(4), 424–428.

Ohlsson, S. (1984). Restructuring revisited: II. An information processing theory of

restructuring and insight. *Scandinavian Journal of Psychology*, *25*(2), 117–129.

Perez, D. C., Dworetsky, A., Braga, R. M., Beeman, M., & Gratton, C. (2023). Hemispheric Asymmetries of Individual Differences in Functional Connectivity. *Journal of Cognitive Neuroscience*, *35*(2), 200–225.

Poppenk, J., Evensmoen, H. R., Moscovitch, M., & Nadel, L. (2013). Long-axis specialization of the human hippocampus. *Trends in Cognitive Sciences*, *17*(5), 230–240. https://doi.org/10.1016/j.tics.2013.03.005

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048

Rouhani, N., Niv, Y., Frank, M. J., & Schwabe, L. (2023). Multiple routes to enhanced memory for emotionally relevant events. *Trends in Cognitive Sciences*.

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *Neuroimage*, *52*(3), 1059–1069.

Schacter, D. L. (2022). Constructive memory: Past and future. *Dialogues in Clinical Neuroscience*.

Schiffer, A.-M., Ahlheim, C., Wurm, M. F., & Schubotz, R. I. (2012). Surprised at all the entropy: Hippocampal, caudate and midbrain contributions to learning from prediction errors. *PloS One*, *7*(5), e36445.

Schilling, M. A. (2005). A" small-world" network model of cognitive insight. *Creativity Research Journal*, *17*(2–3), 131–154.

Schöne, M., Seidenbecher, S., Kaufmann, J., Antonucci, L. A., Frodl, T., Koutsouleris, N., Schiltz, K., & Bogerts, B. (2022). Appetitive aggression is associated with lateralized activation in nucleus accumbens. *Psychiatry Research: Neuroimaging*, *319*, 111425.

Shen, W., Yuan, Y., Liu, C., & Luo, J. (2017). The roles of the temporal lobe in creative insight: An integrated review. *Thinking & Reasoning*, *23*(4), 321–375.

Sinclair, A. H., Manalili, G. M., Brunec, I. K., Adcock, R. A., & Barense, M. D. (2021). Prediction errors disrupt hippocampal representations and update episodic

memories. *Proceedings of the National Academy of Sciences*, *118*(51), e2117625118.

Tik, M., Sladky, R., Luft, C. D. B., Willinger, D., Hoffmann, A., Banissy, M. J., Bhattacharya, J., & Windischberger, C. (2018). Ultra-high-field fMRI insights on insight: Neural correlates of the Aha!-moment. *Human Brain Mapping*, *39*(8), 3241–3252.

Tukey, J. (1953). Multiple comparisons. *Journal of the American Statistical Association*, *48*(263), 624–625.

Tustison, N. J., Avants, B. B., Cook, P. A., Yuanjie Zheng, Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

Webb, M. E., Little, D. R., & Cropper, S. J. (2016). Insight Is Not in the Problem: Investigating Insight in Problem Solving across Task Types. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01424

Weiner, K. S., & Zilles, K. (2016). The anatomical and functional specialization of the fusiform gyrus. *Neuropsychologia*, *83*, 48–62. https://doi.org/10.1016/j.neuropsychologia.2015.06.033

Whitfield-Gabrieli, S., & Nieto-Castanon, A. (2012). Conn: A functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connectivity*, *2*(3), 125–141.

Zalesky, A., Fornito, A., & Bullmore, E. T. (2010). Network-based statistic: Identifying differences in brain networks. *Neuroimage*, *53*(4), 1197–1207.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, 818–833.

Zhang, W., Wang, J., Fan, L., Zhang, Y., Fox, P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2016). Functional organization of the fusiform gyrus revealed with connectivity profiles. *Human Brain Mapping*, *37*(8), 3003–3016.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE*

Transactions on Medical Imaging, *20*(1), 45–57. https://doi.org/10.1109/42.906424

STAR METHODS

Key Resources Table

Reagent or Resource	Source	Identifier
Deposited Data		
HO atlas (CONN)	Desikan et al., 2006	https://cma.mgh.harvard.edu/
Mooney images	in house, Imamoglu et al., 2013	https://github.com/MaxiBecker/Insight_Memory_Effect
Behavioral and fMRI data	in house	https://github.com/MaxiBecker/Insight_Memory_Effect
Software and Algorithms		
fMRIprep	Esteban et al., 2019	https://fmriprep.org
Psychtoolbox	Kleiner, Brainard & Pelli, 2007	http://psychtoolbox.org/
Matlab 2016a	Higham & Higham, 2016	https://au.mathworks.com/
SPM 12	Ashburner et al., 2014	https://www.fil.ion.ucl.ac.uk/spm/software/spm12/
CONN	Whitfield-Gabrieli & Nieto-Castanon, 2012	https://web.conn-toolbox.org/
RSA toolbox	Nili et al., 2014	https://git.fmrib.ox.ac.uk/hnili/rsa

Experimental model and subject details

The full sample consisted of 38 participants [age (in years): range = 20 – 34, 23 females: M= 26.6; 15 males: M= 25.7] recruited via an online student platform in Berlin. Inclusion criteria were German as mother language, normal or corrected to normal vision, between 18 – 35 years, no prior neurological or psychiatric diseases and MRI compatibility. Consent was obtained for all participants prior to study begin. All participants received monetary compensation according to their time on task, and

the local ethics committee of the Humboldt University Berlin approved the study. We excluded 7 subjects from data analysis due to technical issues at the scanner (N=1), excessive head movement in the scanner (N=3), pathological findings in brain anatomy (N=2) andtoo low performance (0% correct recall) in the subsequent memory test (N=1) resulting in a final sample of N=31 [age (in year): range = 19 – 33, 20 females: M= 25.13; 11 males: M= 26.27].

Method details

Stimulus Material

To induce perceptual insight in the MR scanner, we used high-contrast, black and white (henceforth Mooney) images, which have been successfully used to elicit insight (see Fig.1-A; (Imamoglu et al., 2013; Kizilirmak, Galvao Gomes da Silva, et al., 2016; Ludmer et al., 2011). Those images depict a clearly identifiable object but, due to their high-contrast thresholding, they are usually hard to identify at first, with response times taking several seconds (Imamoglu et al., 2013). Visually rechunking the abstracted shapes can lead to sudden object identification, often involving an insight (Kizilirmak, Galvao Gomes da Silva, et al., 2016). Examples of Mooney images used in this study and their real-world pictures are shown in Fig.S1 (Supplementary Material).

The materials consisted of 120 Mooney images used for the problem solving task in the scanner and later included as "old" items in the memory test plus 60 additional Mooney images included in the memory test as "new" (distractor) items (180 in total). A fraction of these images (N=34) were selected from Imamoglu and colleagues (2013) and the rest were created using a custom software (freely available at https://github.com/MaxiBecker/Insight_Memory_Effect.) The final set of images was selected across several pilot studies to ensure a solution accuracy of at least 30%, a response time in the 2-13 sec range, and sufficiently high rate and variation of self-rated insight. The final set had a mean accuracy of 56%(*SD*=18.2%), mean response time 6.23sec (*SD*=3.11sec) and probability of self-rated insight of 62% (*SD*=17%, range [20%-100%]).

Behavioural data acquisition

Assessing insight. Originally, insight was only binarized as present or absent (Jung-Beeman et al., 2004; for review Kounios & Beeman, 2014). However, recent studies showed that the overall subjective insight (AHA) experience is more continuous and can be decomposed into three different dimensions: (1) positive emotional response or internal reward upon finding the solution, (2) experienced suddenness of the solution and (3) certainty about the solution (Danek & Wiley, 2017, 2020). To have better control over these three dimensions, we assessed them separately, and then combined them into a compound insight measure (see section: Mooney image paradigm) using the classical binary and a continuous approaches. In the "*insight median split"* (binary) approach, we added up the ratings of positive emotion, suddenness, and certainty (1-4 each, 3-12 for the total), and used a median split of this measure to distinguish high- and low-insight trials. The advantage of this approach is that it yields a roughly equal number of trials for high- and low-insight, but its disadvantage is it is not the most sensitive insight measure as it is binarized. Thus, in the "*insight sum"* (continuous) approach, we added up the ratings of positive emotion, suddenness and certainty and used this as a continuous measure for insight. Both measures produced similar results except for where it is specifically stated in the results section.

