Creativity and memory: Cortical representational change along with amygdala activation predict the insight memory effect

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 the 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 and analysed the data. M.B., T.S. & R.C. wrote the manuscript and contributed to revising the manuscript.

Abstract (251 words)

Creative solutions often involve *insight*, which refers to a rapid reorganisation of internal representations (representational change - RC) combined with an “Aha” feeling including positive emotions and certainty about the solution. Although there is evidence that problems eliciting insight are better remembered, the neural mechanisms of this “insight memory effect” (IME) remain unclear. Given that insight involves sudden RC, positive emotions and certainty, we hypothesised that the IME would be associated with a change in cortical areas representing the solution (RC) and increased activity in the amygdala. While under-going fMRI scanning, participants identified obscure (Mooney) images of real-world objects, whose identification (solution) is often accompanied by insight. Five days later, participants’ memory for the real-world objects (solutions) was tested. The study yielded four main findings. First, multivariate pattern analyses showed that when solving Mooney images was accompanied by insight, posterior fusiform gyrus (pFusG) and inferior lateral occipital cortex (iLOC) showed sudden RC. Second, at the time of the solution, the amygdala - a region associated with positive emotions - showed increased activity as a function of insight. Third, insight was also associated with increased functional connectivity between amygdala and the aforementioned visual areas (pFusG and iLOC). Most importantly, these three effects predicted subsequent memory (i.e., the IME). This is the first study providing direct empirical evidence of sudden RC during insight problem-solving and how this reorganisation together with the amygdala enhances long-term memory formation. These findings highlight the importance of fostering self-motivated and insightful problem-solving in educational environments to promote better learning outcomes.

Importance statement

Based on prior research, we hypothesised that the neural mechanism of better memory for solutions accompanied by insight (insight memory effect) consists of the neural substrates eliciting representational change and brain regions associated with positive emotions and surprise.

Consistent with our hypotheses, we found that cortical representational change (visual areas given the visual stimuli) before the insight solution was coupled with amygdala activity and amygdala-visual cortex connectivity, and all these changes predicted the insight memory effect.

These findings provide missing evidence regarding the presumed internal reorganisation of insightful problem solving and why it enhances memory for the problem.

Introduction

Creativity and memory have a synergistic relationship. On one hand, finding creative solutions depend on having relevant preexistent knowledge, and, on the other hand, novel solutions are typically well-remembered, promoting new knowledge of the world and future adaptation. *Insight* is a special phenomenon of creativity often studied in the context of scientific discoveries (Kounios & Beeman, 2015) where a non-obvious problem is suddenly solved via a novel solution approach by reorganising preexistent knowledge structures (Danek et al., 2020; Dietrich & Kanso, 2010). This reorganisation of internal (conceptual or perceptual) representations known as *representational change* (RC) is assumed to lead to a sudden solution that is often followed by the solver's conviction that the solution is correct, sudden and it is associated with a positive emotion (Danek & Wiley, 2017; Kizilirmak & Becker, 2022). Compared to being provided the solution, finding the solution by oneself is typically associated with enhanced memory for the solution (Bertsch et al., 2007; Burns, 1992; Danek & Wiley, 2020; Dominowski & Buyer, 2015; McCurdy et al., 2020; Rowland, 2014; Wills et al., 2000). Over and above this effect, if the solution is accompanied by insight, subsequent long-term memory is further boosted, sometimes after a single experience, unlike other forms of learning requiring multiple repetitions (Ash et al., 2012; Auble et al., 1979; Danek et al., 2013; Kizilirmak, Galvao Gomes da Silva, et al., 2016; Kizilirmak, Thuerich, et al., 2016; Ludmer et al., 2011). Hence, the study of insight is not only relevant in the context of scientific discoveries, but it can shed light on mechanisms of rapid learning and memory formation and has practical implications in educational settings emphasising the importance of self-motivated problem-solving accompanied by insights (Gruber et al., 2014; Ludmer et al., 2011).

What explains this *insight memory effect* (IME)? IME has been attributed to the two aforementioned aspects of insight: (1) the cognitive component represented by RC, and (2) the metacognitive ("Aha") component, represented by the subjective positive emotions but also certainty and suddenness (Ash & Wiley, 2008; Auble et al., 1979; Danek & Wiley, 2020). Most evidence regarding these two components is behavioral,

and hence, the neural mechanisms of the IME remain largely unknown (Kizilirmak et al., 2019).

Regarding the cognitive component of insight, RC is assumed to refer to changing the pattern of currently activated cell assemblies encoding the current problem (Becker, Cabeza, et al., 2022; Knoblich et al., 1999). For example for visual problems requiring object recognition, RC refers to breaking up perceptual chunks into their constituent elements to make them available for a novel recombination (Tang et al., 2015). In fMRI studies, this process can be measured as changes in the spatial distribution (pattern) of activity in modality-specific cortices coding for problem components (Haxby et al., 2001; Kriegeskorte et al., 2008; Norman et al., 2006). Memory-related changes in activation patterns of object representations has been investigated by multivariate fMRI studies before (Clarke et al., 2016; Schapiro et al., 2012; Schlichting et al., 2015) finding specifically changes in ventral occipito-temporal cortex (VOTC) including Fusiform Gyrus (Clarke et al., 2016; Schlichting et al., 2015). Importantly however, there is no study investigating insight-related sudden changes in content representations during problem solving nor whether this RC mechanism contributes to the IME.

Regarding the metacognitive component of insight, several studies have linked insight to regions associated with positive emotions, reward and surprise, such as the amygdala (Janak & Tye, 2015; Kizilirmak, Thuerich, et al., 2016; Oh et al., 2020; Shen et al., 2017; Tik et al., 2018). Ludmer et al. (2011) and Kizilirmak et al. (2019) found amygdala activity to predict better memory for sudden solutions in insight tasks but without relating brain activity to measures of insight. Hence, no evidence is currently available on whether and how the amygdala is involved in *insight*-related better subsequent memory. Given that the amygdala has been shown to enhance memory by enhancing learning-related sensory stimulus representations (Chavez et al., 2009), this region may also enhance content representations of the solution in cortical RC areas facilitating insight-related subsequent memory. We hypothesized that insight and the IME will be associated with increased activity in the amygdala, in close collaboration with VOTC regions showing RC.

To investigate the neural mechanisms of insight and the IME, we asked participants to perform a *visual insight task* during fMRI scanning. In this task, 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). Because of the ambiguous figure/ground distinction in Mooney images, black/white contours are often falsely perceptually grouped to meaningless shapes hindering successful object identification. Rechunking or perceptually regrouping the black/white contours can lead to sudden object identification eliciting insight (Kizilirmak, Thuerich, et al., 2016; Ludmer et al., 2011). Given prior neuroscience evidence of object recognition and visual RC, we assume that this perceptual regrouping occurs in VOTC regions involved in visual object identification, such as the inferior Lateral Occipital Cortex (iLOC), the Fusiform Gyrus (FusG), and the Inferior Temporal Gyrus (ITG) (Clarke et al., 2016; DiCarlo et al., 2012; Schlichting et al., 2015). Further, we assumed that successful perceptual regrouping of the black/white contours (=RC) should be associated with a sudden conceptual update of the Mooney object in these regions.

Following Mooney object identification, participants provided ratings of the perceived suddenness, emotion, and certainty of 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.

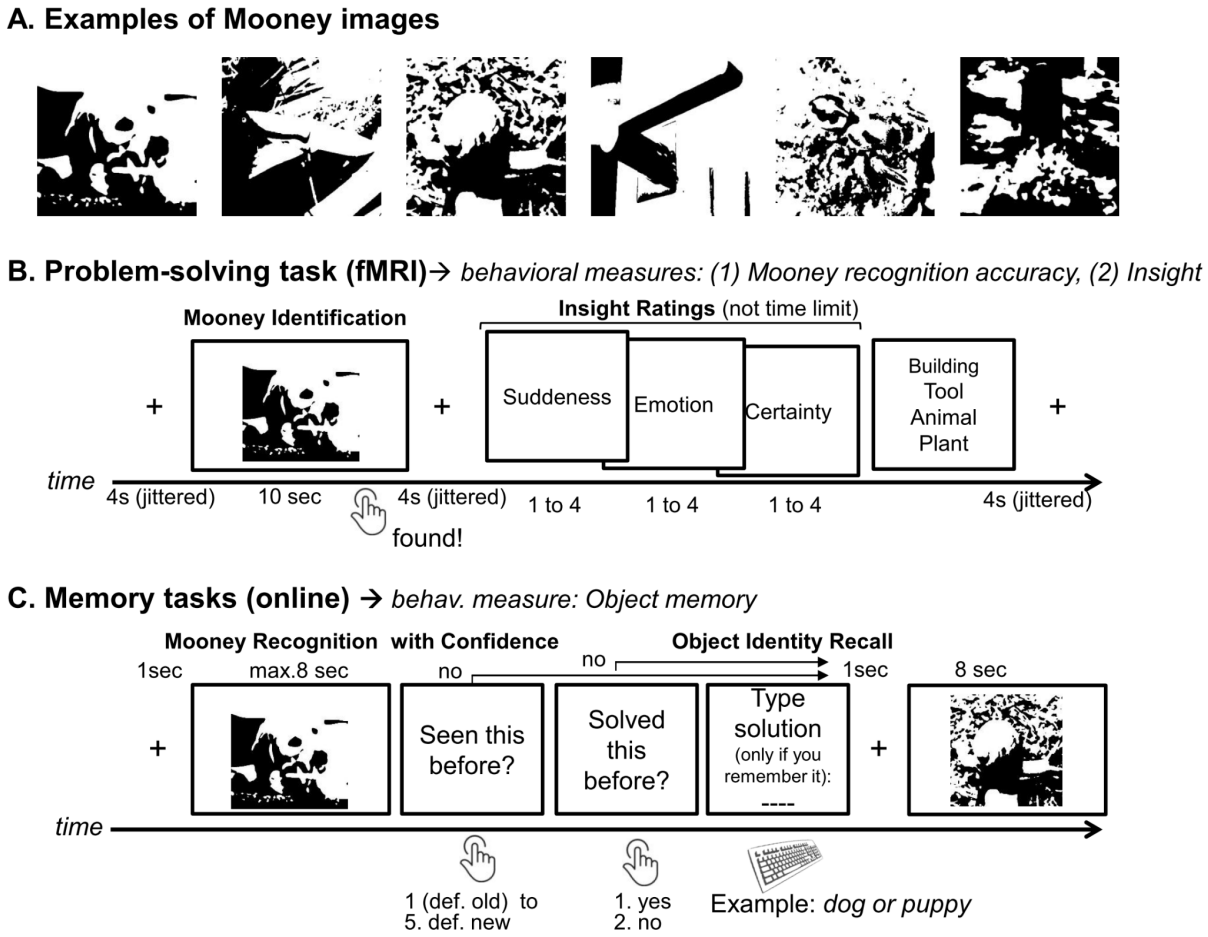

***Figure 1**. Experimental design*

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.

