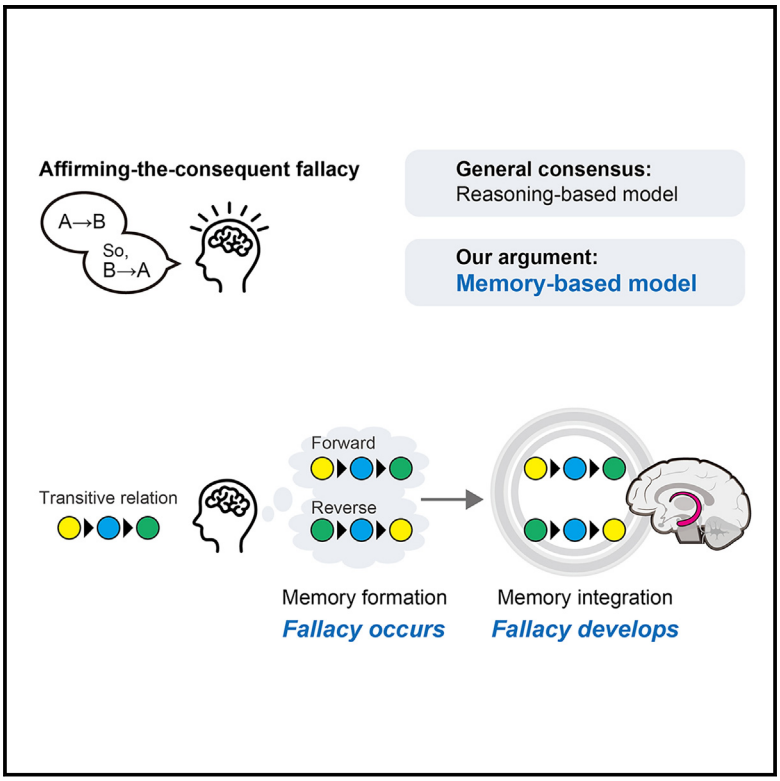
iScience

Article

The role of memory in affirming-the-consequent fallacy

Graphical abstract



Authors

Yoko Higuchi, Ethan Oblak, Hiroko Nakamura, Makiko Yamada, Kazuhisa Shibata

Correspondence

yokohiguchi0114@gmail.com (Y.H.), kazuhisa.shibata@riken.jp (K.S.)

In brief

Medical imaging; Behavioral neuroscience; Cognitive neuroscience

Highlights

- The affirming-the-consequent fallacy can arise through memory mechanisms
- When a sequence is encoded, the memory of the reversed sequence is formed
- The memories of the forward and reversed sequences are integrated over time
- The hippocampus is associated with the integration of the memories

Higuchi et al., 2025, iScience *28*, 111889
February 21, 2025
https://doi.org/10.1016/j.isci.2025.111889

CellPress

iScience

CellPress
OPEN ACCESS

Article

The role of memory in affirming-the-consequent fallacy

Yoko Higuchi,[1,2,*] Ethan Oblak,[1] Hiroko Nakamura,[3,4] Makiko Yamada,[5] and Kazuhisa Shibata[1,6,*]
[1]RIKEN Center for Brain Science, RIKEN, Wako, Saitama, Japan
[2]Department of Cognitive and Information Sciences, Chiba Institute of Technology, Narashino, Chiba, Japan
[3]Japan Society for the Promotion of Science, Chiyoda, Tokyo, Japan
[4]School of Science and Engineering, Tokyo Denki University, Adachi, Tokyo, Japan
[5]Institute for Quantum Life Science, National Institute for Quantum Science and Technology, Inage, Chiba, Japan
[6]Lead contact
*Correspondence: yokohiguchi0114@gmail.com (Y.H.), kazuhisa.shibata@riken.jp (K.S.)
https://doi.org/10.1016/j.isci.2025.111889

SUMMARY

People tend to recognize that a transitive relation remains true even when its order is reversed. This affirming-the-consequent fallacy is thought to be uniquely related to human intelligence. It is generally thought that this fallacy is a byproduct of explicit reasoning at the moment of recognition of the reversed order. Here, we provide evidence suggesting a reconsideration of this account using an implicit memory paradigm, which minimizes the involvement of explicit reasoning. Specifically, we tested a two-stage memory model: (1) when a sequence of events is encoded, the memory of the reversed sequence is formed, resulting in the affirming-the-consequent fallacy, and (2) the memories of the forward and reversed sequences are integrated over time, reinforcing the fallacy. Results of behavioral and functional magnetic resonance imaging experiments were consistent with this memory-based model. Our findings suggest that the affirming-the-consequent fallacy may begin unwittingly when individuals memorize a transitive relation.

INTRODUCTION

Humans learn transitive relations from experiences.[1] In the inference of transitive relations, a fallacy known as "affirming the consequent" is often observed.[2–10] For example, while rain can make the ground wet, the ground being wet does not necessarily mean it rained. Nevertheless, we tend to infer that it must have rained if the ground is wet. Importantly, affirmation of the consequent is not commonly observed in non-human animals.[11] In the case of humans, even 8-month infants show the affirmation of the consequent.[4] Thus, this fallacy reflects the characteristics of human inference and intelligence.[12]

Studies of logical reasoning suggest that the affirming-the-consequent fallacy generally occurs when people reason about transitive relations that are often expressed as conditional propositions.[13] According to this reasoning-based model, the fallacy occurs at the moment that individuals erroneously interpret that the reverse of the given transitive relation holds true.[1,3,5,6,13–18] This model suggests that the fallacy is a byproduct of humans' reasoning system that is biased toward generalizing a learned transitive relation in the reverse direction.[3,4]

In contrast to this reasoning-based model, the present study introduces a novel memory-based model and provides evidence suggesting that the affirming-the-consequent fallacy can also occur through mechanisms of memory. In this memory-based model, the fallacy arises from the memory of the reversed relation that is unwittingly made soon after a transitive relation is encoded in the human memory system. Then, the fallacy is ultimately realized when individuals are asked about the reversed relation. Although not previously associated with the fallacy, similar phenomena have been observed in research on implicit memory, which minimizes involvement of explicit reasoning during recognition of a transitive relation.[19,20] Turk-Browne and Scholl (2009) demonstrated that, using an implicit memory paradigm, learning of a transitive relation involves the formation of memories for the reversed relation. This finding is consistent with the memory-based model.

Suppose this memory-based model we introduce serves as a complementary model to the reasoning-based model. In that case, the memory-based model should explain a variant of the affirming-the-consequent fallacy that cannot be accounted for by the reasoning-based model. Previous studies have reported at least two variants of the fallacy. In the first variant, individuals can distinguish the encoded transitive relation A→B from its reverse B→A.[17,18,21–23] This variant can be explained by the reasoning-based model in which the fallacy occurs when individuals evaluate the reversed relation.[1,13,15,16,24,25] In the second variant, individuals are no longer able to distinguish the encoded transitive relation from its reverse.[4,12,26] This variant cannot be accounted for by the reasoning-based model, which assumes the explicit evaluation of the reversed relation. To comprehensively explain the two variants, the present study postulates that the affirming-the-consequent fallacy can emerge as a result of memories that change over time.

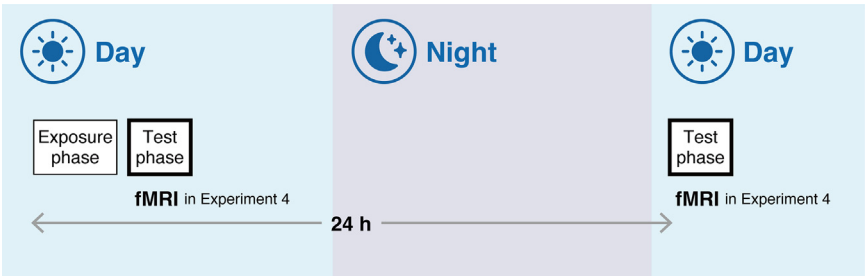


Figure 1. Experimental design to investigate temporal changes in memory

In all experiments, there were exposure phases followed by two test phases. To investigate temporal changes in memory, we conducted the test phases immediately after the exposure phase and 24 h later. In Experiment 4, to explore the neural basis of memory changes, fMRI scans were performed during the two test phases.

Specifically, we propose a two-stage memory model in which the characteristics of the memory system account for the affirming-the-consequent fallacy. First, when a transitive relation is encoded, a memory of the reversed relation is also formed. At this stage, the memory of the transitive relation is distinguished from that of the reversed relation, resulting in the first variant of the affirming-the-consequent fallacy. Second, memories of the encoded transitive relation and its reverse are integrated over time. As a result, the first variant, where the encoded transitive relation and reversed relation were distinguishable right after encoding, develops into the second variant, where these relations are no longer distinguishable.