Mooney image paradigm. Participants had two MRI sessions on two consecutive days and in each session they completed two runs with 30 trials each. The presented images per run and session were counterbalanced between participants to avoid order effects. Before entering the MRI scanner, participants were instructed and received three test trials to assure they understood the task. The paradigm was presented on a black screen (resolution 1280*960 pixel) using Matlab (2016a) and Psychtoolbox-3 (v3.0.17; Kleiner et al., 2007). The black/white Mooney images had a size of 800x800 pixels. During each trial a Mooney image was presented for 10 seconds after a jittered fixation cross (mean 3sec +/- 1sec) (see Fig.1-B). The subjects had a four-button response box in their right hand and were instructed to press the "solution" button (right index finger) as soon as they found a solution even though the Mooney image would stay on the screen for the remaining time. The Mooney images were presented as an invariant 10 second block to be able to compare time series of equal length for subsequent functional connectivity analyses. If the subjects did not press the "solution" button during the 10 second stimulus

presentation, a new trial would start, and the subjects were instructed to press a “no solution” button (middle finger) as soon as they saw the fixation cross to keep the amount of button presses constant for each trial. If the participants did press the “solution” button during stimulus presentation, they would be first presented with a jittered fixation cross (mean 3.5sec +/- 1.5sec), followed by four questions regarding the solution and their insight experience: First, four out of 13 broad categories, e.g. “mammal”, “human being”, “body & health” or “objects in the house” (see Supplementary Material) were shown and the participants had to indicate the category of their solution. On average, they needed 3.13sec (*SD* = 2.5sec) to make a response. This way of assessing the solution was used to test whether subjects correctly identified the Mooney object, because they could not type their response in the scanner. Subsequently, they were asked about their suddenness (On a scale from 1 – 4: Did the solution come to you in a sudden or more gradual manner?), emotional response (On a scale from 1 – 4: How strong was your positive emotional response upon solution?”) and certainty (“On a scale from 1 – 4: How certain are you that the solution is correct?”) regarding the solution on a 4-point Likert scale. There was no time limit for these 4 questions, and subsequently, the new trial would start. Note, the 30 Mooney images per run were presented interleaved together with 30 anagram riddles which are not part of the current study.

Subsequent memory test. Five days after the second fMRI session (4.62 days [*SD*=0.49]), participants completed the subsequent memory test online at home. All 120 Mooney images they had seen in the scanner were randomly presented mixed with the 60 new Mooney images they had not seen before. Participants were explicitly instructed not to identify the Mooney objects again but only rely on their memory. In each trial the respective Mooney image was shown after fixation cross (one sec) for maximally eight seconds. During this time, subjects had to decide whether they had seen this Mooney image before during the MRI sessions, i.e. whether they recognized the high-contrast black and white picture. In particular, they were asked to rate their confidence on a 5-point scale (definitely old, probably old, don’t know, probably new, definitely new). If they responded ‘don’t know’, ‘probably new’ or ‘definitely new’, the next trial would start. If they responded old (definitely old, probably old), they were asked whether they had identified this Mooney object before in the scanner again using a 5-point confidence scale without a time limit (definitely

solved, probably solved, don't know, probably not solved, definitely not solved). If they responded 'solved' (definitely solved or probably solved), they were finally asked to enter the name of the Mooney object (no time limit). Subsequently the next trial would start.

MRI data acquisition

Task-based functional (henceforth fMRI) and structural images were collected on a Siemens Magnetom Prisma 3T scanner (Erlangen, Germany) and a standard 64-channel head coil was used. All sequences were adapted according to the Human Connective Project (Harms et al., 2018). Multiband functional images were collected using a T2*-weighted echo planar imaging (EPI) sequence sensitive to blood oxygen level dependent (BOLD) contrast (TR=0.8ms; TE=37ms, voxel size=2.0mm³, flip angle=52°, FoV=208mm, acquisition matrix = 208×208×144; 72 slices). The structural images were obtained using a three-dimensional T1-weighted magnetization prepared gradient-echo sequence (MPRAGE) (TR = 2500ms; TE = 2.22ms; TI = 1000ms, acquisition matrix =240×256×167, FoV=256mm, flip angle = 8˚; 0.8 mm³ voxel size, 208 slices). Additionally a spin echo field map was acquired to account for the B0 inhomogeneities (TR=8000ms; TE=66ms; flip angle =90°, 72 slices; FoV=208mm; 2mm³ voxel size, acquisition matrix = 208x208x144).

Quantification and statistical analysis

Behavioural data analysis

Subsequent Memory Measure. We were more interested in recall (naming the correct Mooney object) than just recognition (remembering to have seen the black/white Mooney dots) because in the context of problem solving the more relevant memory measure is the ability to recall the content of the solution than just to recognize the problem, i.e. Mooney object. To reduce the likelihood that subjects just identify the Mooney image again instead of remembering the depicted object, we used all three questions as a compound memory measure. Therefore, the identification of the Mooney object was considered as *remembered*, when it was identified in the MR scanner *and* 1) the Mooney image was correctly recognized as

seen in the scanner, 2) the Mooney image was correctly recognized of having been identified in the scanner and finally 3) the hidden Mooney object was correctly named in the Subsequent Memory Test (see section *Subsequent Memory Test*).

Insight-related better memory for all fMRI analyses was measured as an interaction between insight (HI-I vs LO-I) and memory (Remembered vs. Forgotten). Forgotten trials were defined as trials that were subsequently not or not correctly remembered. On average, 53% (*SD*=17%) were not remembered and 47% (*SD*=17%) were incorrectly remembered from all *Forgotten* trials. The false alarm rate was calculated as the percentage of incorrectly recognized Mooney images of all new images (n=60) per subject.

All behavioural and fMRI 2nd level analyses (except for the exploratory whole brain analysis and functional connectivity analysis) were conducted in R (v4.2.0, Crawley, 2012). To control for random variance of subject and the different items, we calculated random mixed effects models using the lmer-function (v1.1-29). Posthoc marginal mean differences between conditions were estimated using the emmeans package (v.1.7.5) adjusting for multiple comparison via the Tukey HSD method (Tukey, 1953) where it applies. P-values for the nested mixed effects models were calculated via likelihood-ratio tests.

Insight Memory Effect. Two binomial mixed effect models were estimated with correct subsequent memory (binary) as dependent measure and a factor that differentiated between three conditions – not solved trials, solved trials with high (*HI-I*) as well as low insight (*LO-I*) – as independent measure controlling for the run order (1-4) and including two random intercepts for subject and item (see equation below). The baseline model excluded the independent measure. To exclude the fact that the difference between high and low insight trials on memory are just because of differences in response time, we repeated the analyses described further above but with response time included as covariate in both models (note, unsolved trials do not have response time and are therefore not part of the latter analysis).

I. Subsequent memory ~ run + (1| ID) + (1 | item) + Ɛ

II. Subsequent memory ~ run + insight + (1| ID) + (1 | item) + Ɛ

fMRI data preprocessing

Functional as well as structural images were preprocessed with fMRIPrep version 20.2.5 using default parameters (Esteban et al., 2019). Each T1w (T1-weighted) volume was corrected for INU (intensity non-uniformity) using N4BiasFieldCorrection v2.1.0 (Tustison et al., 2010) and skull-stripped using antsBrainExtraction.sh v2.1.0 (using the OASIS template). Spatial normalisation to the ICBM 152 Nonlinear Asymmetrical template version 2009c [Fonov et al., 2009; RRID:SCR_008796] was performed through nonlinear registration with the antsRegistration tool of ANTs v2.1.0 [(Avants et al., 2008), RRID:SCR_004757], using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and grey-matter (GM) was performed on the brain-extracted T1w using fast [(Y. Zhang et al., 2001), FSL v5.0.9, RRID:SCR_002823].