Given prior neuroscience evidence, we made four fMRI predictions. First, *VOTC displays insight-related RC*. We measured RC during Mooney object identification by comparing multivariate activity patterns before and after the solution was found, and determined if the change in patterns was greater for high than low insight trials. We focused on six VOTC regions-of-interest (ROIs): posterior FusG (pFusG), anterior FusG (aFusG), inferior LOC (iLOC), as well as anterior ITG (aITG), mid ITG (mITG), and posterior ITG (pITG). Second, *the amygdala shows insight-related activity*. To test this hypothesis, we identified univariate amygdala activity at the moment the participants made an identification response, and determined if the activity was greater for high than low insight trials. Third, *VOTC and the amygdala display insight-related functional connectivity*. This hypothesis is based on the assumption

that insight-related VOTC-RC and amygdala activity are interrelated: VOTC-RC leading to the solution elicits an emotion/surprise-related amygdala activity (feedforward), which in turn strengthens visual representations in VOTC (feedback). We determined if VOTC-amygdala connectivity was stronger for high than low insight trials. Finally, and most importantly, we hypothesised *all three insight-related effects (VOTC-RC, amygdala activity, and VOTC-amygdala connectivity) predict the IME*. The IME was identified by behavioural and neural differences between problems solved with 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 three insight ratings (*positive emotion* [M=2.8, SD=.41], *suddenness* [M=3.0, SD=.39] and *certainty* [M=3.14, SD=.34]; 4-point Likert scales) following Mooney object identification (Danek & Wiley, 2017, 2020). However, due to their significant factor loadings onto a latent insight 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 more conservative *Insight cut-off* measure, where a trial was categorized as “high insight” if the solution was rated as a 3 or 4 on all three scales, and as “low insight” otherwise. 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 in 51.8% (SD=12.8%) of these cases correct. Of all correctly identified Mooney images, 65.2% (SD=6.6%) were solved with high insight and 34.7% with low insight. 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% 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 has been shown to covary with the difficulty of Mooney images (Becker et al., 2021). Because we observed a systematic difference in response times between the *HI-I* and *LO-I* condition but were not interested in differences in difficulty, response time was included as covariate of no interest into all subsequent analyses if appropriate.

Subsequent memory of each Mooney object identification was measured seven 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.

To investigate the behavioral IME, 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 the respective run order (1-4). This insight factor (*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 the IME. This IME effect remained significant when controlling for response time ($Chi^2(1)$=56.96, *p*<.001, odds ratio: *HI-I*=2.17).

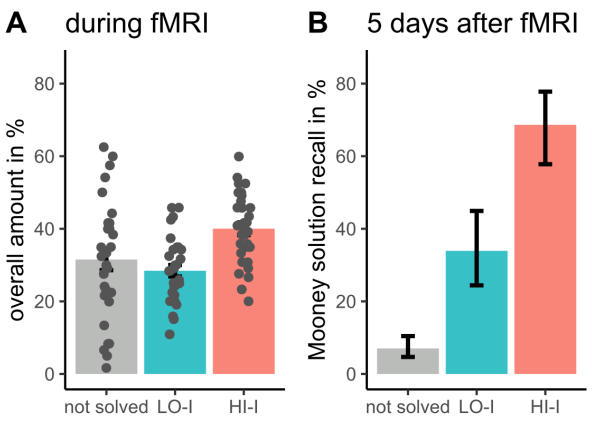

Figure 2. Distribution of events and behavioural Insight Memory Effect.

Note. *HI-I* = solved trials with high insight; *LO-I* = solved trials with low insight; **Panel A**. overall proportion of solved trials divided by condition; **Panel B.** Behavioural insight memory effect; estimates are marginal means from the general linear mixed model without response time as covariate. Error bars represent 95% confidence intervals.

Hypothesis 1: VOTC displays 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 and 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).

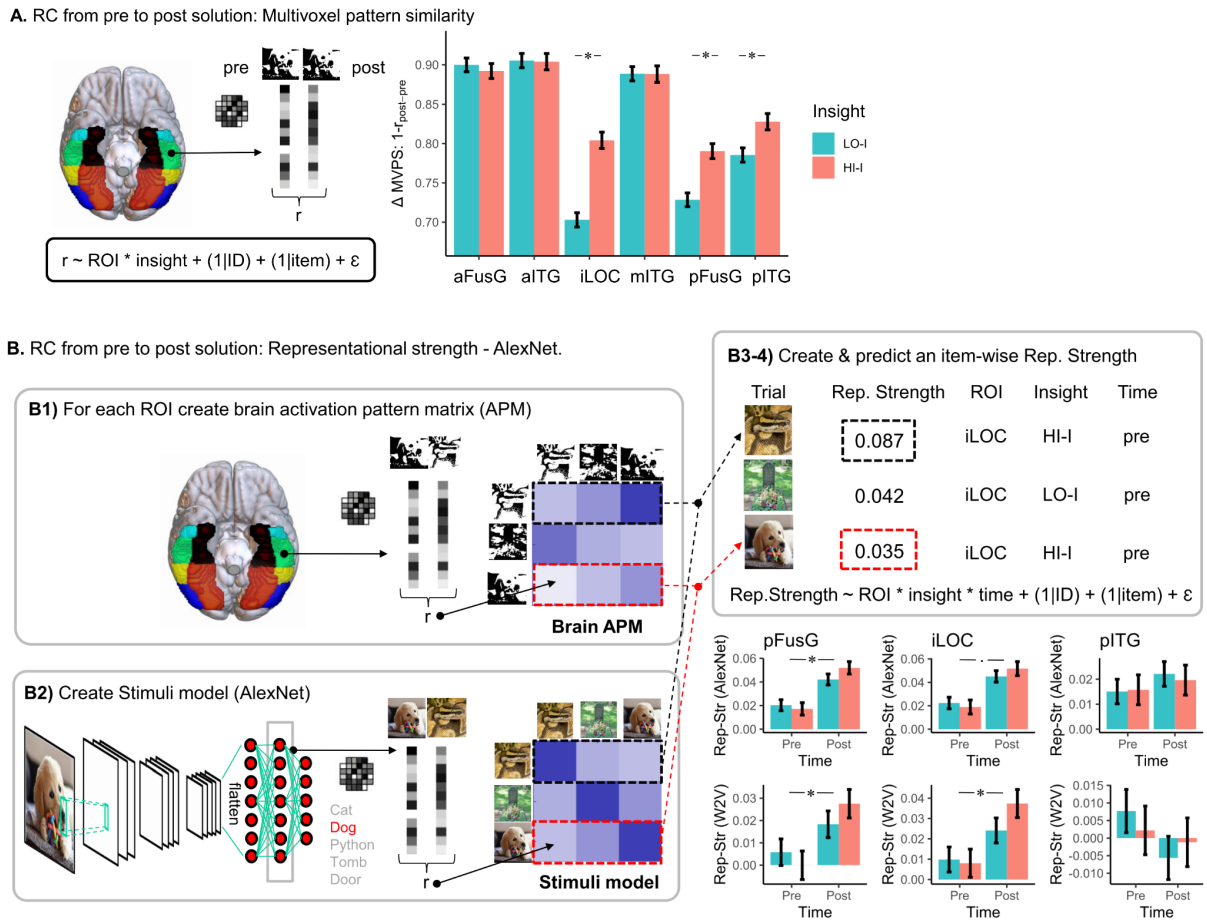

Figure 3. Identification of VOTC regions showing Representational Change (RC)

Note. Values = estimated marginal means ± SEM; asterisk = statistical significance; dot represents trend without statistical significance. *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. (**B1**) A brain activation pattern matrix (APM, 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. (**B2**) A conceptual *stimuli model* (here using AlexNet, size 120x120) is generated where each cell is representing a similarity value for each Mooney object pair. (**B3**) For each brain region and each time point, the row of each stimulus (~120) in the *stimuli model* and in the APM are correlated yielding a representational strength measure (i.e. brain-model fit) per region and time point. (**B4**) 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 *stimuli models* (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 and the run order. The interaction between ROI and insight (*HI-I* vs. *LO-I)* significantly explained changes in pre- and post solution activity patterns ($Chi^2(5)=267.20$, $p<.001$, $ß=-.08$) (Fig.3-A). Posthoc analyses revealed that bilateral iLOC (z=-14.80, *p-Bonferroni*<.001), bilateral pFusG (z=-12.54, *p-Bonferroni*<.001) as well as bilateral pITG (z=-6.19, *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 (p>.11).

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 and the respective run order (1-4).

The model with the three-way insight condition*time*ROI interaction significantly explained variance in representational strength ($Chi^2(6)=41.91$, $p<.001$). Posthoc

analyses revealed a significant insight condition*time interaction effect for bilateral pFusG (*Chi*²(1)=7.87, *p-Bonferroni*<.05) while bilateral iLOC (p>.11) and pITG (p>.60) did not reach significance. Note, using the more conservative insight measure (*insight cuff-off*), iLOC did show a significant condition*time interaction in the hypothesised direction: *Chi*²(1)=9.09, *p-Bonferroni*<.01) (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 three-way insight condition*time*ROI interaction was significant (Chi²(6)=73.91, *p*<.001). Posthoc analyses revealed a significant condition*time interaction effect for bilateral pFusG (*Chi²(*1)=12.49, *p-Bonferroni*<.01) and bilateral iLOC (*Chi²(*1)=6.46, *p-Bonferroni*=.033) 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 (to a lesser degree) 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 metacognitive component of insight, namely the positive emotion, certainty, and suddenness associated with the insight experience. In line with previous research, we hypothesised that this metacognitive component is mediated by the amygdala.