We conducted a series of experiments to test the two-stage memory model of the affirming-the-consequent fallacy. We combined an established paradigm used for the investigation of the affirming-the-consequent fallacy[4,9,10,12,27–35] with a protocol used in implicit memory research.[36–44] In our experiments, specific transitive relations consisting of a series of objects are repeatedly presented unbeknownst to participants, inducing unintentional learning of the relations. As participants are not instructed about the transitive relations, the possibility of participants engaging in explicit reasoning about the relations is minimal. Indeed, participants remain unaware of the presentations of transitive relations in this paradigm.[42,45–52] Thus, this paradigm minimizes the involvement of explicit reasoning in the affirming-the-consequent fallacy in our study. To further confirm the involvement of the human memory system of the brain in the affirming-the-consequent fallacy, we conducted a functional magnetic resonance imaging (fMRI) experiment with particular emphasis on regions implicated in memory processing in the brain.[53–62]

The results of the behavioral experiments are consistent with the two-stage memory model. Participants were able to distinguish forward sequences (encoded transitive relations) and backward sequences (reversed relations) from novel sequences immediately after exposure to sequences and even 24 h later, indicating the formation and retention of memories for both forward and backward sequences. However, while participants could distinguish forward and backward sequences immediately after exposure, they were unable to do so 24 h later.

The fMRI experiment aimed at identifying brain regions involved in the time-dependent changes in memory found in the behavioral experiments. The activation in the right anterior hippocampus was found to be associated with differentiation between forward and backward sequences immediately after exposure. Consistent with the behavioral findings, the activation differences in the right anterior hippocampus between the forward and backward sequences diminished 24 h later.

These results align with the two-stage memory model, proposing that in the process of learning transitive relations, the brain initially forms memories of both the transitive relations and their reverse. Over time, these memories are integrated. The right anterior hippocampus plays a crucial role in the changes of such memory relationships. Collectively, our findings provide evidence that memory may underlie the affirming-the-consequent fallacy.

RESULTS

Four experiments were conducted to test the two-stage memory model of the affirming-the-consequent fallacy. We combined an established paradigm in research on the fallacy[4,27–30,35] with a protocol for implicit memory.[36–44] Each experiment consisted of an exposure phase for memory formation and two test phases for examining memory retention and integration over time (Figure 1). In the exposure phase (Figure 2A), participants observed a series of objects and performed a category judgment task on each object (manmade vs. natural) as a cover task. The objects were assigned to fixed sequences that reflect specific transitive relations repeatedly presented to the participants. Each fixed sequence consisted of three objects (e.g., ABC, DEF, and GHI). The fixed sequences were repeatedly presented unbeknownst to the participants. During each trial in the test phase (Figure 2B), participants were presented with two different sequences. They were asked to report which of the two sequences they had already seen during the exposure phase. The conditions for the two sequences varied for each experiment (Figure 2C). To examine temporal changes in memory, test phases were conducted immediately after the exposure phase and 24 h later.

Experiment 1a: Memory of forward sequences is formed and retained even 24 h after exposure

In the first experiment, we tested whether the memory of forward sequences (e.g., ABC) is formed using the implicit memory paradigm and retained over time. Specifically, in each trial of the test phases, participants (N = 30) were presented with a forward sequence and a randomly generated new sequence (e.g., ABC vs. AEI; see STAR Methods for details). They were then asked to indicate which of the two sequences they had seen during the exposure phase. The test phases were conducted immediately after the exposure phase (day 1) and 24 h later (day 2).

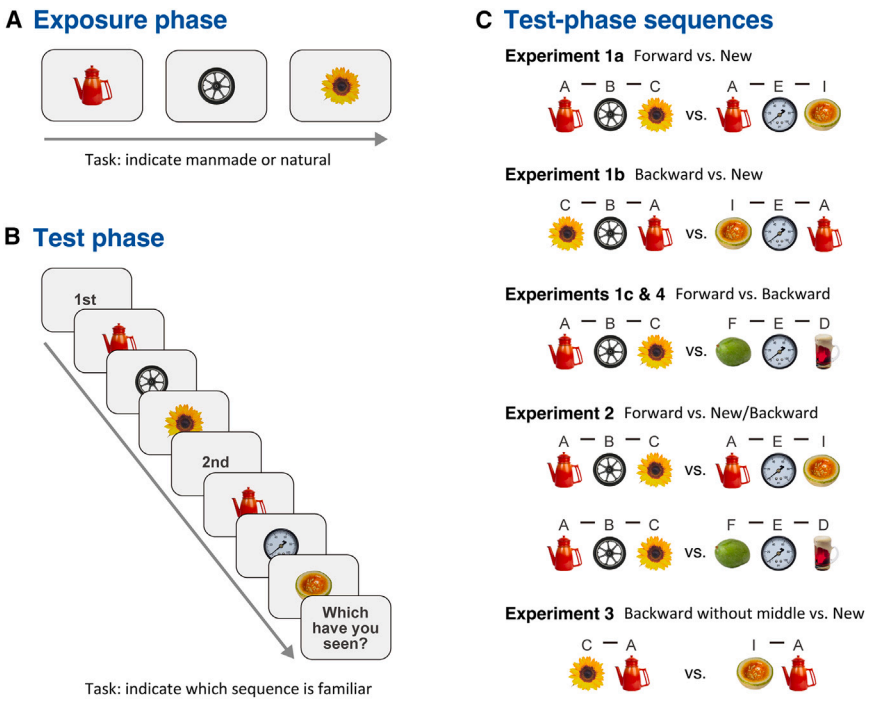


Figure 2. Stimuli and tasks used in exposure and test phases

(A) Exposure phase. Participants were presented with a series of objects and performed a category judgment task (referred to as the "cover task"), indicating whether the current object was manmade or natural by pressing a key on the keyboard.

(B) Test phase. On each trial in the test phase, participants were presented with two sequences and were asked to report which of the two sequences they had previously seen during the exposure phase.

(C) Example sequences/pairs used in the test phase of each experiment. Sequences/pairs were individually generated for each participant by combining manmade and natural objects. For illustrative purposes, alphabets (e.g., ABC) are labeled in the figure; however, these letters were not displayed during the actual experiments. To prevent any potential learning between the two test phases, we used different sets of stimuli from the exposure phase for the first and second test phases.

Figure 3 (left panel) shows the proportion of the choice for the forward sequences. Participants chose the forward sequences significantly more often than the chance level on both day 1 (one-sample t test with Bonferroni corrected threshold, $\alpha/2$; $t_{29} = 5.407$, $p < 10^{-4}$, $d = 0.987$) and day 2 ($t_{29} = 6.590$, $p < 10^{-4}$, $d = 1.203$). No significant difference in the proportion of the choice for the forward sequences was found between day 1 and day 2 (paired t test; $t_{29} = 1.444$, $p = 0.160$, $d = 0.240$). These results suggest that the memory of the forward sequences was formed during the exposure phase and retained stable over time.

Experiment 1b: Memory of backward sequences is formed and retained even 24 h after exposure

We tested whether the memory of backward sequences (e.g., CBA), which were not actually presented during the exposure phase, is formed as a result of exposure to the forward sequences and retained over time. In other words, we investigated whether exposure to the forward sequences results in the affirming-the-consequent fallacy. In the test phases, a new set of participants ($N = 30$) were presented with a backward sequence and a new sequence (e.g., CBA vs. AEI) and were then asked to indicate which of the two sequences they had seen during the exposure phase. Participants chose the backward sequences significantly more often than the chance level on both day 1 (one-sample t test with Bonferroni corrected threshold, $\alpha/2$; $t_{29} = 3.518$, $p = 0.002$, $d = 0.642$) and day 2 ($t_{29} = 4.393$, $p = 10^{-4}$, $d = 0.802$) (Figure 3, middle panel). No significant difference in the proportion of the choice for the backward sequences was found between day 1 and day 2 (paired t test; $t_{29} = 0.124$, $p = 0.902$, $d = 0.027$). These results suggest that the memory of the backward sequences, which were not actually observed during the exposure phase, was formed and retained stable over time.

Experiment 1c: Memories of forward and backward sequences are integrated over time

We tested whether the forward and backward sequences became less distinguishable as time passed. In the test phases, a forward sequence was compared against a backward sequence ($N = 30$; e.g., ABC vs. FED). Figure 3 (right panel) shows the proportion of choice for the forward sequences. On day 1, the proportion of choice for the forward sequences was significantly higher than the chance level (one-sample t test with Bonferroni corrected threshold, $\alpha/2$; $t_{29} = 3.568$, $p = 0.001$, $d = 0.651$). However, on day 2, the proportion of choice for the forward sequences was not statistically different from chance ($t_{29} = 1.305$, $p = 0.202$, $d = 0.238$). Moreover, the proportion of choice for the forward sequences on day 2 was significantly lower than that on day 1 (paired t test; $t_{29} = 2.321$, $p = 0.028$, $d = 0.554$). These results suggest that the memories of the forward and backward sequences are integrated over time.

The results of Experiments 1a, b, and c provide evidence consistent with the two-stage memory model of the affirming-the-consequent fallacy. The memories of the forward (Experiment 1a) and backward sequences (Experiment 1b) are formed and retained 24 h after encoding. However, memories of the forward and backward sequences became indistinguishable over time (Experiment 1c). These findings align with the idea that the affirming-the-consequent fallacy can occur and develop through two stages of memory processing in the brain.