Functional data was motion corrected using MCFLIRT (FSL v5.0.9, (Jenkinson et al., 2002). Distortion correction was performed using an implementation of the TOPUP technique (Andersson et al., 2003) using 3dQwarp [AFNI v16.2.07; (Cox, 1996)]. This was followed by co-registration to the corresponding T1w using boundary-based registration (D. N. Greve & Fischl, 2009) with six degrees of freedom, using bbregister (FreeSurfer v6.0.1). Motion correcting transformations, field distortion correcting warp, BOLD-to-T1w transformation and T1w-to-template (MNI) warp were concatenated and applied in a single step using antsApplyTransforms (ANTs v2.1.0) using Lanczos interpolation.

Physiological noise regressors were extracted by applying anatomical CompCor (aCompCor) (Behzadi et al., 2007). Here, six components were calculated within the intersection of the subcortical mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run. Framewise displacement (Power et al., 2014) was calculated for each functional run using the implementation of Nipype.

To increase the signal to noise ratio, the preprocessed functional data in standard as well as in native space were smoothed with an 2mm full width half maximum Gaussian kernel (*FWHM*) for the ROI-based analysis and with an 8mm FWHM for

the exploratory whole-brain analysis using SPM12 (Welcome Department of Cognitive Neurology, London, UK).

Univariate fMRI data analysis

The amygdala and hippocampus have been previously suggested to be involved in the evaluative component of insight (Aru et al., 2023; Cabeza et al., 2020; Kizilirmak, Thuerich, et al., 2016; Tik et al., 2018). Therefore, we used those brain areas as ROIs for the univariate analysis.

First-level analysis. To estimate the effect of insight on amygdala and hippocampal activity, we contrasted brain activity during *HI-I* and *LO-I* trials conducted in the framework of general linear models (GLMs) following a mass univariate approach as implemented in SPM using standard parameter settings. Regressors were created by convolving the onsets of the respective conditions with the canonical hemodynamic response function (HRF) and their first and second temporal derivative. The time series were corrected for baseline drifts by applying a high-pass filter (128 s) and for serial dependency by an AR(1) autocorrelation model.

For the first-level analysis, a total of 22 separate regressors was created (the numbers in brackets [1–22] indicate the number of the respective regressors including first and second temporal derivative): We modelled the solution button presses for the Mooney images in two different conditions: *HI-I* [1:3], *LO-I* [4:6]. To reduce variance from the implicit baseline, we added two nuisance regressors; one for all solution button presses of the anagram trials [7:9] (which are not reported here) and another one for all remaining button presses [10:12] (related to the insight ratings and selecting a solution category). All those events were modelled for one second starting from solution button press. Additionally, we separately modelled six motion parameters [13:18] and separately modelled the mean for each of the four runs [19:22].

We used the *marsbar* toolbox using its default parameter settings to extract the averaged beta values representing the respective onset regressors from the ROIs per session and block (Brett et al., 2002). The left and right ROI masks for amygdala and hippocampus were extracted from the FSL Harvard-Oxford (HO) Atlas (Desikan et al., 2006). Furthermore due to functional differences, the hippocampus mask was

divided into an anterior and posterior part based on the conventional landmark of the uncal apex (y = -21/22 in MNI space, Poppenk et al., 2013).

Second Level analysis. To statistically compare insight-related mean activity differences in amygdala and hippocampal brain activity, we estimated three nested mixed effect models as depicted below including the averaged beta value per ROI as dependent variable and subjects as random intercepts. To control for potential differences between the sessions and blocks, a factor specifying the order of the respective run order (1-4) was always added as covariate of no interest. A factor that differentiated between two conditions –solved trials with high (*HI-I*) and low insight (*LO-I*) and a variable specifying the respective ROI served as independent measure.

I. Beta values ~ run + ROI + (1| ID) + Ɛ
II. Beta values ~ run + ROI + insight + (1 | ID) + Ɛ
III. Beta values ~ run + ROI * insight + (1 | ID) + Ɛ

Single trial data analyses including response time and accuracy as covariate. To exclude the fact that the observed amygdala and hippocampal effects are just due to differences in response time and accuracy between the high and low insight condition, we conducted a control analysis including response time and accuracy as covariates of no interest. The use of covariates to regress out unwanted variance is not possible within the SPM GLM framework. Therefore, we conducted the analyses with the data from the single-trial analyses containing a beta estimate of univariate amygdala activity for the solution button press at each trial (see Method section, *Multivariate fMRI data analysis – multivoxel pattern similarity analysis: Extraction of ROI-wise multivoxel patterns).*

To statistically compare insight-related mean activity differences in amygdala and hippocampus (ROI variable) during solution (time point of solution button press), we estimated three nested mixed effect models (I.-III.) as depicted below including the averaged amygdala, anterior and posterior hippocampus beta value per trial as dependent variable and subjects as well as items as random intercepts. To control for potential differences between the sessions and blocks, a factor specifying the order of the respective run order (1-4), response time (RT), and accuracy were added as covariates of no interest. A factor that differentiated between two

conditions – solved trials with high (*HI-I*) and low insight (*LO-I*) served as independent measure. To calculate insight-related better memory on amygdala and hippocampal activity during solution, we additionally estimated memory for the solved trials and their interaction with insight and ROI (see equation IV. - VI. below).

I. Beta values ~ run + acc + RT + ROI+ (1| ID) + (1| item) +ε

II. Beta values ~ run + acc + RT + ROI + insight + (1 | ID)+ (1| item) + ε

III. Beta values ~ run + acc + RT + ROI* insight + (1 | ID) + (1| item)+ ε

IV. Beta values ~ run + acc + RT + ROI+ insight + memory + (1 | ID) + (1| item)+ ε

V. Beta values ~ run + acc + RT + ROI+ insight * memory + (1 | ID) + (1| item)+ ε

VI. Beta values ~ run + acc + RT + ROI* insight*memory + (1 | ID) + (1| item)+ ε

Note. acc = accuracy, RT = solution time; ID = random

The effect of insight on amygdala and hippocampal brain activity was tested by comparing the nested models I.-III via log-likelihood tests. To additionally test the effect of insight-related memory on brain activity, we compared the nested models II. and IV. - VI via log-likelihood tests. Posthoc marginal mean differences between conditions were estimated using the emmeans package (v.1.7.5). Posthoc results regarding insight-related better memory were additionally adjusted for multiple comparison via the Tukey HSD method (Tukey, 1953).

Multivariate fMRI data analysis – representational similarity analyses

We hypothesised that 1) RC should be associated with changes in distributed, i.e. multivoxel, patterns from pre- to post-solution, 2) those patterns should start representing the meaningful solution object after solution and 3) visual RC in Mooney images (perceptual regrouping of black/white contours leading to sudden object recognition) should occur in regions involved in visual object recognition along the visual ventral pathway such as the temporal-occipital [pITG], anterior [aITG] and posterior [mITG] Inferior Temporal Gyrus; inferior Lateral Occipital Cortex [iLOC]; anterior, posterior, occipito-temporal [toFusG] and occipital [oFusG] Fusiform Gyrus (DiCarlo et al., 2012; Grill-Spector et al., 2001; Kanwisher et al., 1997). The masks for those ROIs were extracted from the FSL Harvard-Oxford (HO) Atlas (Desikan et al., 2006). Note, to reduce the number of Fusiform Gyrus ROIs but respect functional

differences (Weiner & Zilles, 2016; W. Zhang et al., 2016), we merged this region into only two ROIs: anterior [aFusG] and posterior [pFusG] Fusiform Gyrus. To create the aFusG, the anterior and posterior Fusiform Gyrus ROIs from the HO Atlas were merged into one. To create the pFusG, the toFusG and oFusG ROIs were merged into one.