Hypothesis 2: the amygdala shows insight-related activity

To examine our second hypothesis, that the metacognitive component of insight is mediated by the amygdala, we compared amygdala activity during Mooney object identification in *HI-I* vs. *LO-I* trials. Note, we used univariate activity because the amygdala is 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 activity was 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). To control for differences in response time 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 as covariate of no interest which did not change the significance of the results (see Supplementary Material).

As illustrated by Fig.4-A, amygdala activity significantly distinguishes between *HI-I* and *LO-I* trials during solution ($Chi^2(1)=10.99$, $p<.001$). Posthoc analyses revealed that Amygdala activity was significantly higher for *HI-I* compared to *LO-I* trials ($t(215)=3.342$, $p=.001$, ß= .28). To investigate potential smoothing and normalisation artefacts in the small amygdala ROI, we also conducted an ROI analysis in native space (see Supplementary Material) and the *HI-I>LO-I* difference in amygdala remained significant.

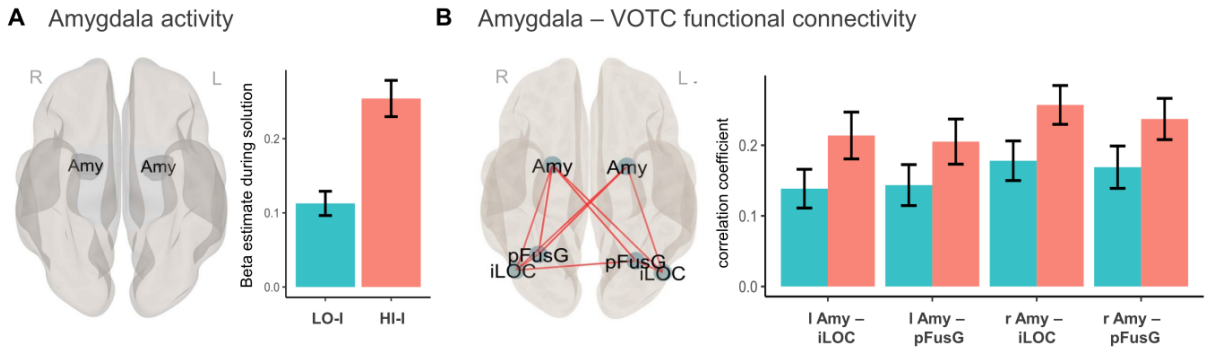

Figure 4. Amygdala activity during solution and amygdala VOTC-RC functional connectivity between during stimulus presentation predicting insight

Note. Values = estimated marginal means ± SEM. (l,r) Amy = Amygdala (left, right); **Panel A**: Amygdala activity during solution divided by condition (HI-I, LO-I); **Panel B:** red = *HI-I* = solved trials with high insight; *turquoise*= solved trials with low insight. Values represent marginal means of correlation coefficients between right or left Amygdala and averaged left and right iLOC or FusG from bilateral Amygdala VOTC-RC network split by condition.

Hypothesis 3: Amygdala and VOTC show insight-related functional connectivity

Due to the theoretical relationship between the cognitive (RC) and metacognitive components of insight, we hypothesized that the functional relationship between the amygdala and VOTC-RC areas should be higher for *HI-I* compared to *LO-I* trials. Therefore, we created a network consisting of bilateral amygdalae, pFusG and iLOC (i.e., 6 ROIs) during stimulus presentation (ten seconds block while participants tried

to identify the Mooney images) and computed functional connectivity (FC) between each pair of these areas as well as 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 left and right hemisphere but FDR-controlled for multiple comparison. Within this network, left (Mass=13.22, *p-FDR*=.019) and right (Mass=16.57, *p-FDR*=.018) amygdalae were significantly more strongly functionally connected to VOTC-RC regions during insight (*HI-I>LO-I)*. In particular, left amygdala was more strongly functionally connected to left ($t(30)=2.45$, $p=.01$) and right ($t(30)1.95$, $p=.03$) iLOC and right pFusG ($t(30)=1.84$, $p=.038$) during insight. Right amygdala was also significantly connected to left ($t(30)=2.16$, $p=.019$) and right ($t(30)=2.23$, $p=.017$) iLOC as well as left ($t(30)=1.80$, $p=.041$) and right ($t(30)=1.93$, $p=.032$) pFusG (see Fig.4-B). The statistics with all connections are listed in the Supplements, Table S2.

Hypothesis 4: All three insight-related effects (VOTC-RC, amygdala activity, and VOTC-amygdala connectivity) predict the IME

Having linked insight to RC in VOTC, increased amygdala activity, and enhanced VOTC-amygdala connectivity, we examined whether each of three insight effects also predicted better subsequent memory (IME).

Insight-related RC predicts the IME.

First, we tested whether insight-related better memory is associated with 1) the changes of distributed activity patterns (i.e., decrease in multivoxel pattern similarity) from pre to post solution and 2) representational strength of the object during post solution in those regions that show patterns of RC (pFusG, iLOC). For this, we used the ordered insight-memory-factor that included subsequently remembered *HI-I* trials (*Rem_HI-I*), subsequently remembered *LO-I* trials (*Rem_LO-I*) and subsequently forgotten (*Forgotten*). Subsequently forgotten trials were not subdivided into *HI-I* and *LO-I* because many participants produced too few *HI-I* forgotten Mooney images. We examined if, 1) and 2) varied as a function of these three conditions and the ROI

controlling for response time, run order (1-4) and random subject and item intercepts.

1) RC from pre to post solution: Multivoxel pattern similarity. The linear trend of the insight-memory-factor significantly accounted for a general reduction in pre to post solution activity pattern similarity ($Chi^2(2)=41.77$, $p<.001$, ß=.12) in both ROIs suggesting that insight-related subsequent memory predicts the amount of change in pre to post activity pattern similarity. The interaction between the insight-memory-factor and ROI was also significant ($Chi^2(2)=10.41$, $p<.01$) suggesting a difference between both ROIs. Post-hoc analyses revealed that pre to post solution activity pattern similarity for *Rem_HI-I* was reduced compared to *Rem_LO-I* in iLOC ($z=-4.16$; $p=.001$) but only trendwise in pFusG ($z=-2.16$ $p=.07$) and while *Rem_HI-I* was significantly reduced to *Forgotten* in both ROIs: pFusG ($z=-6.46$; $p<.001$), iLOC ($z=-3.91$; $p<.001$) (see Fig.5-A). *LO-I* and *Forgotten* events did not significantly differ from each other in both ROIs: pFusG ($p>.29$) and iLOC ($p>.40$).

2) Representational strength at post solution. The linear trend of the insight-memory-factor significantly accounted for the correlation of activity pattern over both brain regions (pFusG, iLOC) with the conceptual AlexNet model during solution ($Chi^2(2)=9.66$, $p<.01$, ß=.07), suggesting that the representational strength of the Mooney object during post solution is associated with insight-related better memory (see Fig.5-B). There was no insight-memory-factor*ROI interaction ($p>.67$) indicating no significant differences between both ROIs regarding the relationship between the insight-memory factor and representational strength of the solution object. Posthoc analyses revealed that the representational strength of the solution object in *Rem_HI-I* trials was significantly higher than for *Forgotten* trials ($z=-3.11$, $p<.01$) while Rem_LO-I trials did not differ from *Forgotten* trials ($p>.35$). However, representational strength of the solution between *Rem_HI-I* and *Rem_LO-I* trials did not differ significantly ($p>.34$).

Those results were replicated when using the Word2Vec instead of the AlexNet conceptual model: the linear trend of the insight-memory-factor was significant ($Chi^2(2) = 12.39$, $p<.01$) over both brain regions (pFusG, iLOC) suggesting that this linear trend is robust. There was also no indication of an insight-memory-factor*ROI interaction ($p>.57$) suggesting that both brain regions do not contribute differently to

the relationship between the insight-memory-factor and the representational strength of the Mooney object during post solution (see Fig.5-C). Posthoc analyses revealed that the representational strength of the Mooney object in *Rem_HI* (z=-3.48, p<.01) but not *Rem_LO-I* (p>.100) differed from *Forgotten* trials. However, *Rem_HI-I* and *Rem_LO-I* trials did not differ significantly (p>.55).

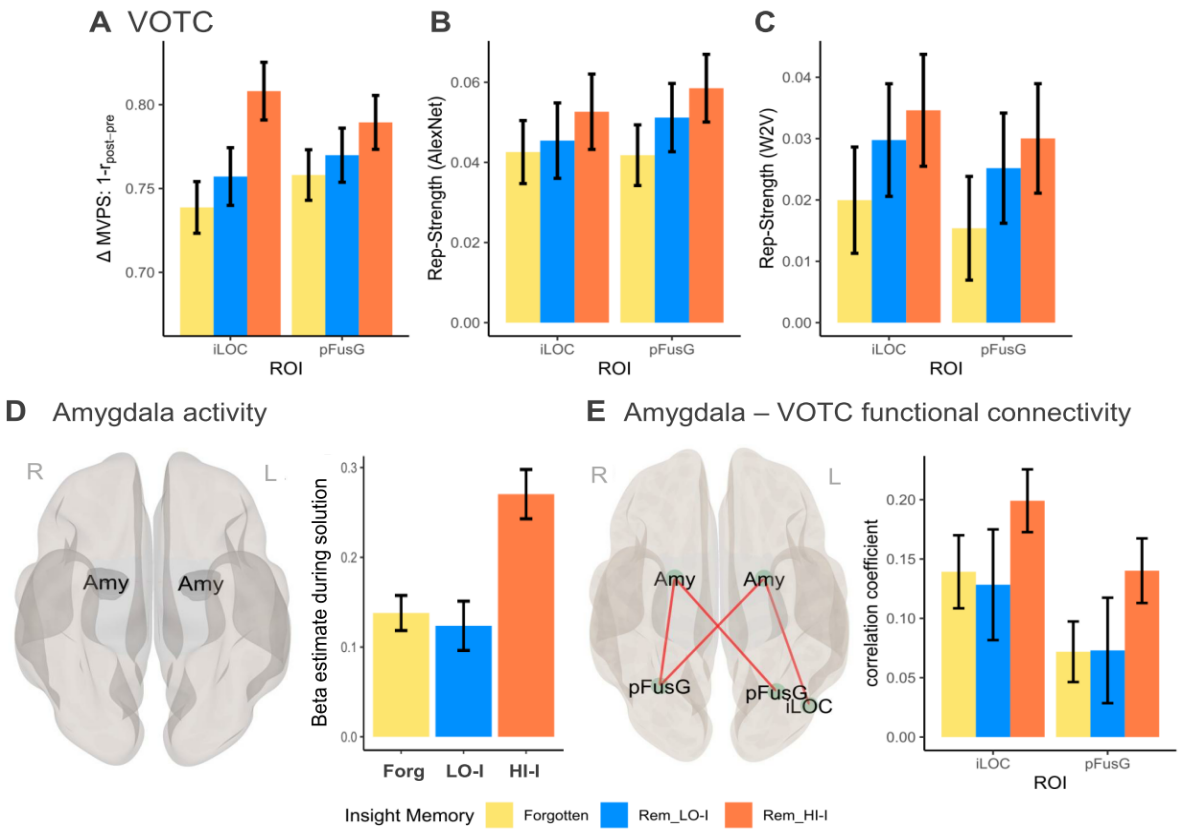

Figure 5. Insight memory factor predicts multivariate activity in VOTC, univariate activity in amygdala and Amygdala-VOTC functional connectivity.