Note that forgetting cannot explain the inability to distinguish the forward from backward sequences on day 2 compared to day 1 in Experiment 1c. In Experiments 1a and b, participants

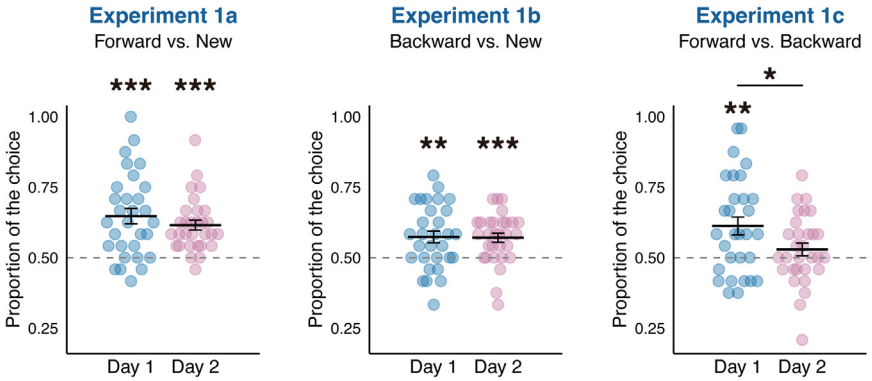


Figure 3. Results of Experiments 1a, 1b, and 1c, conducted as between-participant experiments examining temporal changes in memory

In Experiment 1a, participants were able to indicate forward sequences against new sequences significantly above the chance level on both day 1 and day 2 (one-sample t tests, Bonferroni corrected threshold, $^{***}p < 0.001$). In Experiment 1b, participants were able to indicate backward sequences against new sequences significantly above the chance level on both day 1 and day 2 (one-sample t tests, Bonferroni corrected threshold, $^{**}p < 0.01$, $^{***}p < 0.001$). In Experiment 1c, on day 1, participants were able to identify forward sequences against backward sequences significantly above the chance level (one-sample t test, Bonferroni corrected threshold, $^{**}p < 0.01$). However, on day 2, the proportion of choice for forward sequences did not statistically differ from the chance level. Furthermore, on day 2, the proportion of choice for forward sequences was significantly lower than that on day 1 (paired t test, $^{*}p < 0.05$). Horizontal lines represent group means, and error bars represent standard errors.

were still able to identify the forward and backward sequences from new sequences 24 h after the exposure phase, indicating the memories of the forward and backward sequences were retained. These results support that it is the relationship between the memories of the forward and backward sequences that changed over time.

One might think that the differential changes in the proportion of the choices for forward sequences observed in Experiments 1a, b, and c could be explained by differences in reaction times (RTs) to the tasks. We argue that this is unlikely for the following two reasons. First, there were no time limits imposed on the tasks. Second, the instructions given to participants were identical across these experiments. To verify our claim, we analyzed the RTs. A two-way ANOVA on RTs for experiment (Experiments 1a, 1b, and 1c) and day (day 1, day 2) revealed no significant main effects of experiment ($F_{2,\ 27} = 0.302$, $p = 0.742$, $\eta_p^2 = 0.022$) or day ($F_{1,\ 27} = 0.361$, $p = 0.553$, $\eta_p^2 = 0.013$), nor a significant interaction between experiment and day ($F_{2,\ 27} = 0.414$, $p = 0.665$, $\eta_p^2 = 0.030$). These results rule out the possibility that changes in the proportion of the choices observed in Experiments 1a, b, and c can be explained by differences in RTs. Furthermore, the non-significant main effect of experiment suggests that participants performed the tasks during the test phases of these experiments in the same or similar manner as instructed.

Experiment 2: The results are replicated in a within-participant design

Experiments 1a, b, and c were conducted with different sets of participants, respectively. It is possible that participants only in Experiment 1c might forget the forward and backward sequences over time, and therefore, were unable to distinguish these sequences on day 2. Thus, Experiment 2 ($N = 30$) employed a within-participant design to test if the relationship between memories of the forward and backward sequences selectively changes over time while the memories of the forward and backward sequences are retained. In the test phase, there were two types of trials: comparing a forward sequence against a new sequence (e.g., ABC vs. AEI) and comparing a forward sequence against a backward sequence (e.g., ABC vs. FED).

As shown in Figure 4, the results replicated the findings in Experiments 1a and c. A two-way ANOVA for trial type (forward vs. new, forward vs. backward) and day (day 1, day 2) revealed a significant interaction ($F_{1,29} = 5.197$, $p = 0.030$, $\eta_p^2 = 0.152$; see Table S1 for all the results of the ANOVA). The simple main effect of day was not significant for the trial in which the forward and new sequences were compared ($F_{1,29} = 2.058$, $p = 0.162$, $\eta_p^2 = 0.066$). As in Experiment 1a, participants chose the forward sequences significantly more often than the chance level on both day 1 (one-sample t test with Bonferroni corrected threshold, $\alpha/4$; $t_{29} = 7.719$, $p < 10^{-4}$, $d = 1.409$) and day 2 ($t_{29} = 7.099$, $p < 10^{-4}$, $d = 1.296$). On the other hand, when the forward and backward sequences were compared, the simple main effect of day was significant ($F_{1,29} = 21.890$, $p = 10^{-4}$, $\eta_p^2 = 0.430$). As in Experiment 1c, the proportion of choice for the forward sequences was significantly higher than the chance level on day 1 (one-sample t test with Bonferroni corrected threshold, $\alpha/4$; $t_{29} = 7.448$, $p < 10^{-4}$, $d = 1.360$), but not on day 2 ($t_{29} = 1.668$, $p = 0.106$, $d = 0.304$). These results confirmed that, while the memory of the forward sequences was maintained on both day 1 and day 2, the relationship between the memories of the forward and backward sequences selectively changed over time.

One may argue that Experiment 2 did not test the memory of the backward sequences, leaving the possibility that participants in Experiment 2 had forgotten the backward sequences on day 2. However, this is unlikely. If the memory of the backward sequences was not retained on day 2, it is expected that the proportion of choice for the forward sequences over the backward sequences would be higher than the chance level on day 2. This expectation is inconsistent with the actual results (Figure 4).

As in Experiments 1a, b, and c, we analyzed the RTs. A two-way ANOVA on RTs for trial type (forward vs. new, forward vs. backward) and day (day 1, day 2) revealed no significant main effects of trial type ($F_{1,\ 29} = 0.065$, $p = 0.901$, $\eta_p^2 = 0.002$) or day ($F_{1,\ 29} = 1.459$, $p = 0.237$, $\eta_p^2 = 0.048$), nor a significant interaction between trial type and day ($F_{1,\ 29} = 0.642$, $p = 0.430$, $\eta_p^2 = 0.022$). These results support the following two claims. First, it is unlikely that the changes in the proportion of the choices are driven by

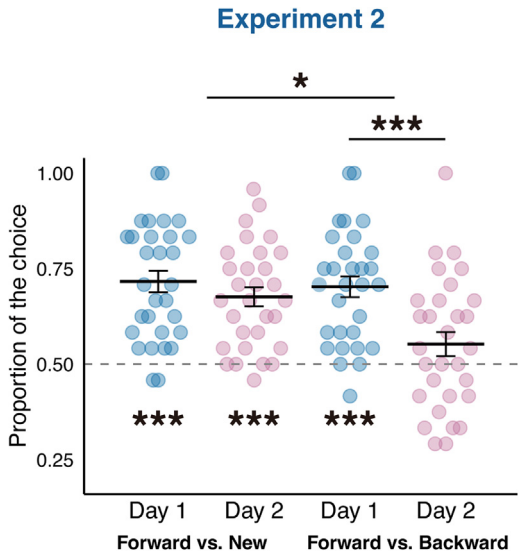


Figure 4. Results of Experiment 2, conducted as a within-participant experiment examining temporal changes in memory

A two-way ANOVA for trial type (forward vs. new, forward vs. backward) and day (day 1, day 2) revealed a significant interaction (*$p < 0.05$). In the trials distinguishing forward sequences from new sequences, the proportion of choice for forward sequences did not decrease on day 2 compared to day 1. In contrast, in the trials distinguishing forward sequences from backward sequences, the proportion of choice for forward sequences decreased on day 2 compared to day 1 (***$p < 0.001$). For the forward vs. new trials, participants were able to indicate forward sequences significantly above the chance level on both day 1 and day 2 (one-sample t tests, Bonferroni corrected threshold, ***$p < 0.001$). However, for the forward vs. backward trials, participants were able to indicate forward sequences significantly above the chance level on day 1 (one-sample t test, Bonferroni corrected threshold, ***$p < 0.001$), but on day 2, the proportion of choice for forward sequences was not statistically different from the chance level. Horizontal lines represent group means, and error bars represent standard errors.

changes in RTs. Second, participants performed the task in a consistent manner across the different types of trials.

Experiment 3: Forgetting the sequential order of objects does not account for the results

The results of Experiments 1 and 2 suggest that the memories of the forward and backward sequences become integrated over time. However, these findings could also be attributed to forgetting of the sequential order of the objects and the effects of temporal proximity among the objects. In other words, due to the forgetting of the sequential order (e.g., ABC), the objects in a sequence (e.g., A, B, and C) might have been stored as a single unit, that is, "as a whole," regardless of their order.