We used RSA to test whether and which ROIs along the visual ventral pathway (VOTC areas) exhibit those properties during insightful problem solution. We first extracted the multivoxel pattern in the above named ROIs per subject, per Mooney image and for a pre- as well as post-solution time point. Subsequently, we conducted two different types of representational similarity analysis: 1) pre- to post-solution similarity to identify ROIs showing a stronger decrease in correlation for *HI-I* than *LO-I* trials and 2) a model-based RSA (Kriegeskorte et al., 2008; similar to Davis et al., 2021) where distributed activity patterns were correlated with the output of two very distinct neural network models, AlexNet and Word2Vec.

Extraction of ROI-wise multivoxel patterns. To obtain multivoxel patterns, we first estimated beta values in each voxel for each pre- and post-solution event. Due to the short response time in particular for *HI-I* events, the pre- and post-solution regressors would be highly collinear, i.e. explain shared variance of the time-series, which results in unreliable parameter estimation and an hence an underestimation of the similarity of activity patterns (Mumford et al., 2015). Therefore, we conducted two separate first-level models - one for the pre- and one for the post-solution time points in which each event was represented by an individual regressor (its onset convolved with the HRF). Pre-solution onsets were 0.5 sec after onset of the Mooney images and post-solution onsets were the time point of the button presses; both events were modelled with a duration of 1 sec. Because there was no button press for unsolved trials during the stimulus presentation, we created a jittered onset regressor for the “post-solution stage” of the unsolved trials that on average matched the response time of the solved trials (~3.7sec). These two univariate single-trial models were conducted as general linear models as implemented in SPM. Due to the sluggish HRF, one would expect a higher intrinsic auto-correlation of the activity patterns of pre and post events when they occur close in time which is on average more often the case for *HI-I* than *LO-I* events. Importantly though, this auto-correlation rather counteracts the hypothesised greater decrease in similarity in *HI-I* trials because

they were solved faster and therefore do not confound the results. We added an additional nuisance regressor for all remaining button presses (related to the insight ratings and selecting a solution category) for one second starting with the respective button press. Additionally, we separately modelled six motion parameters and separately modelled the mean for each of the four runs. The time series were corrected for baseline drifts by applying a high-pass filter (128 sec) and for serial dependency by an AR(1) autocorrelation model.

Subsequently, the corresponding pre- and post-solution activation patterns for each Mooney image from the corresponding beta images per ROI (iLOC, aFusG, pFusG, aITG mITG, pITG) were extracted and vectorized. Those ROI-wise multivoxel patterns were used for 1) Pre- to post-solution RSA and 2) Pre- to post-solution model-based RSA. All analyses were conducted using in-house MATLAB scripts.

1. Pre- to post-solution RSA. To compare the trial-wise change in multivoxel pattern from pre- to post-solution, we first matched the extracted pre- and post-solution activation patterns for each Mooney image from the corresponding beta images per ROI into a pre- to post solution vector pair and subsequently correlated it (Spearman's Rho) (see Fig.3-A).

For statistical analysis, a linear mixed effects model was applied estimating the multivoxel pattern similarity between the pre- and post-solution phase (dependent variable) as a function of the insight factor (*HI-I, LO-I*) and respective ROI. Additionally, accuracy, response time and the run order (1-4) were added as covariate of no interest. Item and subject variables were estimated as random intercepts. To estimate differences in multivoxel pattern similarity between ROIs as a function of the high versus low insight, another model was estimated including an interaction term between ROI and the insight factor (see equations below).

I. PPSS ~ acc + RT+ run + ROI + (1| ID) + (1|item) + Ɛ

II. PPSS ~ acc + RT+ run + ROI + insight + (1| ID) + (1|item) + Ɛ

III. PPSS ~ acc + RT+ run + ROI * insight + (1| ID) + (1|item) + Ɛ

Note. PPSS = Pre-Post Solution Similarity; acc = accuracy; RT = response time.

Resulting p-values were FDR-corrected because we did not have a hypothesis which ROI may exhibit insight-related changes in multivoxel pattern similarity.

2. Pre- to post-solution model-based RSA. Here, representational similarity between the activity pattern elicited by the stimuli is compared to the computationally model-based representational similarity between the same stimuli. This analysis involves four steps: 1) creating representational similarity matrices for each ROI based on the correlation of the multivoxel activity patterns elicited by the stimuli, 2) creating a model-based similarity matrix based on the correlation of the output of the computational model for each stimulus 3) correlating the model-based similarity matrix with the ROI-specific representational similarity matrices for each time point 4) associating the brain-model correlation for each trial with insight or insight*memory.

1) Creating Neural Activity-derived RSMs (Brain RSM)

For each subject, ROI (iLOC, aFusG, pFusG, aITG, mITG, pITG) and time point (pre- and post solution), a *Neural activity-derived RSM* (120x120) was created representing the similarity in the multivoxel activity pattern across stimuli. That is to say, each cell in a matrix represents the Spearman correlation between the vectorized multivoxel patterns of two Mooney images (see Fig.3-B1).

2) Creating Model RSMs

Because we assumed that *conceptual* similarities between the Mooney objects increases after problem solution (=object identification), we created a *Model RSM* consisting of pair-wise *conceptual* similarity values for each Mooney image pair represented in a 120x120 matrix. To assure robustness of the results, we used two different methods to quantify conceptual similarity between the Mooney images - a deep convolutional network (DNN) AlexNet (Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, 2015; Krizhevsky et al., 2017) and a Word2Vec model (Becker, Wiedemann, et al., 2020; Goldberg & Levy, 2014; Mikolov et al., 2013) resulting in two different conceptual *Model RSMs*.

Conceptual similarity via the DNN AlexNet *Model RSM* is based on categorical features (Krizhevsky et al., 2017). DNNs consist of layers of convolutional filters that are trained to classify images into categories with a high level of accuracy. While filters from early layers predominately detect simple visual features, later layers organise items by their categorical features (Zeiler & Fergus, 2014). According to previous research, DNNs like AlexNet model visual representations

along the ventral visual pathway (Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, 2015) and outperform traditional theoretical models of the ventral visual pathway (e.g., HMAX, object-based models; Cadieu et al., 2014; Groen et al., 2018) in their capacity to identify specific objects with the appropriate category-level label. They may therefore be particularly well suited to investigate RC of visual objects using Mooney images. Here, we used the pre-trained 8-layer DNN, the AlexNet, which was successfully trained to classify 1.2 million images into 1000 categories (Krizhevsky et al., 2017). AlexNet consists of eight layers, the first five are convolutional (some are followed by max-pooling layers) and the last three are fully connected layers. Conceptual similarity between the Mooney objects was quantified as follows: First, the actual non-abstracted, colored object images (not their abstracted Mooney counterpart) were entered into the pre-trained AlexNet and the activations of its penultimate layer for each image were extracted and converted into one activation vector (note, activation from this layer can be regarded as conceptual because it is the last layer before the images are explicitly categorised into the trained categories (Bankson et al., 2018; Devereux et al., 2018; Groen et al., 2018). Subsequently, for each pair of images, the similarity (Spearman correlation) was computed between the activation vectors resulting in a 120x120 matrix (=*Model RSM*). For a schematic visualisation of this process, see Fig.3-B2).

In contrast, in the Word2Vec model, pair-wise stimuli similarity is based on cosine similarity derived from statistical co-occurrences in text data via a pre-existing word embedding model as previously published (see Becker, Davis, et al., 2022; Becker, Wiedemann, et al., 2020 for more details). Word embeddings represent words as dense numerical vectors derived from a neural network being trained with a huge text corpus using a word2vec algorithm (see Mikolov et al., 2013). Cosine similarity is defined as the angle between two 300 dimensional vectors that represent both stimuli (concepts, e.g. “tiger” and “dog”) in this vector space. The *Model RSM* that was created with Word2Vec consisted of pair-wise conceptual similarity values for each Mooney concept (word) represented in a 120x120 matrix.