Note. Values represent estimated marginal means ± (between subject) SEM. *Forgotten* = subsequently forgotten trials; *Rem_LO-I:* subsequently remembered trials originally solved with low insight; *Rem_HI-I* = subsequently remembered trials originally solved with high insight. **Panel A:** 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:** Representational strength of solution object (post solution) is measured via a conceptual model created out of the penultimate layer of AlexNet. **Panel C:** Representational strength is measured via a conceptual Word2Vec [W2V] model. **Panel D:** Amygdala mean activity at post solution divided by insight memory conditions. **Panel E:** Functional connectivity between Amygdala and VOTC. Values represent averaged correlation coefficients between left and right amygdala and left and right iLOC or pFusG.

Insight-related amygdala activity predicts the IME

Next, to examine the role of the amygdala in the IME, we contrasted amygdala activity during solution for the ordered insight-memory-factor (*Rem_HI-I>Rem_LO-I>Forgotten*) (see Fig.5-D). This analysis revealed a significant condition effect ($Chi^2(2)=25.34$, $p<.001$, ß=.33). Posthoc analyses revealed that amygdala activity was significantly higher for *Rem_HI-I* compared to *Rem_LO-I* ($t(338)=4.69$, $p<.001$) and *Forgotten* ($t(338)=4.09$, $p<.001$) while *Rem_LO-I* and *Forgotten* trials did not significantly differ ($p>.81$). For exploratory reasons, this ordered factor was also examined in VOTC areas which did not reveal any effect (see Supplementary Material).

Exploratory whole-brain analysis of the ordered IME factor. To explore other brain regions that predicted IME besides the amygdala, we also performed a whole brain analysis comparing *Rem-HI, Rem-LO-I*, and *Forgotten* trials. This analyses yielded multiple regions where activity predicted the IME (see Supplementary Material, Table S1, Fig.6), including areas belonging to the Default Mode Network (DMN) such as the Medial Prefrontal Cortex (mPFC; Peak voxel: mni_{xyz}[6 66 26], $T=6.30$, $p=.001$), bilateral Hippocampus (Peak voxel: mni_{xyz}[-29 -21 -21], $T=5.30$, $p=.033$) but also bilateral Nucleus Accumbens (Peak voxel: mni_{xyz}[6 10 9], $T=6.21$, $p=.002$) and bilateral Middle Temporal Gyrus (Peak voxel: mni_{xyz}[-55 -3 -21], $T=5.33$, $p=.030$).

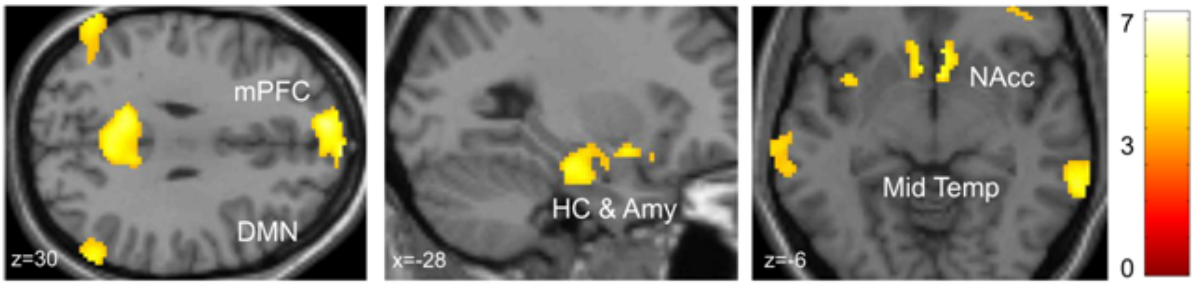

Figure 6. Significant clusters from exploratory whole brain analysis supporting the insight memory effect.

Note. Results stem from *Rem-HI>Rem-LO-I>Forgotten* 2nd level contrast. Voxelwise threshold = $p<.0001$; clusterwise-threshold p-FWE<.05. DMN = Default Mode Network; mPFC = Medial Prefrontal Cortex; HC=Hippocampus; Amy = Amygdala; NAcc = Nucleus Accumbens; Mid Temp= Middle Temporal Gyrus.

Insight-related amygdala-VOTC connectivity predicts the IME

Finally, to examine whether insight-related functional connectivity between amygdala and VOTC-RC regions predicts the IME, we estimated functional connectivity between amygdala activity and pFusG and iLOC during stimulus presentation as a function of the ordered insight-memory-factor (*Forgotten>Rem_LO-I>Rem_HI-I*) (see Fig.5-E). Note, due to similar functional connectivity results in the contrast *HI-I>LO-I*, both amygdala hemispheres were averaged. Functional connectivity between averaged amygdala and the left (*ß*=.06; *t*(30)=2.12, *p*=.043) and right (*ß*=.08; *t*(30)=2.83, *p*=.0082) pFusG as well as the left iLOC (*ß*=.06; *t*(30)=2.12, *p*=.043) but not right iLOC (p>.13) significantly increased as a function of the contrast *Rem_HI-I>Rem_LO-I>Forgotten*.

Discussion

We investigated the neural basis of (visual) insight and the insight memory effect by scanning participants with fMRI during the identification of Mooney images. The objects in those black/white abstracted images are difficult to identify due to their ambiguous figure/ground distinction which hinders the appropriate grouping of perceptual elements. Regrouping those perceptual elements, i.e. reorganising internal (conceptual or perceptual) representations known as representational change (RC) can sometimes lead to sudden object recognition (solution), followed by a feeling of positive emotion and suddenness. One week later outside of the scanner, we tested memory for these Mooney object identification trials. We examined the neural mechanisms of RC and the enhancing effect of insight on subsequent memory (insight memory effect—IME). We hypothesised that insightful solutions are associated with greater RC in ventral occipito-temporal cortex (VOTC), boosted amygdala activity, and enhanced amygdala-VOTC connectivity, and that all these three effects would contribute to the IME. Consistent with these hypotheses, the study yielded four main findings: (1) multivariate pattern analyses showed insight-related RC in VOTC regions; (2) univariate analyses revealed that amygdala activity was increased during high insight trials; (3) functional connectivity analyses demonstrated that functional connectivity between VOTC and the amygdala was also enhanced by insight; (4) finally, all three foregoing effects predicted the IME.

These findings provide novel evidence regarding the neural mechanisms of insight and the IME in the visual domain, and they are discussed below, organised according to brain region or analysis type.

VOTC regions showed insight-related RC

This study provided novel evidence for a neural mechanism of insight-related RC in cortical areas. We assessed this mechanism by comparing multivariate activity patterns before and after the solution, under the assumption that the before-after representation difference should be more pronounced for solutions accompanied by high insight. Given the visual nature of Mooney images and underlying objects, we investigated the ventral visual pathway for object processing. This pathway shows a posterior-anterior gradient with early visual cortex mediating the analysis of simple visual features, VOTC regions (pFusG, iLOC, pITG and mITG) processing the conjunction of visual features into integrated visual objects, and the anterior temporal lobes (ATL: aFusG and aITG) processing abstract concepts and semantic associations (Bussey et al., 2005; Clarke & Tyler, 2014). We limited our analyses to mid-level visual processing in VOTC and high-level visuo-semantic processing in ATL where object processing is sufficiently complex to allow for a potential regrouping of visual/conceptual features. We found significant insight-related RC in VOTC, in particular in pFusG and (less consistently) iLOC, but not in ATL. Although the latter finding could reflect the typically lower signal-to-noise ratio (SNR) in fMRI measures of ATL (due to air-tissue boundaries, see Embleton et al., 2010), it could also reflect a differential contribution of VOTC and ATL to visual insight. For example, it is possible that insight during Mooney object identification is more related to perceptual RC (e.g., perceptual re-chunking of Mooney elements into integrated objects, such as the shape of a tiger) than to abstract semantic RC (e.g., new links to concepts and semantic associations, such as “tiger is related to lion”; (Peelen & Caramazza, 2012; Simmons & Barsalou, 2003). Although our RSA analyses included a *stimuli model* where conceptual similarity is based on co-occurrences in text data (Word2Vec), the insight-related RC effects for this model were found in VOTC rather than ATL. It is possible that insight-related RC effects can be found in ATL for other insight tasks, such as those involving verbal stimuli. However, besides

the specific brain regions involved, the important point is that this study reports the first evidence for a potential neural mechanism of insight-related perceptual RC.

What remains an open question is what causes insight-related RC in VOTC. As mentioned above, the function of VOTC suggests that perceptual RC involves re-chunking visual elements of the Mooney image into an integrated shape. This could be driven primarily by a relatively automatic, bottom-up process. For example, there is evidence that pyramidal cells in cortical layer 5 create a bursting signal when they receive matching information from a feature-specific external input (i.e. Mooney elements) and context (memory/knowledge of a particular object) and communicate a match to the rest of the brain (Larkum, 2013). Alternatively, re-chunking visual elements of the Mooney image could also be a top-down process driven by a control structure systematically biasing those firing patterns in pFusG (and iLOC) in VOTC (Becker, Sommer, et al., 2020a; Tang et al., 2015; Wu et al., 2013).