Assume that the sequential order (e.g., ABC) is forgotten on day 2 while the memory of the temporal proximity (A, B, and C appear close in time) is retained. In terms of the temporal proximity among A, B, and C, the memories of the forward and backward sequences are indistinguishable. Thus, if such forgetting were indeed occurring, any pair of sequences consisting of objects A, B, and C should be indistinguishable from each other. This prediction also applies to a skipped and backward sequence such as CA (see STAR Methods for details). If the

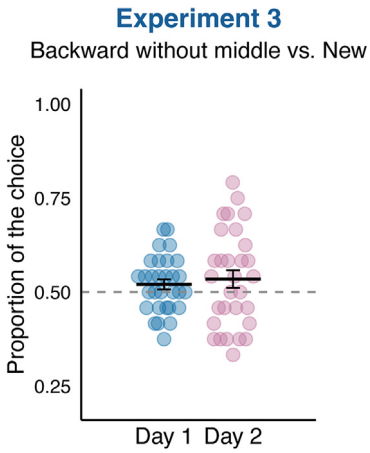


Figure 5. Results of Experiment 3 examining whether forgetting of the sequential orders occurred

The proportion of choice for CA pairs was not statistically different from the chance level. Additionally, no significant difference was observed in the proportion of choice for CA pairs over new pairs between day 1 and day 2. Horizontal lines represent group means, and error bars represent standard errors.

sequential order (e.g., ABC) is selectively forgotten on day 2, it is expected that the skipped and backward sequence (e.g., CA) becomes familiar to participants compared to a random and new pair (e.g., IA) due to the temporal proximity between the objects A and C during the exposure phase. On the other hand, if the order of the forward and backward sequences were retained, the skipped and backward sequences should not be recognized as familiar, even after the passage of time.

To test which is the case, we conducted Experiment 3 ($N = 30$). The exposure phase was identical to Experiments 1a, 1b, 1c, and 2. During the test phases, participants were presented with a CA pair and a new combination and were asked to report which of the two pairs they had previously seen during the exposure phase. Figure 5 shows the proportion of the choice for the CA pairs. The proportion of choice for CA pairs was not statistically different from the chance level on day 1 (one-sample t test with Bonferroni corrected threshold, $\alpha/2$; $t_{29} = 1.542$, $p = 0.134$, $d = 0.281$) or day 2 ($t_{29} = 1.479$, $p = 0.150$, $d = 0.270$). We found no significant difference in the proportion of choice for CA pairs between day 1 and day 2 (paired t test; $t_{29} = 0.494$, $p = 0.625$, $d = 0.133$). These results are inconsistent with the prediction based on the forgetting of the sequential orders. Thus, the findings from Experiments 1c and 2 cannot be attributed to a simple explanation involving the forgetting and effects of the temporal proximity that result in learning the objects in a sequence "as a whole."

The results of the behavioral experiments revealed the following. First, the forward and backward sequences are distinguishable from new sequences immediately after exposure and 24 h later. Second, while the forward and backward sequences are distinguishable immediately after encoding, they become indistinguishable 24 h later. Lastly, these results cannot be explained by forgetting the sequential orders. These findings align with the two-stage memory model in which the two variants of the affirming-the-consequent fallacy emerge through a

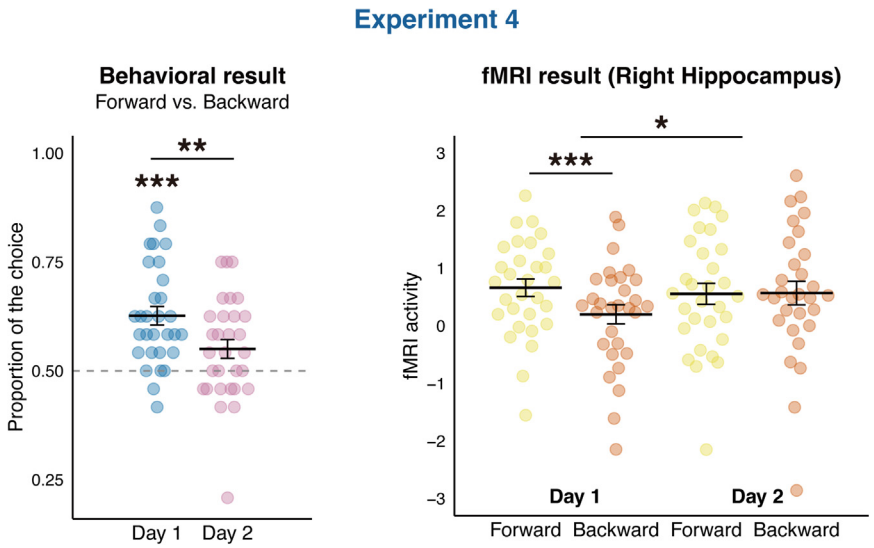


Figure 6. Results of Experiment 4 examining memory system involvement in affirming-the-consequent fallacy

The left figure shows the proportion of choice for the forward sequences. The proportion of choice for forward sequences was significantly lower on day 2 than on day 1 (paired t test, $^{**}p < 0.01$). Participants indicated forward sequences significantly above chance on day 1 (one-sample t test, Bonferroni corrected threshold, $^{***}p < 0.001$), whereas on day 2, this difference was not statistically significant. The right figure shows the fMRI activity (defined as the contrasts between the forward and backward sequences and the inter-trial interval as a baseline) in the right hippocampus. A two-way ANOVA for sequence type (forward, backward) and day (day 1, day 2) revealed a significant interaction only in the region contained in the right hippocampus (Bonferroni corrected threshold, $^{*}p < 0.05$). On day 1, the right hippocampus exhibited a significantly stronger activity for forward sequences compared to backward sequences ($^{***}p < 0.001$), whereas on day 2, there was no significant difference in activity between forward and backward sequences. Horizontal lines represent group means, and error bars represent standard errors.

two-stage mechanism originating from the characteristics of the memory system. Following exposure to the forward sequences, the memory of the unobserved backward sequences is formed, resulting in the first variant of the fallacy. Over time, the memories of the forward and backward sequences are integrated, leading to the development of the first variant into the second variant of the fallacy. All of these results were obtained using the implicit memory paradigm, which is designed to minimize the involvement of explicit reasoning during the test phases. These results are consistent with the memory-based model of the affirming-the-consequent fallacy.

Experiment 4: The right hippocampus is involved in the time-dependent memory integration

To test the memory-based model in which the memory system is involved in the affirming-the-consequent fallacy at the neural level, we measured fMRI in Experiment 4. If the memory-based model is correct, memory-related brain regions should be associated with time-dependent integration of the memories found in the behavioral experiments. The growing body of evidence suggests that the integration of memories is primarily associated with the hippocampus.[63–67] In the hippocampus, neural activity related to past experiences is replayed in both forward and backward sequences.[68–70] This neuronal replay provides a potential mechanism for memory integration and the learning of knowledge structures.[66] In humans, it has also been shown that such hippocampal replay occurs and contributes to memory integration.[55,71] Thus, we employed 14 memory-related brain regions (Figure S1), including the bilateral hippocampus and caudate, as the regions of interest (ROIs) by using NeuroSynth[72] (see STAR Methods for details).

The procedure of Experiment 4 was identical to that of Experiment 1c, except that the test phases were conducted in an MRI scanner. As in Experiment 1c, a forward sequence was compared against a backward sequence ($N = 30$; e.g., ABC vs. FED) in the test phases. The behavioral results (Figure 6, left panel) replicated those of Experiment 1c. The proportion of choice for the forward sequences on day 2 was significantly lower than on day 1 (paired t test; $t_{29} = 2.949$, $p = 0.006$, $d = 0.655$). Participants selected the forward sequences significantly more often than the chance level on day 1 (one-sample t test with Bonferroni corrected threshold, $\alpha/2$; $t_{29} = 5.992$, $p < 10^{-4}$, $d = 1.094$), but not on day 2 ($t_{29} = 2.330$, $p = 0.027$, d = 0.425).

In the fMRI analyses, we examined which ROI(s) exhibit changes in activation patterns comparable to the memory integration found in the behavioral experiments. For this aim, a general linear model (GLM) analysis was performed for each ROI (see STAR Methods for details). We found activation changes in the right hippocampus corresponded to the results of the behavioral experiments (Figure 6, right panel; see Figure S2 for other regions). A two-way ANOVA with repeated measures for sequence type (forward, backward) and day (day 1, day 2) was performed for each ROI. A significant interaction was found only in the right hippocampus (Bonferroni corrected threshold, $\alpha/14$; $F_{1,29} = 10.99$, $p = 0.003$, $\eta_p^2 = 0.275$). The simple main effect of sequence type was significant on day 1 ($F_{1,29} = 21.19$, $p = 10^{-4}$, $\eta_p^2 = 0.422$), indicating that the right hippocampus exhibited stronger fMRI activity for the forward sequences compared to the backward sequences. However, on day 2, the simple main effect of sequence type was not significant ($F_{1,29} = 0.013$, $p = 0.912$, $\eta_p^2 < 10^{-3}$). None of the other regions than the right hippocampus showed this pattern of activation changes (see Table S2 for full ANOVA results). These results suggest that the right hippocampus is particularly involved in the development of the first variant of the affirming-the-consequent fallacy into the second variant.