3) Computing an item-wise representational strength

Subsequently, the item-wise rows (one item = one Mooney image object) from the *Neural-activity-derived RSM* for each subject, each time point and each ROI were partially correlated (Pearson's r) with the same item-wise rows of each *Model RSM* (AlexNet and Word2Vec). Note, to control for the confound of having presented the Mooney images in different scans, i.e. blocks and sessions, those item-wise row correlations between the brain and model data were partialled out by a variable indicating the respective run order (1-4). Such an item-wise approach differs from the classical method of assessing such second-order correlations between brain and model activation patterns (Kriegeskorte & Kievit, 2013), which typically relate the entire item × item matrix at once (note, this item-wise approach has been successfully applied elsewhere, Davis et al., 2021). Furthermore, the item-wise row correlations were further filtered into solved and unsolved Mooney images because we assumed that only solved Mooney images (successful object identification) should exhibit conceptual representational similarity among each other. The resulting representational strength ("2nd order correlation") identifies brain regions that process and/or store conceptual representations emphasising either categorical visual features (AlexNet) or similar meaning based on co-occurrences (Word2Vec). Hence, higher representational strength values in a brain region indicate higher conceptual similarity (see Fig.3-B3).

4) Predicting the representational strength values

The statistical analysis was similar to the *pre- to post-solution similarity analysis* using mixed effects models in a nested way to estimate the effect of changes in representational similarity (i.e. representational strength or brain-model fit) as a function of insight (*HI-I vs LO-I*), ROI and time (see equations below). Time was a factor that indicated the pre- (0.5sec after stimulus presentation) or the post-solution (solution button press) phase. Note, the ROI factor only contained pFusG, pITG and iLOC because only those visual brain areas exhibited changes in pre- to post solution similarity (see A.). Additionally, response time (RT) was added as covariate of no interest. Next to the ROI*insight interaction, we additionally modelled a 3-way interaction (ROI*insight*time) to test which ROI exhibits insight-related changes in representational similarity from pre- to post-solution (IV., see equations below).

I. Rep. Strength ~ acc + RT+ run + time + ROI + (1|ID) + (1|item) + Ɛ

II. Rep. Strength ~ acc +RT+ run + time + ROI + insight + (1|ID) + (1|item) + Ɛ

III. Rep. Strength ~ acc +RT+ run + time + ROI * insight + (1|ID) + (1|item) + Ɛ

IV. Rep. Strength ~ acc +RT+ run + time * ROI * insight + (1|ID) + (1|item) + Ɛ

Note. Rep.Strength = Representational Strength, i.e. brain-model fit; acc = accuracy.

Finally, we tested whether a) differences in pre to post solution multivoxel pattern similarity as well as b) the representational strength of the solution object after solution in those ROIs that show all described properties of RC (pFusG, iLOC) are related to insight-related better memory. To test this, we estimated another set of nested linear mixed models with a) pre to post solution similarity (PPSS) and b) the changes in representational strength (Rep.Strength: AlexNet or Word2Vec) from pre to post solution as dependent variables and ROI as well as an interaction between (time and) insight and memory as independent factors (see equations below). Note, only pFusG and iLOC exhibited all properties of RC, for this reason we only estimated the described models for those two ROIs. Response time, accuracy and the run order were entered as covariates of no interest and items and subjects were estimated as random intercepts. To test for differences in insight-related memory between ROIs, another model with ROI*insight*memory interaction term was additionally modelled (IV., see equations below).

I. PPSS ~ acc + RT+ run + ROI + insight + (1| ID) + (1|item) + Ɛ

II. PPSS ~ acc +RT+ run + ROI + insight + memory + (1| ID) + (1|item) + Ɛ

III. PPSS ~ acc + RT+ run + ROI + insight * memory + (1| ID) + (1|item) + Ɛ

IV. PPSS ~ acc +RT+ run + ROI * insight * memory + (1| ID) + (1|item) + Ɛ

Note. PPSS = Pre-Post Solution Similarity;acc = accuracy; RT = solution time.

I. Rep.Strength ~ acc + RT+ run + ROI + time + insight + (1| ID) + (1|item) + Ɛ

II. Rep.Strength ~ acc +RT+ run + ROI + time + insight + memory + (1| ID) + (1|item) + Ɛ

III. Rep.Strength ~ acc + RT+ run + ROI + time + insight * memory + (1| ID) + (1|item) + Ɛ

IV. Rep.Strength ~ acc +RT+ run + ROI + time * insight * memory + (1| ID) + (1|item) + Ɛ

V. Rep.Strength ~ acc +RT+ run + ROI * time * insight * memory + (1| ID) + (1|item) + Ɛ

Note. Rep. Strength = Representational strength of solution object at post solution (button press); acc = accuracy; RT = solution time.

Multivariate fMRI data analysis – functional connectivity analysis

We expected insight-related increased functional connectivity between those areas that process the solution object (VOTC) and those that likely evaluate it such as the amygdala and hippocampus. To estimate functional connectivity (FC) between those areas, we used the fMRIprep preprocessed fMRI data (see above) and first denoised and subsequently analysed it in CONN - an open source connectivity toolbox (version 21.a, (Whitfield-Gabrieli & Nieto-Castanon, 2012).

Denoising. To remove confounding effects to the estimated BOLD signal for each subject in each voxel and at each time point, i.e. session and block, CONN's implemented anatomical component-based noise correction procedure (CompCor) was used. Those confounding effects involve six subject-motion parameters including their first-order derivatives, identified outlier scans, constant or linear condition effects as well as noise components from areas of cerebral white matter and cerebrospinal fluid. Note, to avoid the risk of artificially introducing anticorrelations (negative connectivity) into the FC estimates, no global signal regression was applied (Chai et al., 2012). Finally, the resulting time series were band-pass filtered to .008 - .09 Hz.

First-level analysis. FC during high versus low insight conditions (*HI-I>LO-I*) was estimated. We were only interested in ROIs processing the solution object and therefore constructed a specific *solution network* comprising left and right aFusG, pFusG, aITG, mITG, pITG, iLOC including Amygdala and anterior as well as posterior hippocampus (18 ROIs in total). Amygdala and hippocampus were included because we were additionally interested in whether those brain areas that showed insight-related differences in univariate BOLD activity are also part of this *solution network*. All ROIs taken from the FSL Harvard Oxford Atlas were identical to the ones already used for univariate and the other multivariate analyses but analysed separately by hemisphere (see Perez et al., 2023; Schöne et al., 2022). The timeseries for each ROI were acquired by averaging the BOLD time series of each voxel belonging to the respective ROI. ROI-to-ROI connectivity matrices (*solution network*: 18x18; subnetwork: 10x10) were computed for each participant (31), for each condition (1) *HI-I vs. LO-I*, each session (2) and each block (2) separately. Each element in each connectivity matrix represents the strength of FC between a

pair of ROIs and was defined as the Fisher-transformed bivariate correlation coefficient between the preprocessed and denoised BOLD time series of two ROIs.

Second-level analysis. To estimate FC for the contrast *HI-I>LO-I* in the *solution network*, the within-subjects contrast in the GLM was specified as follows y~ 1*HI-I – 1*LO-I + ε). For statistical inference, the 18x18 ROI connectivity matrix was thresholded at the connection-level at p<.05 (p-uncorrected, two-sided) and at the cluster-level at p<.05 (ROI-p FDR corrected) (Benjamini & Hochberg, 1995). Note, for cluster-level inference we adopted the Network Based Statistics (NBS) approach (Zalesky et al., 2010).

To quantify whether the cognitive (RC) and evaluative component of insight share direct functional connections, we additionally specified a *subnetwork* of the *solution network* only including left and right pFusG, iLOC, amygdala and anterior/posterior hippocampus (10 ROIs in total) and repeated the analysis using the same connection- and cluster-level threshold as for the *solution network*.