We also found evidence that insight-related RC in VOTC contributes to IME. In particular, the representational strength of the identified solution object increased from subsequently forgotten over remembered objects originally solved with low to high insight. The link to the IME is not as strong as for amygdala activity, which is discussed next.

The amygdala is strongly linked to the IME

While the current study replicated enhanced amygdala activity during insight (Kizilirmak, Thuerich, et al., 2016; Shen et al., 2017), another novel finding is the clear link between amygdala activity and the IME. Amygdala activity at the time of the solution was significantly higher for subsequently remembered objects solved with high- vs. with low-insight, and significantly higher for both of these objects than for subsequently forgotten objects. Amygdala activity had been associated with better subsequent memory in different insight paradigms before (Kizilirmak et al., 2019; Ludmer et al., 2011). However, those studies did not relate the amygdala activity predicting subsequent memory to measures of insight. Hence, the present results showing that amygdala activity is involved in insight and specifically *insight*-related better memory provide stronger evidence of the role of the amygdala in the IME.

The amygdala has been long known to enhance memory for emotional information (Dolcos et al., 2004; Hamann, 2001; Vuilleumier et al., 2004), primarily through modulatory processes (Dolcos et al., 2012). There is evidence that in response to emotional arousal, amygdala neurons increase their firing rate, facilitating neural plasticity and memory consolidation in other brain regions, such as the HC (Fujise et al., 2017; Huff et al., 2013; McGaugh, 2004; Popescu et al., 2007; Roozendaal et al., 2009). The amygdala is highly functionally interconnected with the HC but also projects to all levels of the ventral stream (Freese & Amaral, 2005) enhancing sensory representations both via direct connections to sensory areas (Freese & Amaral, 2005; Gschwind et al., 2012); as well as via indirect pathways involving brainstem neuromodulatory systems and frontoparietal regions (Janak & Tye, 2015; Lim et al., 2009; Pourtois et al., 2013). In particular, stronger amygdala-fusiform coupling during successful encoding of emotional memories has been observed (Gieske & Sommer, 2023; Schümann & Sommer, 2018). Building on these findings, we hypothesize the sudden solution (RC) during insightful problem solving elicits heightened emotional arousal in the amygdala enhancing visual object representations in VOTC facilitating memory encoding and/or consolidation. Our results show increased functional connectivity between amygdala and pFusG as well as left iLOC predicting the IME supporting this assumption.

Additionally, the amygdala modulates memory processes by projecting to basal forebrain cholinergic systems (e.g., mPFC) and midbrain dopaminergic reward systems (NAcc) (Janak & Tye, 2015; McDonald, 1998; Packard et al., 1994; Roozendaal & McGaugh, 1996). Next to HC, NAcc and mPFC, as part of the DMN, also exhibited increased activity linked to the IME and are highly interconnected with the amygdala (Janak & Tye, 2015). Thus, activity in these regions predicting IME are likely functionally related to amygdala activity and will be discussed next.

Neural mechanism of IME – other brain regions

Increased HC activity has been predictive of insight (Becker, Sommer, et al., 2020b; Kizilirmak, Thuerich, et al., 2016; Luo & Niki, 2003; Shen et al., 2017; Tik et al., 2018) and was consistently related to episodic long-term memory formation (Burgess et al., 2002; Moscovitch et al., 2016; Squire et al., 2004). The NAcc is a primary

reward region (Ikemoto, 2007) that has also been found to be predictive for insight (Tik et al., 2018). Amygdala projections to the NAcc modulate not only stimulus-reward associations (Cador et al., 1989; Everitt et al., 1989; Janak & Tye, 2015) but also interact with HC effects on memory consolidation via this pathway (Roozendaal et al., 2001).

The DMN, a network that mediates spontaneous cognition, has not only been causally linked to idea generation and creative thinking, i.e. the ability to link seemingly unrelated elements (Shofty et al., 2022). The DMN is also thought to contribute to flexible retrieval of memories (Andrews-Hanna et al., 2010) and mediate interactions between HC and neocortex for consolidation of new memories (Kaefer et al., 2022). Consistently, the DMN has been found to predict the IME in verbal tasks (Kizilirmak et al., 2019). Moreover, the mPFC, as part of the DMN, plays a pivotal role in schema-related enhanced memory encoding (Sommer, 2017; Sommer et al., 2022; Van Kesteren et al., 2012). This process involves anchoring novel information to preexisting knowledge (schemas), thereby facilitating later retrieval (Tse et al., 2011). Consequently, during moments of insight, the mPFC may detect a profound alignment between new information (e.g., black/white contours of the Mooney image) and existing knowledge (e.g., the underlying object category) (Kizilirmak et al., 2019).

Finally, exploratory analyses also revealed increased activity of bilateral Middle Temporal Gyrus as a function of the IME which could be interpreted as insight promoting enhanced semantic processing of the cue which additionally helped later remembering (Davey et al., 2015; Wei et al., 2012).

Limitations

One limitation of the study relates to the subsequent memory measure of the Mooney objects. Even though subjects were specifically instructed not to identify the Mooney images again but just to indicate whether they remember the solution object, we cannot exclude the fact that the subjects attempted to re-solve those visual problems (however, recognition time for old Mooney images in the subsequent memory test was approx. 1.9sec shorter than average response time to identify the Mooney images in the scanner). Hence, the presented results may reflect a mixture

of insight-related better memory and increased ability to correctly identify the Mooney images again.

Another limitation is that fMRI-based functional connectivity is not informative regarding the direction of the amygdala-VOTC connectivity. Thus, it is unclear if the insight-related increase in functional connectivity and its impact on IME reflects sudden RC in VOTC triggering an emotional response in the amygdala, or a modulatory effect of amygdala on RC in VOTC. Our hypothesis is that both effects are synergistic. Future studies need to further investigate the causal directions between those functionally different but interrelated brain areas.

Conclusion

This work enhances our mechanistic understanding of insight-related reorganisation of internal representations, i.e. RC and the role of the amygdala in the IME. We showed that the IME is associated with (1) the representational strength of the solution object in VOTC showing RC, (2) insight-related amygdala activity reflecting positive emotion, certainty and suddenness, as well as (3) increased functional connectivity between amygdala and these cortical areas. Given those results, we have proposed that the amygdala is involved in binding content representations in long-term memory by enhancing solution-related representations in FusG and iLOC after solution. Furthermore, this finding has practical implications for the field of education, highlighting the importance of creating an environment that cultivates not only self-motivated learning but also fosters insightful problem-solving, ultimately leading to improved learning outcomes.

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.

Andrews-Hanna, J. R., Reidler, J. S., Huang, C., & Buckner, R. L. (2010). Evidence for the default network's role in spontaneous cognition. *Journal of Neurophysiology*, *104*(1), 322–335.

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

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.

Becker, M., Cabeza, R., & Kizilirmak, J. M. (2022). *A cognitive neuroscience perspective on insight as a memory process: Searching for the solution.* [Preprint]. PsyArXiv. https://doi.org/10.31234/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. (2020a). Inferior frontal gyrus involvement during search and solution in verbal creative problem solving: A parametric fMRI study. *Neuroimage*, *206*, 116294.

Becker, M., Sommer, T., & Kühn, S. (2020b). 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

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*, 201–210.

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

Burgess, N., Maguire, E. A., & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron*, *35*(4), 625–641.

Burns, D. J. (1992). The consequences of generation. *Journal of Memory and Language*, *31*(5), 615–633.

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.

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.

Cador, M., Robbins, T. W., & Everitt, B. J. (1989). Involvement of the amygdala in stimulus-reward associations: Interaction with the ventral striatum. *Neuroscience*, *30*(1), 77–86.

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.

Chavez, C. M., McGaugh, J. L., & Weinberger, N. M. (2009). The basolateral amygdala modulates specific sensory memory representations in the cerebral cortex. *Neurobiology of Learning and Memory*, *91*(4), 382–392. https://doi.org/10.1016/j.nlm.2008.10.010

Clarke, A., Pell, P. J., Ranganath, C., & Tyler, L. K. (2016). Learning warps object representations in the ventral temporal cortex. *Journal of Cognitive Neuroscience*, *28*(7), 1010–1023.

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.

Davey, J., Cornelissen, P. L., Thompson, H. E., Sonkusare, S., Hallam, G., Smallwood, J., & Jefferies, E. (2015). Automatic and controlled semantic retrieval: TMS reveals distinct contributions of posterior middle temporal gyrus and angular gyrus. *Journal of Neuroscience*, *35*(46), 15230–15239.

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.

Dolcos, F., Denkova, E., & Dolcos, S. (2012). Neural correlates of emotional memories: A review of evidence from brain imaging studies. *Psychologia*, *55*(2), 80–111.

Dolcos, F., LaBar, K. S., & Cabeza, R. (2004). Interaction between the amygdala and the medial temporal lobe memory system predicts better memory for emotional events. *Neuron*, *42*(5), 855–863.

Dominowski, R. L., & Buyer, L. S. (2015). Retention of problem solutions: The re-solution effect. *Higher Mental Processes*, *113*(2), 152.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113*(28), 7900–7905.

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.

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.

Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., & Peters, T. M. (1993). 3D statistical neuroanatomical models from 305 MRI volumes. *1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference*, 1813–1817.

Everitt, B. J., Cador, M., & Robbins, T. W. (1989). Interactions between the amygdala and ventral striatum in stimulus-reward associations: Studies using a second-order schedule of sexual reinforcement. *Neuroscience*, *30*(1), 63–75.

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

Fujise, Y., Kubota, M., Suzuki, M., Tamano, H., & Takeda, A. (2017). Blockade of intracellular Zn2+ signaling in the basolateral amygdala affects object recognition memory via attenuation of dentate gyrus LTP. *Neurochemistry International*, *108*, 1–6.

Gieske, A., & Sommer, T. (2023). Independent effects of emotional arousal and reward anticipation on episodic memory formation. *Cerebral Cortex*, *33*(8), 4527–4541.

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

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

Hamann, S. (2001). Cognitive and neural mechanisms of emotional memory. *Trends in Cognitive Sciences*, *5*(9), 394–400.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

Huff, M. L., Miller, R. L., Deisseroth, K., Moorman, D. E., & LaLumiere, R. T. (2013).