The right anterior hippocampus is associated with the integration of memories

Theoretical studies have suggested that the anterior hippocampus is primarily involved in implicit memory.[73,74] The voxels obtained from NeuroSynth[72] are located in the anterior part of the

right hippocampus in a standard space (Figure S1), which is consistent with these theoretical studies. However, these voxels were provided based on previous functional mapping studies. Thus, the voxels are not necessarily aligned with the anterior hippocampus defined by the anatomical structures of individual participants. Given the significant variability in the anatomy of hippocampus among individuals,[75–82] it is important to test whether anatomically defined anterior hippocampus is also associated with the memory integration.

To test this, we used ITK-SNAP[83] and Automatic Segmentation of Hippocampal Subfields (ASHS) toolbox[84] to define the anterior and posterior parts of the right hippocampus. A two-way ANOVA with repeated measures for sequence type (forward, backward) and day (day 1, day 2) was performed for each of the anterior and posterior parts in the right hippocampus. We found a significant interaction only in the anterior part (Bonferroni corrected threshold, $\alpha/2$; $F_{1,29} = 6.096$, $p = 0.020$, $\eta_p^2 = 0.174$), but not in the posterior part ($F_{1,29} = 1.089$, $p = 0.305$, $\eta_p^2 = 0.037$) (Figure S3; see Table S3 for full ANOVA results). The simple main effect for the interaction in the right anterior hippocampus of sequence type was significant on day 1 ($F_{1,29} = 15.96$, $p < 10^{-3}$, $\eta_p^2 = 0.355$), indicating that the right anterior hippocampus exhibited stronger activity for the forward sequences compared to the backward sequences on day 1. However, no significant simple main effect of sequence type was found on day 2 ($F_{1,29} = 0.089$, $p = 0.767$, $\eta_p^2 < 10^{-3}$). Taken together, these results suggest that the development of the first variant of the affirming-the-consequent fallacy into the second variant is specifically associated with the anterior part of the right hippocampus.

DISCUSSION

In this study, we experimentally tested the memory-based model in which the two variants of affirming-the-consequent fallacy emerge through a two-stage memory mechanism. The key findings in the current study are as follows. Immediately after exposure and even 24 h later, memories of the forward and backward sequences are distinguishable from new sequences (Experiments 1a and b). While the forward sequences can be distinguished from the backward sequences immediately after exposure, they became indistinguishable 24 h later (Experiments 1c, 2, and 4). The memory changes cannot be solely explained by the forgetting of the sequential orders or temporal proximity (Experiment 3). The right anterior hippocampus is primarily associated with the memory changes (Experiment 4). These results were obtained using an implicit memory paradigm, which minimizes the involvement of explicit reasoning during the recognition of the backward sequences. Collectively, the findings of the current study suggest that the affirming-the-consequent fallacy can originate from memories that change over time in the brain, though these findings do not exclude the role of reasoning processes in the fallacy.

The results of this study indicate that the affirming-the-consequent fallacy arose immediately after learning and develops within 24 h. First, when a transitive relation was encoded into memory, its reverse was also formed, giving rise to the affirming-the-consequent fallacy.[85] At this stage, participants were still able to distinguish the original memory of the transitive relation from its reverse, if asked. However, these two memories became integrated in 24 h after encoding. As a result, the separability of these memories was weakened, making the fallacy stronger.

The results of the fMRI experiment showed the right hippocampus showed activation changes corresponding to the development of the first variant of the affirming-the-consequent fallacy into the second variant. Previous research on episodic memory has implicated the role of the right hippocampus in memory integration. The activity of the right hippocampus is associated with combining multiple concepts.[86] It is also reported that the right hippocampus is involved in recalling integrated and abstracted memories, while the left hippocampus is implicated in recalling specific episodic memories.[87] The development of the second variant of the fallacy could be attributed to the integration of memories of forward and backward sequences into more generalized memories, which could be why the involvement of the right hippocampus was observed in the current study. The fMRI results emphasized the importance of the anterior part of the right hippocampus. It is suggested that memory integration is primarily associated with the anterior hippocampus.[88–91] The hippocampus has anatomical subregions, with the cornu ammonis (CA1–3) mainly located in the anterior hippocampus, and the dentate gyrus (DG) typically corresponding to the posterior hippocampus.[92] Thus, it is possible that CA1–3 is mainly involved in memory integration, as suggested by previous research findings.[93] In the posterior hippocampus, we did not observe activation changes corresponding to the results of the behavioral experiments. This might be because the DG is less involved in memory integration and more associated with pattern separation.[73,94]

It is worth noting that the results of fMRI experiment do not suggest that the affirming-the-consequent fallacy is solely due to memory in general. While the hippocampus indeed plays a central role in memory processes, it is also involved in logical reasoning.[95] In our fMRI experiment, we tested the hypothesis that if the development of the fallacy is primarily driven by memory mechanisms, memory-related brain regions should be involved in the time-dependent integration of memories observed in the behavioral experiments. However, our current results and the findings in previous studies are not mutually exclusive. Specifically, our results do not rule out the possibility that the hippocampus contributes to logical reasoning in general cases of the affirming-the-consequent fallacy. It is plausible that the hippocampus plays a significant role in both memory and logical reasoning in the context of this fallacy. By employing an implicit memory paradigm, we aimed to minimize the influence of explicit reasoning in this study. Nevertheless, further research is necessary to disentangle the specific contributions of memory and reasoning mechanisms to the fallacy, respectively.

In conclusion, the findings of the current study provide evidence supporting the idea that the affirming-the-consequent fallacy may emerge through memory processing. When learning transitive relations, the memories of forward relations and their reverses are first formed and subsequently integrated over time. Such changes in memory relationships are associated with the hub of the memory system in the brain: the hippocampus.

Limitations of the study

The first limitation of this study is that none of our experiments directly tested logical reasoning. Our experiments used temporal sequences, but not logical relations. Furthermore, we did not include specific logic detection tasks that could help dissociate memory processes from logical reasoning processes. This makes it difficult to rule out the influence of logical reasoning or to conclude that the observed effects are purely memory-based. Participants were exposed to the temporal sequences of objects and were later tested on whether they had incidentally stored these sequences. In this test, participants were not required to engage in explicit logical reasoning. Therefore, it remains unclear what kind of representation participants actually formed through exposure to the temporal sequences. It could be that participants formed a conditional proposition like, "If the teapot appears, then the flower will appear next." Alternatively, they may have formed a non-logical proposition about the sequence, such as, "The teapot is followed by the flower." It could also be that what was encoded was not a linguistic or logical proposition at all, but rather a simple unidirectional association.

Despite this ambiguity, there are the following two reasons to argue that this study is indeed investigating the affirming-the-consequent fallacy. First, prior research provides evidence supporting the possibility that temporal sequences could be encoded as conditionals. Such temporal sequences are known to be frequently learned as causal relationships,[96] and conditional propositions are often used to represent such relationships.[97] Given these findings, it is possible that participants in this study associated the learned sequences with causal relationships, which were subsequently represented as conditional propositions.

Second, and more importantly, even if it is unclear whether the temporal sequences were encoded as conditional propositions, there is considerable research demonstrating that inferences like the affirming-the-consequent fallacy can still occur. Numerous studies have demonstrated that from a sequential order like "A is followed by B," participants can infer both "If A, then B" (modus ponens) and "If B, then A" (affirming the consequent).[2,4–10,12,33,35,98–106] These studies support the notion that inferences like the affirming-the-consequent fallacy can occur even when temporal sequences are not explicitly represented as conditional propositions. Thus, we argue that the results of this study are also applicable to the affirming-the-consequent fallacy.

Another limitation of this study is that it does not provide direct evidence as to why the affirming-the-consequent fallacy is unique to humans. This fallacy is thought to be specific to humans and mostly absent in non-human animals.[11] Both humans and non-humans rely on memory for their thinking processes, and memory representations serve as the basis of inference and intelligence.[107] Differences in inferential processing between humans and animals may arise from variations in memory systems. Human memory appears to differ qualitatively from animal memory.[108] While humans can form a memory of the reversed relation after learning a transitive relation,[109–111] chimpanzees, for example, do not exhibit this ability.[4] In humans, the relationships between multiple memories change over time.[112–115] Indeed, memories are integrated with the passage of time.[53] The ability to reconstruct memory scenarios by integrating memory and semantic information may be unique to humans.[116] Non-human animals are unable to embed memories into future contexts or link present events to future events, resulting in context-independent memory representations.[108,117] Moreover, while humans have concept cells that respond to concepts at the single-cell level,[118] such neurons have not been found in other species.[107] These studies argue that there is a potential to explain uniquely developed cognitive abilities in humans, such as knowledge generalization and creative thinking.[107] Human-specific memory systems might contribute to inference, such as the affirming-the-consequent fallacy. However, further research is needed to conclude whether the memory integration observed in this study is unique to humans.