To further explore whether FC in the *solution network* also predicts insight-related better memory, we estimated another first level analysis identical to the one described above with the only difference that four different conditions were estimated: 1) high insight - remembered trials [*HI-I-Rem*], 2) high insight - forgotten trials [*HI-I-Forg*], 3) low insight - remembered trials [*LO-I-Rem*] and 4) low insight - forgotten trials [*LO-I-Forg*]. For the second level analysis, we specified the following within-subjects contrast in the GLM: y~ 1*HI-I-Rem - 1*HI-I-Forg - 1*LO-I-Rem + 1*LO-I-Forg + ε.

Graph-theoretical measures. To further characterise efficient information integration during insight we computed several network measures quantifying functional integration (increased global efficiency, reduced average path length), segregation (increased local efficiency and clustering coefficient) and centrality (degree) of the *solution network* and its individual ROIs with the contrast *HI-I>LO-I* as implemented in CONN (Whitfield-Gabrieli & Nieto-Castanon, 2012). Note, the description and mathematical formulas for each graph measure is in Table S3.

All ROI-graph measures were based on functional connectivity values of our pre-defined non-directed *solution network* (18x18 Fisher transformed correlation matrix). For each subject and condition (*HI-I, LO-I*) a series of adjacency matrices

are then computed by thresholding the respective Fisher transformed correlation matrix by different correlation coefficients ranging from r=0.15 to r=0.8 in steps of 0.1. Note, only positive correlations were included for better interpretation of the subsequent graph measures (Nieto-Castanon, 2020). From those resulting adjacency matrices, several graph-theoretical network measures of interest were subsequently computed for the contrast *HI-I>LO-I* addressing topological properties of each ROI within the matrix but also of the entire *solution network*. Importantly, to demonstrate the robustness of the results rather than just picking an arbitrary threshold, we aggregated across the different adjacency matrices (acquired via those different correlation coefficient thresholds) to compute the respective graph measure of interest. Note, we tested one-sided due to directed hypotheses.

Data and code availability

The behavioural data including the aggregated univariate and multivoxel pattern fMRI data, as well as the analysis code and the stimulus material have been made publicly available and can be accessed at https://github.com/MaxiBecker/Insight_Memory_Effect.

During the preparation of this work the author(s) used chatgpt 3.5 in order to improve the language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Supplementary Material

Figure S1. Mooney image examples

Note. The black/white abstracted images on the left represent the Mooney images and the corresponding real-world pictures are presented on the right.

Categories as response options for solution in the MRI scanner

1. Leisure & Hobbies,
2. Tool, weapon or technology,
3. Clothing & Accessories,
4. Vehicles & Road Traffic,
5. Landscape or building,
6. Vehicles & Road Traffic,
7. Predators or rodents,
8. Mammal,
9. Reptiles or Insects,
10. Birds or animal that lives in/near the water,
11. Events or Science,
12. Objects in the house,
13. Human Being

Relationship between different dimensions of insight experience

To simplify the analyses, we explored whether it is justified to combine the three insight dimensions (*Suddenness, Emotion* and *Certainty*) into one single measure. For this, we calculated a measurement model for a latent insight factor from those three insight dimensions for correctly identified Mooney images. The latent factor was estimated within a confirmatory factor analysis (CFA) in R using the lavaan package (v. 0.6-15, Eves Rosseel, 2012). The model fit was evaluated via Bentler's comparative fit index (CFI), root mean square error of approximation (RMSEA) and standardized root-mean-square residual (SRMR). Accepted thresholds indicating good model fit are RMSEA ≤ .05, SRMR <0.1 and CFI ≥ = .95 (Hu & Bentler, 1998, 1999; Schermelleh-Engel et al., 2014). Note, due to a lack of degrees of freedom no exact *Chi²* goodness-of-fit statistic could be calculated. Error variance of the latent factor was fixed to one to receive factor loadings from all three insight dimensions.

The model converged normally after 12 iterations. The latent insight factor loaded significantly onto all three variables: *Certainty* (λ=.620, z=27.22; p<.001), *Emotion* (λ=.618, z=26.93; p<.001) and *Suddenness* (λ=.574, z=24.27; p<.001) suggesting that those variables contribute significantly to the latent insight factor. Practical fit indices (CFI= 1.0; RSME = 0.000; SRMR =0.000) suggested a good fit of the model to the data.

Note, the significant factor loadings did not change significantly when including all trials (i.e. also incorrectly identified Mooney images). In sum, those results justified combining the three insight dimensions into one single measure for further analyses.

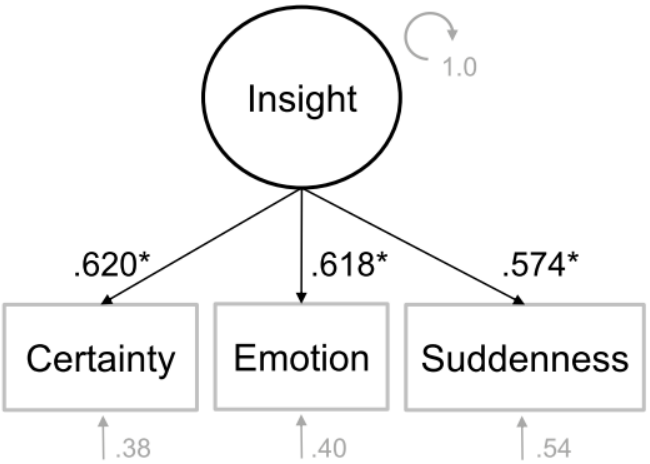

Figure S2. Measurement model for latent insight variable

Note. Asterisk = significance at p<.001;

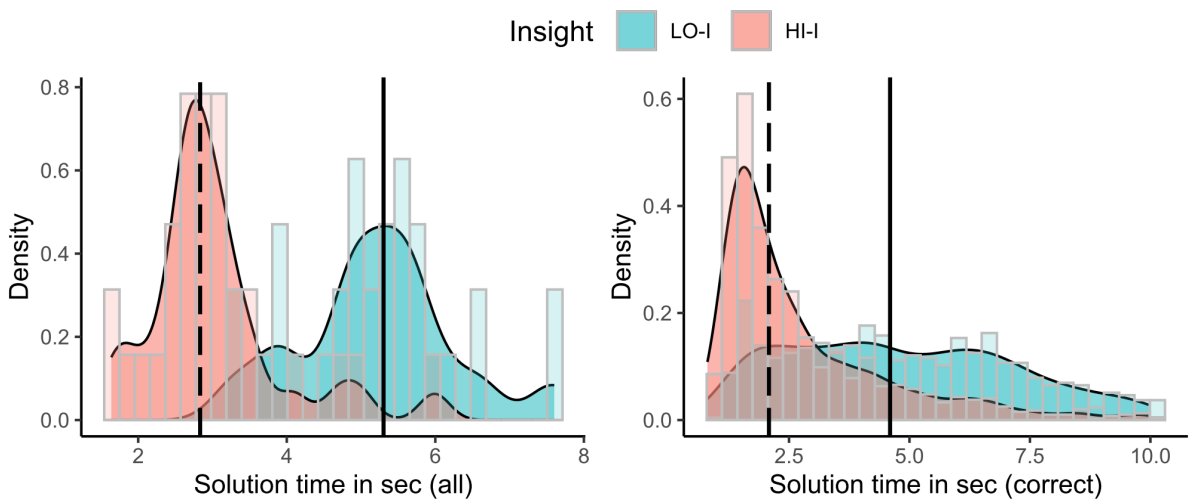

Figure S3. Response time distribution divided by high and low insight solutions (within subject).

Note. *LO-I* = low insight trials; *HI-I* = high insight trials. (all) = all solved trials; (correct) = only correctly solved trials. Dashed line = median response time for *HI-I* trials; Full line = median response time for *LO-I* trials.