Posttraining optogenetic manipulations of basolateral amygdala activity modulate consolidation of inhibitory avoidance memory in rats. *Proceedings of the National Academy of Sciences*, *110*(9), 3597–3602.

Ikemoto, S. (2007). Dopamine reward circuitry: Two projection systems from the ventral midbrain to the nucleus accumbens–olfactory tubercle complex. *Brain Research Reviews*, *56*(1), 27–78.

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.

Kaefer, K., Stella, F., McNaughton, B. L., & Battaglia, F. P. (2022). Replay, the default mode network and the cascaded memory systems model. *Nature Reviews Neuroscience*, *23*(10), 628–640.

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., 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.

Larkum, M. (2013). A cellular mechanism for cortical associations: An organizing principle for the cerebral cortex. *Trends in Neurosciences*, *36*(3), 141–151.

Lim, S.-L., Padmala, S., & Pessoa, L. (2009). Segregating the significant from the mundane on a moment-to-moment basis via direct and indirect amygdala contributions. *Proceedings of the National Academy of Sciences*, *106*(39), 16841–16846.

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

Luo, J., & Niki, K. (2003). Function of hippocampus in “insight” of problem solving. *Hippocampus*, *13*(3), 316–323.

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review*, *27*, 1139–1165.

McGaugh, J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annu. Rev. Neurosci.*, *27*, 1–28.

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

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.

Oh, Y., Chesebrough, C., Erickson, B., Zhang, F., & Kounios, J. (2020). An insight-related neural reward signal. *NeuroImage*, *214*, 116757.

Peelen, M. V., & Caramazza, A. (2012). Conceptual object representations in human anterior temporal cortex. *Journal of Neuroscience*, *32*(45), 15728–15736.

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.

Popescu, A. T., Saghyan, A. A., & Paré, D. (2007). NMDA-dependent facilitation of corticostriatal plasticity by the amygdala. *Proceedings of the National Academy of Sciences*, *104*(1), 341–346.

Pourtois, G., Schettino, A., & Vuilleumier, P. (2013). Brain mechanisms for emotional influences on perception and attention: What is magic and what is not. *Biological Psychology*, *92*(3), 492–512. https://doi.org/10.1016/j.biopsycho.2012.02.007

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

Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J., & Joliot, M. (2020). Automated anatomical labelling atlas 3. *Neuroimage*, *206*, 116189.

Roozendaal, B., Dominique, J.-F., Ferry, B., Setlow, B., & McGaugh, J. L. (2001). Basolateral amygdala–nucleus accumbens interactions in mediating glucocorticoid enhancement of memory consolidation. *Journal of Neuroscience*, *21*(7), 2518–2525.

Roozendaal, B., McEwen, B. S., & Chattarji, S. (2009). Stress, memory and the amygdala. *Nature Reviews Neuroscience*, *10*(6), 423–433.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432.

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities.

Current Biology, *22*(17), 1622–1627.

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications*, *6*(1), 8151.

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.

Schümann, D., & Sommer, T. (2018). Dissociable contributions of the amygdala to the immediate and delayed effects of emotional arousal on memory. *Learning & Memory*, *25*(6), 283–293.

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.

Shofty, B., Gonen, T., Bergmann, E., Mayseless, N., Korn, A., Shamay-Tsoory, S., Grossman, R., Jalon, I., Kahn, I., & Ram, Z. (2022). The default network is causally linked to creative thinking. *Molecular Psychiatry*, *27*(3), 1848–1854.

Simmons, W. K., & Barsalou, L. W. (2003). The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cognitive Neuropsychology*, *20*(3–6), 451–486.

Sommer, T. (2017). The emergence of knowledge and how it supports the memory for novel related information. *Cerebral Cortex*, *27*(3), 1906–1921.

Sommer, T., Hennies, N., Lewis, P. A., & Alink, A. (2022). The assimilation of novel information into schemata and its efficient consolidation. *Journal of Neuroscience*, *42*(30), 5916–5929.

Squire, L. R., Stark, C. E., & Clark, R. E. (2004). The medial temporal lobe. *Annu. Rev. Neurosci.*, *27*, 279–306.

Tang, X., Pang, J., Nie, Q.-Y., Conci, M., Luo, J., & Luo, J. (2015). Probing the cognitive mechanism of mental representational change during chunk decomposition: A parametric fMRI study. *Cerebral Cortex*, *26*(7), 2991–2999.

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.

Tse, D., Takeuchi, T., Kakeyama, M., Kajii, Y., Okuno, H., Tohyama, C., Bito, H., & Morris, R. G. (2011). Schema-dependent gene activation and memory encoding in neocortex. *Science*, *333*(6044), 891–895.

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

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, *15*(1), 273–289.

Van Kesteren, M. T., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211–219.

Vuilleumier, P., Richardson, M. P., Armony, J. L., Driver, J., & Dolan, R. J. (2004). Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nature Neuroscience*, *7*(11), 1271–1278.

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

Wei, T., Liang, X., He, Y., Zang, Y., Han, Z., Caramazza, A., & Bi, Y. (2012). Predicting conceptual processing capacity from spontaneous neuronal activity of the left middle temporal gyrus. *Journal of Neuroscience*, *32*(2), 481–489.

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.

Wills, T. W., Soraci, S. A., Chechile, R. A., & Taylor, A. H. A. (2000). “Aha” effects in the generation of pictures. *Memory & Cognition*, *28*(6), 939–948.

Wu, L., Knoblich, G., & Luo, J. (2013). The role of chunk tightness and chunk familiarity in problem solving: Evidence from ERPs and fMRI. *Human Brain Mapping*, *34*(5), 1173–1186.

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 of the study. We excluded 7 subjects from data analysis due to technical issues at the scanner (N=1), too excessive head movement in the scanner (N=3), due to pathological findings in brain anatomy (N=2) and due to too 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 its 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 comprised 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 measured in a binary way 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 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 two different approaches. In the simple "*insight median split"* 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 conservative insight measure, which can be assumed to be highest in trials that are rated high on all three scales. Thus, in the "*insight cut-off"* approach, we defined as high-insight only those trials that were rated as 3 or 4 in all three dimensions (emotion, suddenness, certainty). 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 one sec fixation cross 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 ordered factor: Trials that were solved in the scanner with high insight and subsequently correctly remembered (*Rem_HI-I*), solved trials with low insight and then subsequently correctly remembered (*Rem_LO-I*), and solved trials that were subsequently not or not correctly remembered (*Forgotten*). 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 (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 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 – Region of interest (ROI) based analysis

The amygdala has been previously suggested to be involved in the metacognitive component of insight (Kizilirmak, Thuerich, et al., 2016; Tik et al., 2018). Therefore, we used it as ROI for the univariate analysis.

First-level analysis. To estimate the effect of insight and insight-related memory on amygdala activity, we conducted two first-level analyses: The goal of the first was to contrast brain activity during *HI-I* and *LO-I* trials. The goal of the second analysis was to relate brain activity to the insight memory effect, i.e. to contrast activity during *Rem_HI-I* with *Rem_LO-I* and *subsequently forgotten*). Both first-level analyses were 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 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].

For the second first-level analysis, a total of 26 separate regressors was created: We modelled the solution button presses for the Mooney images in three different conditions: *Forgotten* [1:3], *Rem_LO-I* [4:6], *Rem_HI-I* [7:9]. To reduce variance from

the implicit baseline, we additionally included a nuisance regressor for all solution button presses of the anagram trials [10:12] and a separate nuisance regressor including all remaining button presses [13:15]. All those events were modelled for one second starting from one second before until the button press. Additionally, we separately modelled six motion parameters [16:22] and separately modelled the mean for each of the four runs [23:26].

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 ROI mask for amygdala was extracted from the FSL Harvard-Oxford (HO) Atlas (Desikan et al., 2006).

Second Level analysis. To statistically compare insight-related mean activity differences in amygdala 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. Because we did not have a specific hypothesis for the VOTC areas, resulting p-values were FDR-corrected for multiple comparison.

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

To estimate the effect of insight-related memory on amygdala brain activity, we repeated the same analyses as described above with the only difference that instead of the insight factor, we included an ordered IME factor representing insight-related better memory (*Rem_HI-I>Rem_LO-I>Forgotten*).

Univariate fMRI data analysis – exploratory whole brain-based analysis

For exploratory purposes, we additionally conducted a *whole brain analysis* to investigate whether there are other relevant brain areas predicting the IME not directly hypothesised in this context.

First Level Analysis. We used the beta values that had already been estimated for the second first level analysis (see further above in section *Univariate fMRI data analysis – Region of interest (ROI) based analysis).* For each individual, simple averaged contrast t-images were calculated from the beta weights of the onset regressors over all four runs.

Second Level Analysis. We included the contrast images corresponding to the three event types (*Forgotten, Rem_LO-I, Rem_HI-I*) during the solution in a random-effects analysis using SPM's one-way ANOVA function. The specific contrasts are reported in the Results section. To correct for multiple comparisons, a specifically conservative cluster threshold of $p<.0001$ ($t>3.96$, uncorrected; see (Eklund et al., 2016) was applied at the voxel level. The resulting clusters were whole-brain family-wise error (FWE) corrected for multiple comparisons at $p<.05$ (extent threshold $k = 78$ voxels). The reported coordinates correspond to the Montreal Neurological Institute (MNI) coordinate system (Evans et al., 1993).

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 does 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 (128sec) 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, response time and the run order (1-4) was added as covariate of no interest. Item and subject variables were estimated as random intercepts. To avoid collinearity artefacts due to the sluggisch hemodynamic response, only solutions that were solved after two seconds were considered (Mumford et al., 2015). 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 ~ RT+ run + ROI + (1| ID) + (1|item) + Ɛ

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

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

Note. PPSS = Pre-Post Solution Similarity. 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 and the IME.

1) Creating brain activity pattern matrices

For each subject, ROI (iLOC, aFusG, pFusG, aITG, mITG, pITG) and time point (pre- and post solution), a brain activity pattern matrix (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 stimuli models

Because we assumed that *conceptual* similarities between the Mooney objects increases after problem solution (=object identification), we created a *stimuli model* 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 *stimuli models*.