RESOURCE AVAILABILITY

Lead contact

Further information for resources should be directed and will be fulfilled by the lead contact, Kazuhisa Shibata (kazuhisa.shibata@riken.jp).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Anonymized fMRI data and summary behavioral data have been deposited at CBS Data Sharing Platform and are publicly available as of the date of publication at https://doi.org/10.60178/cbs.20241223-001.
- All original code has been deposited at GitHub and is publicly available at https://github.com/yokohiguchi/affirming_the_consequent as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We are thankful to Brynn Sherman for guiding us on the analysis of the hippocampal subfields. Our thanks extend to Kenichi Ueno and Chisato Suzuki from the Support Unit for Functional Magnetic Resonance Imaging at RIKEN Center for Brain Science for their assistance with MRI measurements. Special thanks to Yasunori Kinosada for his assistance with data visualization and statistical analysis, Lana Okuma and Julian Matthews for some English advice, and Akito Tsuneda for the data collection and discussions on the prototype of this study. Further, we are grateful to Takahiro Nishio, Shoko Nagatomo, Aya Kokubu, Kiwako Shimada, and Ryohei Mimura for assisting with data collection. This work was supported by JSPS KAKENHI Grant (JP20J01411, JP20K14272, and JP23KJ2231 to Y.H.; JP22KJ2787 to H.N.; JP22H01108, JP22K18265, and JP23H04833 to M.Y.; and JP19H01041 and JP20H05715 to K.S.), JST CREST Grant (JPMJCR23P4 to M.Y.), MEXT QLEAP Grant (JPMXS0120330644 to M.Y.), and JST Moonshot R&D Grant (JPMJMS2295-01 to M.Y. and JPMJMS2013 to K.S.).

AUTHOR CONTRIBUTIONS

Y.H. and K.S. designed the research; Y.H. collected the data; Y.H., E.O., and K.S. analyzed the data; and Y.H. and K.S. wrote the initial draft of the manuscript. All authors contributed to the editing of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Experimental paradigm
  - Experiment 1a
  - Experiment 1b
  - Experiment 1c
  - Experiment 2
  - Experiment 3
  - Experiment 4
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.111889.

Received: May 19, 2024
Revised: October 9, 2024
Accepted: January 22, 2025
Published: January 25, 2025

REFERENCES

1. Goodwin, G.P., and Johnson-Laird, P.N. (2005). Reasoning about relations. Psychol. Rev. *112*, 468–493. https://doi.org/10.1037/0033-295X.112.2.468.
2. Chartier, T.F., and Fagot, J. (2022). Associative symmetry: a divide between humans and nonhumans? Trends Cognit. Sci. *26*, 286–289. https://doi.org/10.1016/j.tics.2022.01.009.
3. Hastie, R., and Dawes, R.M. (2001). Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making (Sage Publications, Inc).
4. Imai, M., Murai, C., Miyazaki, M., Okada, H., and Tomonaga, M. (2021). The contingency symmetry bias (affirming the consequent fallacy) as a prerequisite for word learning : A comparative study of pre-linguistic human infants and chimpanzees. Cognition *214*, 104755. https://doi.org/10.1016/j.cognition.2021.104755.
5. Hattori, M., and Oaksford, M. (2007). Adaptive Non-Interventional Heuristics for Covariation Detection in Causal Induction: Model Comparison and Rational Analysis. Cognit. Sci. *31*, 765–814. https://doi.org/10.1080/03640210701530755.
6. Takahashi, T., Nakano, M., and Shinohara, S. (2010). Cognitive Symmetry: Illogical but Rational Biases. Symmetry: Culture and Science *21*, 275–294.
7. Horne, P.J., and Lowe, C.F. (1996). On the origins of naming and other symbolic behavior. J. Exp. Anal. Behav. *65*, 185–241. https://doi.org/10.1901/jeab.1996.65-185.
8. Luciano, C., Gómez Becerra, I., and Rodríguez Valverde, M. (2007). The role of multiple-exemplar training and naming in establishing derived equivalence in an infant. J. Exp. Anal. Behav. *87*, 349–365. https://doi.org/10.1901/jeab.2007.08-06.
9. Sidman, M., and Tailby, W. (1982). Conditional discrimination vs. matching to sample: an expansion of the testing paradigm. J. Exp. Anal. Behav. *37*, 5–22. https://doi.org/10.1901/jeab.1982.37-5.
10. Sobel, D.M., and Kirkham, N.Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. Dev. Psychol. *42*, 1103–1115. https://doi.org/10.1037/0012-1649.42.6.1103.
11. Lionello-DeNolf, K.M. (2009). The search for symmetry: 25 years in review. Learn. Behav. *37*, 188–203. https://doi.org/10.3758/LB.37.2.188.
12. Ogawa, A., Yamazaki, Y., Ueno, K., Cheng, K., and Iriki, A. (2010). Neural Correlates of Species-typical Illogical Cognitive Bias in Human Inference. J. Cognit. Neurosci. *22*, 2120–2130. https://doi.org/10.1162/jocn.2009.21330.
13. Barrouillet, P., Gauffroy, C., and Lecas, J.-F. (2008). Mental models and the suppositional account of conditionals. Psychol. Rev. *115*, 760–772. https://doi.org/10.1037/0033-295X.115.3.760.
14. Byrne, R.M.J., and Johnson-Laird, P.N. (2009). 'If' and the problems of conditional reasoning. Trends Cognit. Sci. *13*, 282–287. https://doi.org/10.1016/j.tics.2009.04.003.
15. Johnson-Laird, P.N., Byrne, R.M., and Schaeken, W. (1992). Propositional reasoning by model. Psychol. Rev. *99*, 418–439. https://doi.org/10.1037/0033-295X.99.3.418.
16. Johnson-Laird, P.N., and Byrne, R.M.J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. Psychol. Rev. *109*, 646–678. https://doi.org/10.1037/0033-295X.109.4.646.
17. Evans, J.S.B.T., Handley, S.J., Nelzens, H., and Over, D.E. (2007). Thinking about conditionals: A study of individual differences. Mem. Cognit. *35*, 1772–1784. https://doi.org/10.3758/BF03193509.
18. Evans, J.S.B.T., Handley, S.J., Neilens, H., and Over, D. (2010). The influence of cognitive ability and instructional set on causal conditional inference. Q. J. Exp. Psychol. *63*, 892–909. https://doi.org/10.1080/17470210903111821.
19. Yang, J., Xu, X., Du, X., Shi, C., and Fang, F. (2011). Effects of Unconscious Processing on Implicit Memory for Fearful Faces. PLoS One *6*, e14641. https://doi.org/10.1371/journal.pone.0014641.
20. Fang, F., and He, S. (2005). Cortical responses to invisible objects in the human dorsal and ventral pathways. Nat. Neurosci. *8*, 1380–1385. https://doi.org/10.1038/nn1537.
21. Newman, I.R., Gibb, M., and Thompson, V.A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. J. Exp. Psychol. Learn. Mem. Cogn. *43*, 1154–1170. https://doi.org/10.1037/xlm0000372.
22. Evans, J.S.B.T., Handley, S.J., and Bacon, A.M. (2009). Reasoning under time pressure: a study of causal conditional inference. Exp. Psychol. *56*, 77–83. https://doi.org/10.1027/1618-3169.56.2.77.
23. Schwartz, F., Epinat-Duclos, J., Léone, J., and Prado, J. (2017). The neural development of conditional reasoning in children: Different mechanisms for assessing the logical validity and likelihood of conclusions. Neuroimage *163*, 264–275. https://doi.org/10.1016/j.neuroimage.2017.09.029.
24. Barrouillet, P., and Lecas, J.-F. (1999). Mental Models in Conditional Reasoning and Working Memory. Think. Reas. *5*, 289–302. https://doi.org/10.1080/135467899393940.
25. Meiser, T., Klauer, K.C., and Naumer, B. (2001). Propositional reasoning and working memory: the role of prior training and pragmatic content. Acta Psychol. *106*, 303–327. https://doi.org/10.1016/S0001-6918(00)00055-X.
26. Qiu, J., Li, H., Huang, X., Zhang, F., Chen, A., Luo, Y., Zhang, Q., and Yuan, H. (2007). The neural basis of conditional reasoning: An event-related potential study. Neuropsychologia *45*, 1533–1539. https://doi.org/10.1016/j.neuropsychologia.2006.11.014.
27. Acuna, B.D., Eliassen, J.C., Donoghue, J.P., and Sanes, J.N. (2002). Frontal and parietal lobe activation during transitive inference in humans. Cerebr. Cortex *12*, 1312–1321. https://doi.org/10.1093/cercor/12.12.1312.
28. Baggio, G., Cherubini, P., Pischedda, D., Blumenthal, A., Haynes, J.-D., and Reverberi, C. (2016). Multiple neural representations of elementary logical connectives. Neuroimage *135*, 300–310. https://doi.org/10.1016/j.neuroimage.2016.04.061.
29. Brunamonti, E., Mione, V., Di Bello, F., Pani, P., Genovesio, A., and Ferraina, S. (2016). Neuronal Modulation in the Prefrontal Cortex in a Transitive Inference Task: Evidence of Neuronal Correlates of Mental Schema Management. J. Neurosci. *36*, 1223–1236. https://doi.org/10.1523/JNEUROSCI.1473-15.2016.

30. Fugard, A.J.B., Pfeifer, N., Mayerhofer, B., and Kleiter, G.D. (2011). How people interpret conditionals: Shifts toward the conditional event. J. Exp. Psychol. Learn. Mem. Cogn. *37*, 635–648. https://doi.org/10.1037/a0022329.

31. Zhang, X., Qiu, Y., Li, J., Jia, C., Liao, J., Chen, K., Qiu, L., Yuan, Z., and Huang, R. (2022). Neural correlates of transitive inference: An SDM meta-analysis on 32 fMRI studies. Neuroimage *258*, 119354. https://doi.org/10.1016/j.neuroimage.2022.119354.