Univariate fMRI results of VOTC areas

For exploratory purposes, we investigated univariate brain activity in VOTC areas as a function of insight (HI-I>LO-I) and for insight-related memory in VOTC areas. To investigate insight, we performed the same analyses as for amygdala and hippocampus controlling for run order and including random subject and item intercepts. To investigate an insight*memory interaction, we used the beta values from the single trial analysis including accuracy, response time, run order as covariates of no interest and random subject and item intercepts. Because we did not have a specific hypothesis for the VOTC areas, resulting p-values from both analyses were FDR-corrected for multiple comparison.

Insight. Regarding univariate effects of VOTC areas, there was a significant but reversed condition effect ($Chi^2(1)=14.90$, $p<.001$, $ß=-.02$) as well as condition*ROI interaction ($Chi^2(5)=38.89$, $p<.001$) effect. Posthoc analyses revealed that iLOC ($t(1445)=-5.45$, *p-FDR*<.001) and pFusG ($t(1445)=-4.40$, *p-FDR*<.001) showed the reverse pattern with more BOLD activity for *LO-I* compared to *HI-I* trials while BOLD activity did not significant vary as a function of insight in the other ROIs (p>.17) (see Fig.S4-A).

*Insight*Memory*. There was a significant memory effect ($Chi^2(1)=4.56$, $p<.05$, $ß=-.02$) when controlling for insight and a reversed insight*memory effect ($Chi^2(1)=32.44$, $p<.001$, $ß=-.07$) while the ROIs also differed in this insight*memory effect ($Chi^2(15)=171.97$, $p<.001$). Post Hoc analyses revealed that pFusG ($Chi^2(1)=20.55$, $p<.001$, $ß=-.13$), iLOC ($Chi^2(1)=13.62$, $p<.001$, $ß=-.15$), pITG ($Chi^2(1)=4.75$, $p<.05$, $ß=-.09$), aITG ($Chi^2(1)=3.89$, $p<.05$, $ß=.08$), mITG ($Chi^2(1)=4.50$, $p<.05$, $ß=-.09$) showed a significant insight*memory effect but not in the hypothesized direction, i.e. highest activity for remembered insight trials (see Fig.S4-B).

In sum, univariate effects in VOTC regions cannot explain the observed multivariate insight and insight*memory effects in those regions.

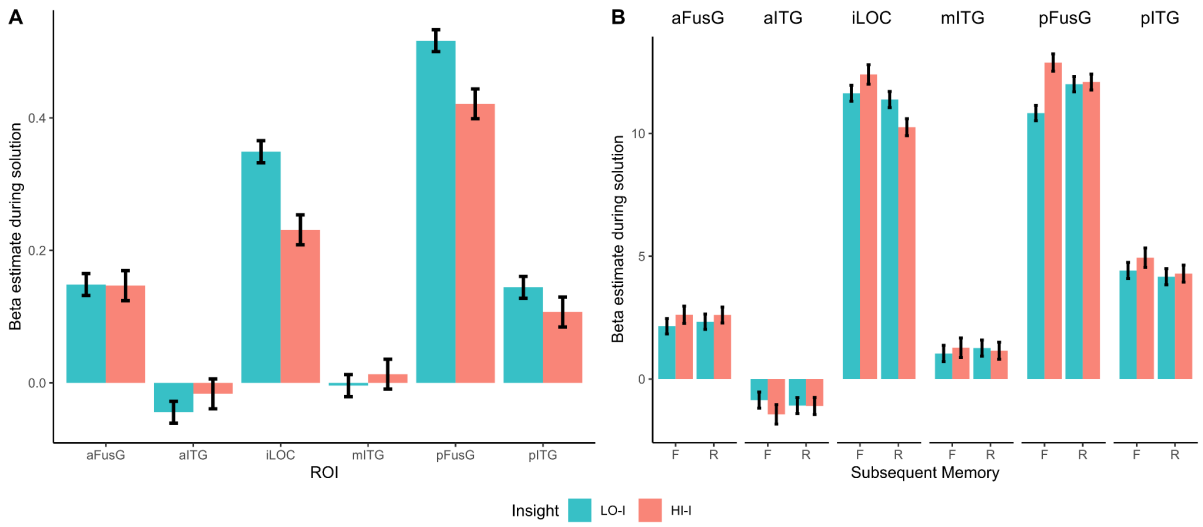

Figure S4. Univariate BOLD activity in VOTC areas (not) predicting insight.

Note. **Panel A.** BOLD activity in VOTC for insight (HI-I>LO-I). **Panel B.** BOLD activity in VOTC for an insight*memory interaction (data taken from single trial analysis). F = forgotten trials; R= remembered trials.

Univariate fMRI analyses and results in subject space

To exclude potential normalisation artefacts due to the small volume of the amygdala and hippocampal ROIs, we additionally conducted another ROI analysis in subject space for the contrast *HI-I>LO-I*.

Functional and structural data were preprocessed using fMRIprep in the same way as for the analysis in standard space (see Methods section *fMRI data preprocessing)*, with the exception that data were resampled into the individual's

anatomical reference generated with the T1 weighted images. All further analyses (first and second level analyses, beta value extraction via marsbar) were identical to the analyses reported for the data in standard space further above. The only exception was that the amygdala and the anterior and posterior hippocampus ROIs were first resampled into subject-space using the ANTs function *antsApplyTransforms* (Avants et al., 2009) before extracting the condition specific mean beta values using marsbar.

Activity in amygdala and hippocampus during solution predicts Insight

As illustrated by Fig.S5, amygdala and hippocampal activity significantly distinguish between *HI-I* and *LO-I* trials during solution ($Chi^2(1)=70.29$, $p<.001$) while there are no significant differences between the ROIs ($Chi^2(1)=4.75$, $p=.093$). Posthoc analyses revealed that amygdala activity ($t(707)=6.40$, $p<.001$, ß=.45), anterior ($t(707)=5.13$, $p<.001$, ß=.47) and posterior ($t(707)=3.34$, $p<.001$, ß=.29) hippocampus activity were significantly higher for *HI-I* compared to *LO-I* trials.

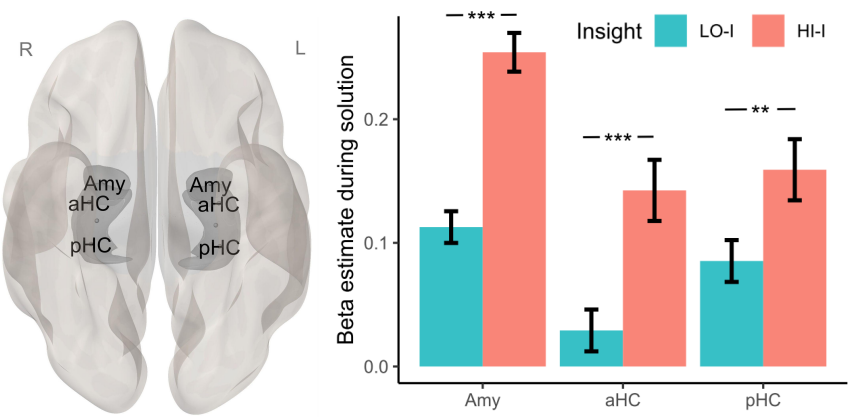

Figure S5. Univariate BOLD activity of amygdala and hippocampus during solution in native space predicted by insight

Note. Values = estimated marginal means ± SEM. *** represents p-value <.0001; ** represents p-value <.001;. Amygdala (Amy), anterior (aHC) and posterior (pHC) hippocampal activity during solution divided by condition (HI-I = high insight, LO-I = low insight).

Table S1.