Conceptual similarity via the DNN AlexNet *stimuli model* 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 (=*stimuli model*). 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 *stimuli model* 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 brain activation pattern matrix for each subject, each time point and each ROI were partially correlated (Pearson's r) with the same item-wise rows of each *stimuli model* (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 but it 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 and only Mooney images that were solved after two seconds were considered (Mumford et al., 2015). 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 ~ RT+ run + time + ROI + (1|ID) + (1|item) + Ɛ
II. Rep. Strength ~ RT+ run + time + ROI + insight + (1|ID) + (1|item) + Ɛ
III. Rep. Strength ~ RT+ run + time + ROI * insight + (1|ID) + (1|item) + Ɛ
IV. Rep. Strength ~ RT+ run + time * ROI * insight + (1|ID) + (1|item) + Ɛ

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

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 (IME). To test this, we estimated another set of two nested linear mixed models with a) pre to post solution similarity (PPSS) and b) the representational strength (AlexNet or Word2Vec) at the post-solution timepoint (solution button press) as dependent variables and ROI as well as an ordered IME factor describing insight-related better memory (*Rem_HI-I >Rem_LO-I> Forgotten)* as independent factor. 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 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*IME interaction term was additionally modelled (III., see equations below).

I. PPSS/Rep.Strength ~ RT+ run + ROI + (1| ID) + (1|item) + Ɛ
II. PPSS/Rep.Strength ~ RT+ run + ROI + IME + (1| ID) + (1|item) + Ɛ
III. PPSS/Rep.Strength ~ RT+ run + ROI * IME + (1| ID) + (1|item) + Ɛ

Note. PPSS = Pre-Post Solution Similarity; Rep. Strength = Representational strength of solution object at post solution (button press).

Multivariate fMRI data analysis – functional connectivity analysis

Given the results from the univariate and multivoxel pattern analyses, we expected amygdala to be more functionally connected to those visual brain areas that displayed RC (pFusG, iLOC). 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. Two first-level analyses were conducted: 1) to estimate FC during high versus low insight conditions (*HI-I>LO-I*) and 2) to estimate FC as a function of insight-related better memory (*Rem_HI-I>Rem_LO-I>Forgotten*). Due to our specific hypotheses regarding the functional relationship between amygdala and VOTC regions eliciting RC, we only estimated FC between bilateral amygdala, pFusG and iLOC. The 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 (6x6) were computed for each participant (31), for each condition (1) *HI-I vs. LO-I*; 2) *Rem_HI-I, Rem_LO-I, Forgotten*), 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 in CONN. 1) To estimate FC for the contrast high versus low insight trials, the within-subjects contrast in the GLM was specified as follows y~ 1*HI-I – 1*LO-I + ε). For statistical inference, the 6x6 ROI connectivity matrix was thresholded at the connection level at p<.05 (p-uncorrected and one-sided, only considering positive correlations) and at the cluster-level at p<.05 (ROI-p FDR

corrected) because we did not know which ROI in this amygdala and VOTC-RC network would be most relevant, i.e. elicit the highest degree (Benjamini & Hochberg, 1995). 2)_Due to similar FC results in the contrast *HI-I>LO-I*, both amygdala hemispheres were averaged and the within-subjects contrast y~ 1**Rem_HI-I* + 0**Rem_LO-I* - 1**Forgotten* + ε for statistical inference was applied to four individual links (amygdala-left or right iLOC; amygdala - left or right pFusG).

Data and code availability

The behavioural data including the aggregated univariate and multivoxel pattern fMRI data, and FC values 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.

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.

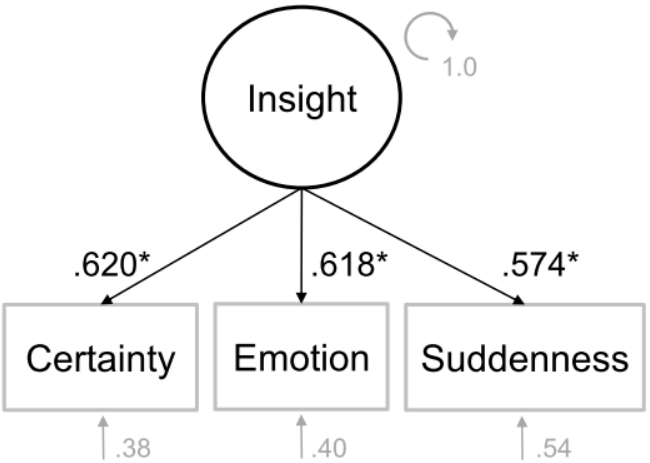

Figure S2. Measurement model for latent insight variable

Note. Asterisk = significance at p<.001;

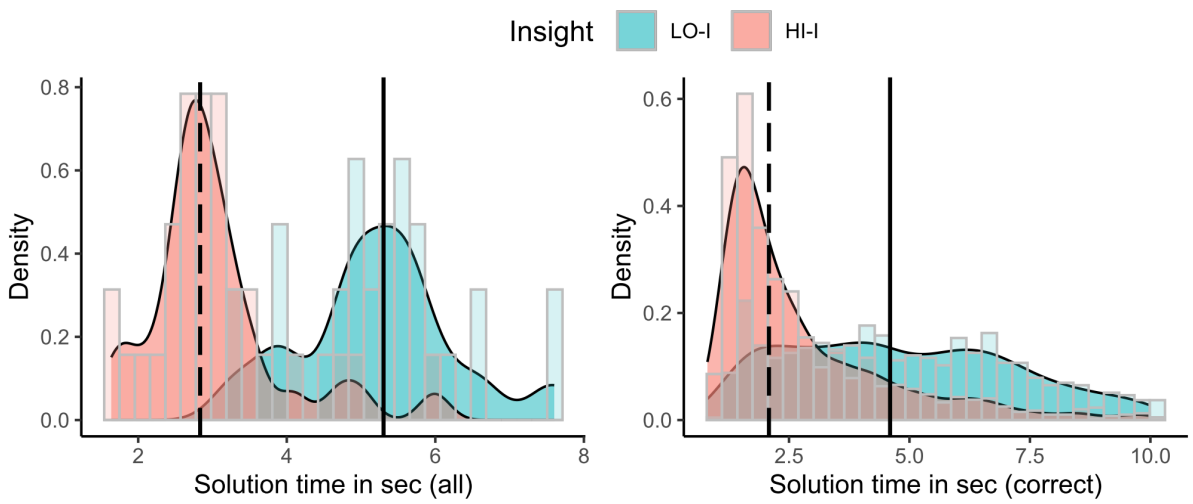

Figure S3. Response time distribution divided by high and low insight solutions (within subject).

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

Regarding univariate effects of VOTC areas, there was a significant condition ($Chi^2(1) = 14.90$, $p<.005$) as well as condition*ROI interaction ($Chi^2(5) = 38.89$, $p<.001$) effect. However, posthoc analyses revealed that iLOC ($t(1492)=-5.45$, *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 see Fig.S4-A).

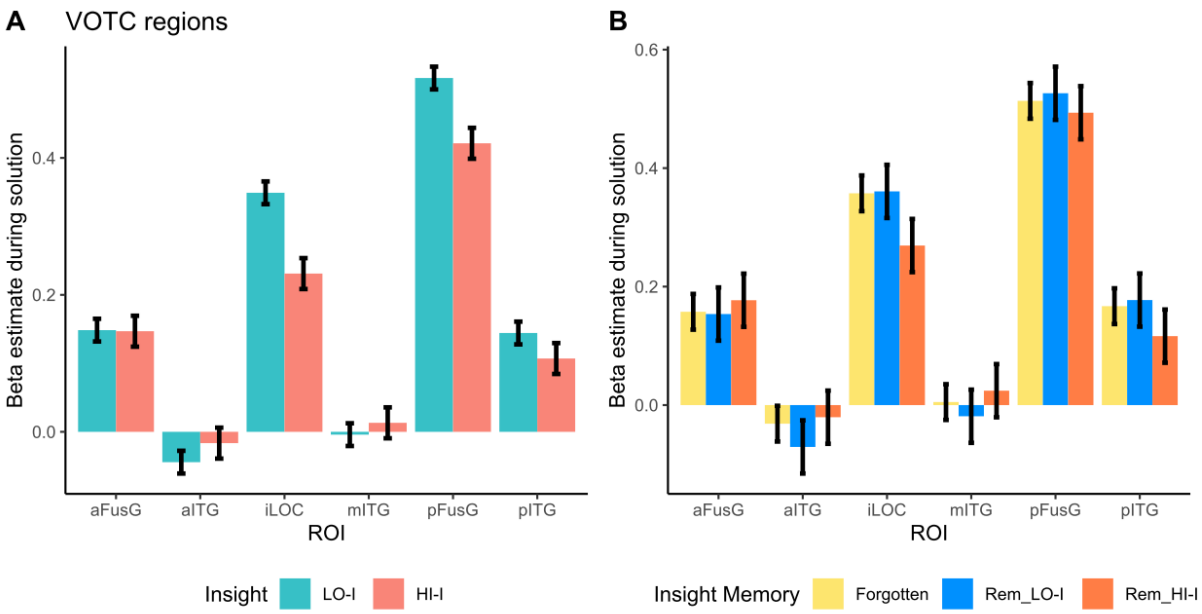

Figure S4. Univariate BOLD activity in VOTC areas (not) predicting insight and the insight memory effect.

Regarding insight-related memory, no significant condition overall effect supporting the IME was found (*p*>.174) but there was a significant condition*ROI interaction effect (Chi²(10)=33.63, *p*<.001). Posthoc analyses revealed no condition effects for any of those ROIs, except for iLOC showing a reverse pattern in the sense that *Rem-HI-I* showed significantly less activity than *Rem-LO-I* (t(2183)= -3.81; *p-FDR*<.01) and Forgotten (t(2183)= -3.67; *p-FDR*<.01) trials (see Fig.S4-B).

Univariate fMRI analyses and results including response time as covariate

To exclude the fact that the observed amygdala effects are just due to differences in response time between the high and low insight condition, we conducted a control analysis including response time as covariate 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 during solution (time point of solution button press), we estimated two nested mixed effect

models as depicted below including the averaged amygdala 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) in addition to response time (RT) was added as covariate of no interest. To exclude pop-up solutions not related to insight (Becker et al., 2021), only solved trials that were solved after two seconds were considered. A factor that differentiated between two conditions –solved trials with high (*HI-I*) and low insight (*LO-I*) served as independent measure. To calculate the insight-related better memory effect (IME) on amygdala activity during solution, we replicated the same modelling procedure as described above with the only difference that the insight factor (*HI-I>LO-I)* was interchanged with the ordered IME factor (*Rem_HI-I>Rem_LO-I>Forgotten*) representing insight-related better memory.