32. Valiña, M.D., and Martín, M. (2021). Reasoning with the THOG Problem: A Forty-Year Retrospective. Psychology *12*, 2042–2069. https://doi.org/10.4236/psych.2021.1212124.

33. Tomonaga, M., Matsuzawa, T., Fujita, K., and Yamamoto, J. (1991). Emergence of symmetry in a visual conditional discrimination by chimpanzees (*Pan troglodytes*). Psychol. Rep. *68*, 51–60. https://doi.org/10.2466/PR0.68.1.51-60.

34. Schusterman, R.J., and Kastak, D. (1993). A California Sea Lion (Zalophus Californianus) is Capable of Forming Equivalence Relations. Psychol. Rec. *43*, 823–839. https://doi.org/10.1007/BF03395915.

35. Ogawa, A., Yamazaki, Y., Ueno, K., Cheng, K., and Iriki, A. (2010). Inferential reasoning by exclusion recruits parietal and prefrontal cortices. Neuroimage *52*, 1603–1610. https://doi.org/10.1016/j.neuroimage.2010.05.040.

36. Turk-Browne, N.B. (2019). The hippocampus as a visual area organized by space and time: A spatiotemporal similarity hypothesis. Vis. Res. *165*, 123–130. https://doi.org/10.1016/j.visres.2019.10.007.

37. Sherman, B.E., and Turk-Browne, N.B. (2020). Statistical prediction of the future impairs episodic encoding of the present. Proc. Natl. Acad. Sci. USA *117*, 22760–22770. https://doi.org/10.1073/pnas.2013291117.

38. Chun, M.M., and Turk-Browne, N.B. (2007). Interactions between attention and memory. Curr. Opin. Neurobiol. *17*, 177–184. https://doi.org/10.1016/j.conb.2007.03.005.

39. Turk-Browne, N.B., Jungé, J., and Scholl, B.J. (2005). The automaticity of visual statistical learning. J. Exp. Psychol. Gen. *134*, 552–564. https://doi.org/10.1037/0096-3445.134.4.552.

40. Sherman, B.E., Graves, K.N., and Turk-Browne, N.B. (2020). The Prevalence and Importance of Statistical Learning in Human Cognition and Behavior. Curr Opin Behav Sci *32*, 15–20. https://doi.org/10.1016/j.cobeha.2020.01.015.

41. Schapiro, A.C., Kustner, L.V., and Turk-Browne, N.B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. Curr. Biol. *22*, 1622–1627. https://doi.org/10.1016/j.cub.2012.06.056.

42. Zhao, J., Ngo, N., McKendrick, R., and Turk-Browne, N.B. (2011). Mutual interference between statistical summary perception and statistical learning. Psychol. Sci. *22*, 1212–1219. https://doi.org/10.1177/0956797611419304.

43. Turk-Browne, N.B., Scholl, B.J., Johnson, M.K., and Chun, M.M. (2010). Implicit Perceptual Anticipation Triggered by Statistical Learning. J. Neurosci. *30*, 11177–11187. https://doi.org/10.1523/JNEUROSCI.0858-10.2010.

44. Turk-Browne, N.B., Isola, P.J., Scholl, B.J., and Treat, T.A. (2008). Multidimensional visual statistical learning. J. Exp. Psychol. Learn. Mem. Cogn. *34*, 399–407. https://doi.org/10.1037/0278-7393.34.2.399.

45. Bertels, J., Franco, A., and Destrebecqz, A. (2012). How implicit is visual statistical learning? J. Exp. Psychol. Learn. Mem. Cogn. *38*, 1425–1431. https://doi.org/10.1037/a0027210.

46. Fiser, J., and Lengyel, G. (2022). Statistical Learning in Vision. Annu. Rev. Vis. Sci. *8*, 265–290. https://doi.org/10.1146/annurev-vision-100720-103343.

47. Fiser, J., and Aslin, R.N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. Psychol. Sci. *12*, 499–504. https://doi.org/10.1111/1467-9280.00392.

48. Lengyel, G., Žalalytė, G., Pantelides, A., Ingram, J.N., Fiser, J., Lengyel, M., and Wolpert, D.M. (2019). Unimodal statistical learning produces multimodal object-like representations. Elife *8*, 1–21. https://doi.org/10.7554/elife.43942.

49. Zhao, J., Al-Aidroos, N., and Turk-Browne, N.B. (2013). Attention Is Spontaneously Biased Toward Regularities. Psychol. Sci. *24*, 667–677. https://doi.org/10.1177/0956797612460407.

50. Luo, Y., and Zhao, J. (2018). Statistical Learning Creates Novel Object Associations via Transitive Relations. Psychol. Sci. *29*, 1207–1220. https://doi.org/10.1177/0956797618762400.

51. Kim, R., Seitz, A., Feenstra, H., and Shams, L. (2009). Testing assumptions of statistical learning: Is it long-term and implicit? Neurosci. Lett. *461*, 145–149. https://doi.org/10.1016/j.neulet.2009.06.030.

52. Jun, J., and Chong, S.C. (2018). Visual statistical learning at basic and subordinate category levels in real-world images. Atten. Percept. Psychophys. *80*, 1946–1961. https://doi.org/10.3758/s13414-018-1566-z.

53. Tompary, A., and Davachi, L. (2017). Consolidation Promotes the Emergence of Representational Overlap in the Hippocampus and Medial Prefrontal Cortex. Neuron *96*, 228–241.e5. https://doi.org/10.1016/j.neuron.2017.09.005.

54. Robin, J., and Moscovitch, M. (2017). Details, gist and schema: hippocampal–neocortical interactions underlying recent and remote episodic and spatial memory. Curr. Opin. Behav. Sci. *17*, 114–123. https://doi.org/10.1016/j.cobeha.2017.07.016.

55. Schapiro, A.C., McDevitt, E.A., Rogers, T.T., Mednick, S.C., and Norman, K.A. (2018). Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. Nat. Commun. *9*, 3920. https://doi.org/10.1038/s41467-018-06213-1.

56. Carlson, T., Grol, M.J., and Verstraten, F.A.J. (2006). Dynamics of visual recognition revealed by fMRI. Neuroimage *32*, 892–905. https://doi.org/10.1016/j.neuroimage.2006.03.059.

57. Kim, H., and Cabeza, R. (2009). Common and specific brain regions in high- versus low-confidence recognition memory. Brain Res. *1282*, 103–113. https://doi.org/10.1016/j.brainres.2009.05.080.

58. Kok, P., and Turk-Browne, N.B. (2018). Associative Prediction of Visual Shape in the Hippocampus. J. Neurosci. *38*, 6888–6899. https://doi.org/10.1523/JNEUROSCI.0163-18.2018.

59. Kok, P., Rait, L.I., and Turk-Browne, N.B. (2020). Content-based Dissociation of Hippocampal Involvement in Prediction. J. Cognit. Neurosci. *32*, 527–545. https://doi.org/10.1162/jocn_a_01509.

60. Ellis, C.T., Skalaban, L.J., Yates, T.S., Bejjanki, V.R., Córdova, N.I., and Turk-Browne, N.B. (2021). Evidence of hippocampal learning in human infants. Curr. Biol. *31*, 3358–3364.e4. https://doi.org/10.1016/j.cub.2021.04.072.

61. Aly, M., and Turk-Browne, N.B. (2016). Attention promotes episodic encoding by stabilizing hippocampal representations. Proc. Natl. Acad. Sci. USA *113*, E420–E429. https://doi.org/10.1073/pnas.1518931113.

62. Hindy, N.C., Ng, F.Y., and Turk-Browne, N.B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. Nat. Neurosci. *19*, 665–667. https://doi.org/10.1038/nn.4284.

63. Ólafsdóttir, H.F., Bush, D., and Barry, C. (2018). The Role of Hippocampal Replay in Memory and Planning. Curr. Biol. *28*, R37–R50. https://doi.org/10.1016/j.cub.2017.10.073.

64. Carr, M.F., Jadhav, S.P., and Frank, L.M. (2011). Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. Nat. Neurosci. *14*, 147–153. https://doi.org/10.1038/nn.2732.

65. Rasch, B., and Born, J. (2007). Maintaining memories by reactivation. Curr. Opin. Neurobiol. *17*, 698–703. https://doi.org/10.1016/j.conb.2007.11.007.

66. Gupta, A.S., van der Meer, M.A.A., Touretzky, D.S., and Redish, A.D. (2010). Hippocampal Replay Is Not a Simple Function of Experience. Neuron *65*, 695–705. https://doi.org/10.1016/j.neuron.2010.01.034.

67. Chun, M.M., and Phelps, E.A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. Nat. Neurosci. *2*, 844–847. https://doi.org/10.1038/12222.

68. Foster, D.J., and Wilson, M.A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature *440*, 680–683. https://doi.org/10.1038/nature04587.

69. Skaggs, W.E., and McNaughton, B.L. (1996). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. Science *271*, 1870–1873. https://doi.org/10.1126/science.271.5257.1870.

70. Wilson, M.A., and McNaughton, B.L. (1994). Reactivation of Hippocampal Ensemble Memories During Sleep. Science *265*, 676–679. https://doi.org/10.1126/science.8036517.