Individual Connections from *solution network* for *HI-I > LO-I*

ROI1	ROI2	T(30)	p-FDR
iLOC l	mITG l	4.53	0.000087
iLOC l	aFusG r	3.12	0.004025
iLOC l	mITG r	2.58	0.015132
iLOC l	Amygdala l	2.45	0.020177
iLOC l	Amygdala r	2.16	0.038953
iLOC l	toITG l	2.09	0.044807
iLOC r	mITG l	2.95	0.006062
iLOC r	aFusG r	2.94	0.006189
iLOC r	mpITG r	2.52	0.017254
iLOC r	Amygdala r	2.23	0.033418
iLOC r	aHC l	2.10	0.044623
pFusG l	toITG r	2.88	0.007313
pFusG l	iLOC l	2.87	0.007388
pFusG l	aITG l	2.71	0.011104
pFusG l	mITG r	2.46	0.019692
pFusG l	mITG l	2.38	0.023779
pFusG r	aHC l	2.70	0.011354
pFusG r	mITG l	2.59	0.014599
pFusG r	aITG l	2.57	0.015275
pFusG r	iLOC r	2.11	0.043319
aHC l	pITG l	2.32	0.027128
aHC r	Amygdala l	2.38	0.024043
aHC r	aITG l	2.44	0.020969
pHC l	pITG l	2.24	0.032769
pHC l	aFusG l	2.20	0.035558

mITG l	pITG l	3.18	0.003438
aITG l	mITG r	2.06	0.047715
pITG l	Amygdala l	2.05	0.049440

Note. The "solution" network contains the ROIs a/pFusG, iLOC, a/m/p ITG, amygdala, anterior (aHC) and posterior (pHC) hippocampus.

Table S2.

Individual Connections from subnetwork for HI-I > LO-I

ROI1	ROI2	T(30)	p-FDR
pFusG r	aHC l	2.70	0.0446
iLOC r	aHC l	2.10	0.0446
iLOC l	Amygdala l	2.45	0.0446
aHC r	Amygdala l	2.38	0.0446
iLOC r	Amygdala r	2.23	0.0446
iLOC l	Amygdala r	2.16	0.0446
pFusG l	iLOC l	2.87	0.0446
pFusG r	iLOC r	2.11	0.0446

Note. The subnetwork contains the ROIs pFusG, iLOC, amygdala, anterior (aHC) and posterior (pHC) hippocampus.

Graph measures for characterising insight-related more efficient information integration within the *solution network*

Degree serves as a reliable indicator of centrality for both networks and ROIs, capturing the level of local interconnectedness within a graph or the entire network. At the ROI level, *Degree* quantifies the edges connected to/from each ROI, while the network degree calculates the mean degree across all graph nodes (Whitfield-Gabrieli & Nieto-Castanon, 2012).

Functional integration within the brain denotes its capacity to rapidly combine specialised information from distributed brain regions, thus illustrating the fluidity of

inter-regional communication (Rubinov & Sporns, 2010). Metrics indicating functional integration or the extent of overall interconnectedness in the network comprise *Average path distance* and *Global efficiency*. The *Average path distance* between pairs of ROIs in a graph refers to the minimum connections to navigate between them along an optimal route. Consequently, the node-specific average path distance pertains to the mean path distance between the node and all other nodes in the connected subgraph. *Global efficiency* is calculated as the average of inverse-distances between a given ROI and all other ROIs within the same graph (Whitfield-Gabrieli & Nieto-Castanon, 2012).

Functional segregation within the brain refers to its capacity for specialized processing to unfold within densely interconnected clusters of brain regions (Rubinov & Sporns, 2010). Indicators of functional segregation or localized integration encompass the *Clustering coefficient* and *local eff*iciency. The *Clustering coefficient* is delineated as the ratio of interconnected edges in the nearby sub-graph corresponding to each ROI. Node-specific *Local efficiency* is defined as the Global efficiency within the neighboring sub-graph of the given ROI (Whitfield-Gabrieli & Nieto-Castanon, 2012).

Table S3.

Mathematical definitions of the used graph measures as implemented in CONN

Function	Graph measure	ROI-level	Network-level
Centrality	Degree (d)	$d_i = \Sigma_j A_{i,j}$	$d = \frac{\Sigma_i d_i}{N}$
Functional integration	Average path length (L)	$L_i = \frac{\Sigma_{j \in \Omega_i} D_{i,j}}{N_i - 1}$	$L = \frac{\Sigma_i L_i}{N}$
	Global Efficiency (GE)	$GE_i = \frac{\Sigma_{i \neq j} 1/D_{i,j}}{N-1}$	$GE = \frac{\Sigma_i GE_i}{N}$
Functional segregation	Local efficiency (LE)	$LE_i = \frac{\Sigma_{j,k \in \Gamma_i} 1/Dy_{j,k}(i)}{d_i(d_i - 1)}$	$LE = \frac{\Sigma_i LE_i}{N}$
	Clustering Coefficient (CC)	$CC_i = \frac{\Sigma_{j \neq k \in \Gamma_i} A_{j,k}(i)}{d_i(d_i - 1)}$	$CC = \frac{\Sigma_i CC_i}{N}$

Note. A = adjacency matrix; N = total number of nodes in a graph; di= degree of a graph at each node; D = shortest-path distance matrix, Dy is the shortest-path distance matrix within the neighboring sub-graph at each node, characterized by all nodes neighboring this node and all existing edges among them.

Relationship between Insight-related efficient information integration and subsequent memory

Finally, for reasons of completeness, we further explored whether different graph measures in the pre-defined *solution network* also predicted insight-related better memory. For this, we repeated FC and graph theoretical analyses but instead of a simple *HI-I>LO-I* contrast, we now calculated an insight (*HI-I vs LO-I)* x memory (remember vs. forgotten) interaction with four conditions: 1) high insight - remembered trials [*HI-I-Rem*], 2) high insight - forgotten trials [*HI-I-Forg*], 3) low insight - remembered trials [*LO-I-Rem*] and 4) low insight - forgotten trials [*LO-I-Forg*] to construct the adjacency matrices. To calculate an interaction, we used the following contrast 1*HI-I-Rem - 1*HI-I-Forg - 1*LO-I-Rem + 1*LO-I-Forg for subsequent FC and graph theoretical analyses.

Functional connectivity. Neither the *solution network* nor its subnetwork (only consisting of bilateral pFusG, iLOC, amygdala and anterior and posterior hippocampus) survived multiple comparisons at the cluster-threshold for an insight x memory interaction.

Graph theoretical measures. Average pathlength was significantly reduced for this interaction ($t(30)=-3.60$, $p<.001$, $ß=-.17$) at the *solution network*-level and also bilateral pFusG (right: $t(29)=-2.66$, *p-FDR*=.025, $ß=-.15$, left: $t(29)=-2.26$, *p-FDR*=.042, $ß=-.15$) and left iLOC ($t(27)=-3.33$, *p-FDR*=.01, $ß=-.13$) exhibited significantly less average pathlength compared to the other ROIs in the network (note, right iLOC did not survive multiple comparison [FDR-p=.069] but was significant at an uncorrected threshold $t(29)=-1.89$, *p-uncor*=.034, $ß=-.11$). In addition, the insight*memory interaction in the *solution network* was associated with a trend towards significance for an increased clustering coefficient ($t(26)=1.70$, $p=.051$, $ß=.06$) and local efficiency ($t(26)=1.54$, $p=.068$, $ß=.05$). No individual ROI survived multiple comparisons for the latter two network measures. Neither degree

(p>.43) nor global efficiency (p>.86) showed significant effects.

In sum, the results from FC and graph theoretical measures provide no consistent evidence but some graph measures (reduced average pathlength, trend for increased clustering coefficient and local efficiency) suggest an association between insight-related efficient information integration and better subsequent memory.

Supplement References

Avants, B. B., et al. (2009). "Advanced normalization tools (ANTS)." Insight j 2(365): 1-35.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, *82*, 1-26.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modelling. Journal of Statistical Software, 48(2), 1 – 36.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure model-ing: Sensitivity to underparameterized model misspecification. Psycho-logical Methods, 3, 424 – 453.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 6(1), 1 – 55.

Schermelleh-Engel, K., Kerwer, M., & Klein, A. G. (2014). Evaluation of model fit in nonlinear multilevel structural equation modeling. Frontiers in Psychology, 5, 181.