I. Beta values ~ run + RT + (1| ID) + ε
II. Beta values ~ run + RT + insight(HI-I, LO-I) + (1 | ID) + ε

I. Beta values ~ run + RT + (1| ID) + ε
II. Beta values ~ run + RT + IME + (1 | ID) + ε

P-values for both nested mixed effects models were calculated via log-likelihood tests. Posthoc marginal mean differences between conditions were estimated using the emmeans package (v.1.7.5). Post Hoc results regarding the IME were additionally adjusted for multiple comparison via the Tukey HSD method (Tukey, 1953).

Activity in Amygdala during solution predicts Insight and Insight Memory Effect when controlling for response time

Amygdala activity was significantly higher for HI-I (M=1.54, 95%CI[1.28 – 1.8]) than for *LO-I* (M=1.04, 95%CI[0.79 – 1.3]) insight trials during solution ($Chi^2(1)$= 25.32, p<.001, ß=.16) when additionally controlling for response time. Furthermore, the model including the ordered IME factor was significant ($Chi^2(2)$=13.38; p<.01, ß=.10). Posthoc analyses revealed that amygdala activity during solution was significantly higher for *Rem_HI-I* compared to *Forgotten* (z=3.30, p<.01) and *Rem-LO-I* (z=2.86,

p<.05) trials while there was no difference between Forgotten and *Rem-LO-I* trials (p>.72).

Univariate fMRI analyses and results in subject space

To exclude potential smoothing and normalisation artefacts due to the small volume of the amygdala ROI, we additionally conducted another ROI analysis in subject space for the contrast *HI-I>LO-I* and *Forgotten>Rem_LO-I > Rem_HI-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 VOTC 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 during solution predicts Insight and Insight Memory Effect

Amygdala activity significantly distinguishes between high and low insight trials during solution ($Chi^2(1)=28.55$, $p<.001$). Posthoc analyses revealed that Amygdala activity was significantly higher for *HI-I* compared to *LO-I* trials ($t(215)=5.50$, $p<.001$, ß=.45) (see Fig.S5, left). Regarding the insight memory effect, there was also a significant condition effect for amygdala ($Chi^2(1)=33.72$, $p<.001$). Posthoc analyses revealed that Amygdala activity was significantly higher for *Rem_HI-I* compared to *Rem_LO-I* ($t(338)=5.38$, $p<.001$) and *Forgotten* ($t(338)=4.85$, $p<.001$) (see Fig. S5, right).

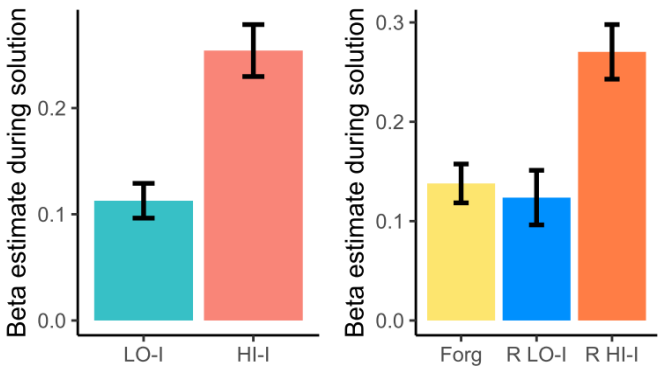

Figure S5. Univariate BOLD activity of amygdala during solution in native space predicting insight and the insight memory effect

Note. Values represent estimated marginal means. Error bars represent standard error. The left panel = amygdala activity for high (*HI-I*) and low (*LO-I*) insight solutions and the right panel = amygdala activity for solved trials that were later either forgotten (*Forgotten*) or remembered and originally solved with low (*Rem_LO-I*) or high (*Rem_HI-I)* insight.

Table S1

Brain areas (label) identified by the results of the whole-brain analysis

xyz MNI	Label	L/R	% Cluster	# Vx Cluster	% Label	# Vx Label
***Rem-HI-I* > *Rem-LO-I* > *Forgotten* (1 0 -1)**						
18 14 -15	Hippocampus	L	13.85	1343	19.96	932
	Hippocampus	R	7.74	1343	10.99	946
	Insula	L	6.63	1343	4.79	1858
	NAcc	R	5.51	1343	51.75	143
	NAcc	L	5.29	1343	43.56	163
	Olfactory	L	5.14	1343	23.71	291
	Parahippocampus	R	4.32	1343	5.12	1132
	Amygdala	L	4.02	1343	24.55	220
-19 44 48	Frontal Sup Med	L	31.60	4339	45.82	2992
	Frontal Sup Med	R	16.50	4339	33.55	2134

	Frontal Sup 2	L	10.42	4339	9.28	4870
	ACC pre	L	9.24	4339	63.96	627
	Frontal Med Orb	R	4.86	4339	24.65	856
	Frontal Med Orb	L	4.10	4339	24.76	719
	ACC pre	R	3.83	4339	25.62	648
-47 -71 46	Angular	L	39.74	843	28.56	1173
	Temporal Mid	L	3.44	843	0.59	4942
52 -73 42	Angular	R	55.61	579	18.38	1752
	Parietal Inf	R	2.42	579	1.04	1345
-7 -37 38	Cingulate Mid	L	25.08	1519	19.63	1941
	Precuneus	L	20.54	1519	8.84	3528
	Cingulate Mid	R	19.49	1519	13.44	2203
	Precuneus	R	13.03	1519	6.06	3265
	Cingulate Post	L	11.26	1519	36.93	463
	Cingulate Post	R	6.12	1519	27.76	335
56 -9 -15	Temporal Mid	R	75.00	292	4.97	4409
	Temporal Sup	R	22.60	292	2.10	3141
-55 -3 -21	Temporal Mid	L	93.19	514	9.69	4942
	Temporal Sup	L	3.89	514	0.87	2296
	Temporal Pole Mid	L	0.58	514	0.40	755
88 -43 -7	Temporal Mid	R	82.12	302	5.62	4409
	Temporal Inf	R	6.95	302	0.59	3557
40 50 -13	Frontal Mid	R	71.79	78	1.15	4860
	Frontal Inf Orb 2	R	20.51	78	1.83	874
	OFC anterior	R	5.13	78	0.62	648
24 -39 2	Precuneus	R	30.23	86	0.80	3265
	Calcarine	R	27.91	86	1.29	1861

	Hippocampus	R	18.60	86	1.69	946

***Forgotten > Rem-LO-I > Rem-HI-I* (1 0 -1)**

32 -97 18	Occipital Mid	R	37.18	3462	61.34	2098
	Occipital Sup	R	10.37	3462	25.41	1413
	Occipital Inf	R	10.23	3462	35.79	989
	Temporal Inf	R	7.42	3462	7.23	3557
	Temporal Mid	R	6.90	3462	5.42	4409
2 -75 -25	Vermis 7		43.70	135	30.41	194
	Cerebellum 6	R	20.74	135	1.56	1795
	Vermis 6		15.56	135	5.66	371
	Cerebellum 6	L	13.33	135	1.06	1694
	Cerebellum Crus1	L	5.93	135	0.31	2603
-39 -75 6	Occipital Mid	L	57.07	1938	33.87	3265
	Occipital Inf	L	8.31	1938	17.11	941
	Temporal Mid	L	4.95	1938	1.94	4942
22 -65 54	Parietal Sup	R	67.45	685	20.79	2222
	Parietal Inf	R	7.01	685	3.57	1345
	Occipital Sup	R	7.01	685	3.40	1413
	Angular	R	5.84	685	2.28	1752
28 -5 46	Precentral	R	54.05	222	3.55	3381
	Frontal Mid 2	R	10.36	222	0.47	4860
	Frontal Sup 2	R	8.11	222	0.35	5126
-9 18 46	Supp Motor Area	L	88.66	97	4.01	2147
	Frontal Sup Med	L	11.34	97	0.37	2992

Note. Voxelwise threshold p<.0001; clusterwise threshold p-FWE <.05; L/R = left or right hemisphere; Vx = voxel; # = Amount of; ACC pre = Anterior cingulate cortex pregenual; OFC = Orbitofrontal Cortex; Mid = Middle; Sup = Superior; Inf = Inferior; Post = Posterior; Ant = Anterior; Orb = Orbital; Med = Medial; NAcc = Nucleus Accumbens; *Rem_HI-I* > *Rem_LO-I* > *Forgotten*: high insight trials subsequently remembered versus low insight trials subsequently remembered versus subsequently Forgotten trials.

Table S2

Functional Network Connectivity Results predicting Insight *HI-I* > *LO-I*

ROI	**t**	**p-FDR**		**t**	**p**
Amy r	2.91	0.010			
			iLOC l	2.45	0.010
			iLOC r	1.95	0.030
			pFusG r	1.84	0.038
Amy l	2.18	0.040			
			iLOC l	2.16	0.019
			iLOC r	2.23	0.017
			pFusG l	1.80	0.041
			pFusG r	1.93	0.032
pFusG r	3.10	0.010			
			iLOC r	2.11	0.022
			Amy l	1.84	0.038
			Amy r	1.93	0.032
pFusG l	2.92	0.010			
			iLOC l	2.87	0.004
			iLOC r	1.75	0.045
			pFusG r	1.84	0.038
			Amy r	1.80	0.041
iLOC r	2.15	0.040			

Table S2

			pFusG l	1.75	0.045
			pFusG r	2.11	0.022
			Amy l	1.95	0.030
			Amy r	2.23	0.017
iLOC l	4.01	0.002			
			pFusG l	2.87	0.004
			Amy l	2.45	0.010
			Amy r	2.16	0.019

Note. HI-I > LO-I: Values stem from contrast high insight trials > low insight trials. All anatomical areas were determined using the AAL3 atlas (Rolls et al., 2020; Tzourio-Mazoyer et al., 2002) based on the percentage of voxels belonging to a respective anatomical area intersected by an activated cluster from the ANOVA analysis and visually tested for plausibility.

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.