71. Singh, D., Norman, K.A., and Schapiro, A.C. (2022). A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation. Proc. Natl. Acad. Sci. USA *119*, e2123432119. https://doi.org/10.1073/pnas.2123432119.

72. Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., and Wager, T.D. (2011). Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods *8*, 665–670. https://doi.org/10.1038/nmeth.1635.

73. Schapiro, A.C., Turk-Browne, N.B., Botvinick, M.M., and Norman, K.A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. Philos. Trans. R. Soc. Lond. B Biol. Sci. *372*, 20160049. https://doi.org/10.1098/rstb.2016.0049.

74. Schapiro, A.C., Turk-Browne, N.B., Norman, K.a., and Botvinick, M.M. (2016). Statistical learning of temporal community structure in the hippocampus. Hippocampus *26*, 3–8. https://doi.org/10.1002/hipo.22523.

75. Brown, T.I., Whiteman, A.S., Aselcioglu, I., and Stern, C.E. (2014). Structural Differences in Hippocampal and Prefrontal Gray Matter Volume Support Flexible Context-Dependent Navigation Ability. J. Neurosci. *34*, 2314–2320. https://doi.org/10.1523/JNEUROSCI.2202-13.2014.

76. Bohbot, V.D., Lerch, J., Thorndycraft, B., Iaria, G., and Zijdenbos, A.P. (2007). Gray Matter Differences Correlate with Spontaneous Strategies in a Human Virtual Navigation Task. J. Neurosci. *27*, 10078–10083. https://doi.org/10.1523/JNEUROSCI.1763-07.2007.

77. Carr, V.A., Bernstein, J.D., Favila, S.E., Rutt, B.K., Kerchner, G.A., and Wagner, A.D. (2017). Individual differences in associative memory among older adults explained by hippocampal subfield structure and function. Proc. Natl. Acad. Sci. USA *114*, 12075–12080. https://doi.org/10.1073/pnas.1713308114.

78. Chrastil, E.R., Sherrill, K.R., Aselcioglu, I., Hasselmo, M.E., and Stern, C.E. (2017). Individual differences in human path integration abilities correlate with gray matter volume in retrosplenial cortex, hippocampus, and medial prefrontal cortex. eNeuro *4*, ENEURO.0346-16.2017. https://doi.org/10.1523/ENEURO.0346-16.2017.

79. van Eijk, L., Hansell, N.K., Strike, L.T., Couvy-Duchesne, B., de Zubicaray, G.I., Thompson, P.M., McMahon, K.L., Zietsch, B.P., and Wright, M.J. (2020). Region-specific sex differences in the hippocampus. Neuroimage *215*, 116781. https://doi.org/10.1016/j.neuroimage.2020.116781.

80. Iaria, G., Petrides, M., Dagher, A., Pike, B., and Bohbot, V.D. (2003). Cognitive Strategies Dependent on the Hippocampus and Caudate Nucleus in Human Navigation: Variability and Change with Practice. J. Neurosci. *23*, 5945–5952. https://doi.org/10.1523/JNEUROSCI.23-13-05945.2003.

81. Iaria, G., Lanyon, L.J., Fox, C.J., Giaschi, D., and Barton, J.J.S. (2008). Navigational skills correlate with hippocampal fractional anisotropy in humans. Hippocampus *18*, 335–339. https://doi.org/10.1002/hipo.20400.

82. Maguire, E.A., Gadian, D.G., Johnsrude, I.S., Good, C.D., Ashburner, J., Frackowiak, R.S., and Frith, C.D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. Proc. Natl. Acad. Sci. USA *97*, 4398–4403. https://doi.org/10.1073/pnas.070039597.

83. Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., and Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage *31*, 1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015.

84. Yushkevich, P.A., Pluta, J.B., Wang, H., Xie, L., Ding, S.-L., Gertje, E.C., Mancuso, L., Kliot, D., Das, S.R., and Wolk, D.A. (2015). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. Hum. Brain Mapp. *36*, 258–287. https://doi.org/10.1002/hbm.22627.

85. Rader, A.W., and Sloutsky, V.M. (2002). Processing of logically valid and logically invalid conditional inferences in discourse comprehension. J. Exp. Psychol. Learn. Mem. Cogn. *28*, 59–68. https://doi.org/10.1037/0278-7393.28.1.59.

86. Luo, J., and Niki, K. (2003). Function of hippocampus in "insight" of problem solving. Hippocampus *13*, 316–323. https://doi.org/10.1002/hipo.10069.

87. Addis, D.R., Moscovitch, M., Crawley, A.P., and McAndrews, M.P. (2004). Recollective qualities modulate hippocampal activation during autobiographical memory retrieval. Hippocampus *14*, 752–762. https://doi.org/10.1002/hipo.10215.

88. Kwon, M., Lee, S.W., and Lee, S.H. (2023). Hippocampal integration and separation processes with different temporal and spatial dynamics during learning for associative memory. Hum. Brain Mapp. *44*, 3873–3884. https://doi.org/10.1002/hbm.26319.

89. Brunec, I.K., Bellana, B., Ozubko, J.D., Man, V., Robin, J., Liu, Z.-X., Grady, C., Rosenbaum, R.S., Winocur, G., Barense, M.D., and Moscovitch, M. (2018). Multiple Scales of Representation along the Hippocampal Anteroposterior Axis in Humans. Curr. Biol. *28*, 2129–2135.e6. https://doi.org/10.1016/j.cub.2018.05.016.

90. Theves, S., Fernandez, G., and Doeller, C.F. (2019). The Hippocampus Encodes Distances in Multidimensional Feature Space. Curr. Biol. *29*, 1226–1231.e3. https://doi.org/10.1016/j.cub.2019.02.035.

91. Morton, N.W., Zippi, E.L., and Preston, A.R. (2023). Memory reactivation and suppression modulate integration of the semantic features of related memories in hippocampus. Cerebr. Cortex *33*, 9020–9037. https://doi.org/10.1093/cercor/bhad179.

92. Malykhin, N.V., Lebel, R.M., Coupland, N.J., Wilman, A.H., and Carter, R. (2010). In vivo quantification of hippocampal subfields using 4.7 T fast spin echo imaging. Neuroimage *49*, 1224–1230. https://doi.org/10.1016/j.neuroimage.2009.09.042.

93. Collin, S.H.P., Milivojevic, B., and Doeller, C.F. (2015). Memory hierarchies map onto the hippocampal long axis in humans. Nat. Neurosci. *18*, 1562–1564. https://doi.org/10.1038/nn.4138.

94. Borzello, M., Ramirez, S., Treves, A., Lee, I., Scharfman, H., Stark, C., Knierim, J.J., and Rangel, L.M. (2023). Assessments of dentate gyrus function: discoveries and debates. Nat. Rev. Neurosci. *24*, 502–517. https://doi.org/10.1038/s41583-023-00710-z.

95. Pudhiyidath, A., Morton, N.W., Viveros Duran, R., Schapiro, A.C., Momennejad, I., Hinojosa-Rowland, D.M., Molitor, R.J., and Preston, A.R. (2022). Representations of Temporal Community Structure in Hippocampus and Precuneus Predict Inductive Reasoning Decisions. J. Cognit. Neurosci. *34*, 1736–1760. https://doi.org/10.1162/jocn_a_01864.

96. Lagnado, D.A., and Sloman, S.A. (2006). Time as a guide to cause. J. Exp. Psychol. Learn. Mem. Cogn. *32*, 451–460. https://doi.org/10.1037/0278-7393.32.3.451.

97. Oaksford, M., and Chater, N. (2010). Causation and Conditionals in the Cognitive Science of Human Reasoning. Open Psychol. J. *3*, 105–118. https://doi.org/10.2174/1874350101003020105.

98. Greene, A.J., Spellman, B.A., Dusek, J.A., Eichenbaum, H.B., and Levy, W.B. (2001). Relational learning with and without awareness: Transitive inference using nonverbal stimuli in humans. Mem. Cognit. *29*, 893–902. https://doi.org/10.3758/BF03196418.

99. D'amato, M.R., Salmon, D.P., Loukas, E., and Tomie, A. (1985). SYMMETRY AND TRANSITIVITY OF CONDITIONAL RELATIONS IN MONKEYS

(CEBUS APELLA) AND PIGEONS (COLUMBA LIVIA). J. Exp. Anal. Behav. *44*, 35–47. https://doi.org/10.1901/jeab.1985.44-35.

100. Holmes, P.W. (1979). Transfer of matching performance in pigeons. J. Exp. Anal. Behav. *31*, 103–114. https://doi.org/10.1901/jeab.1979.31-103.
101. Lipkens, R., Kop, P.F., and Matthijs, W. (1988). A test of symmetry and transitivity in the conditional discrimination performances of pigeons. J. Exp. Anal. Behav. *49*, 395–409. https://doi.org/10.1901/jeab.1988.49-395.
102. McIntire, K.D., Cleary, J., and Thompson, T. (1987). Conditional relations by monkeys: reflexivity, symmetry, and transitivity. J. Exp. Anal. Behav. *47*, 279–285. https://doi.org/10.1901/jeab.1987.47-279.
103. Richards, R.W. (1988). The question of bidirectional associations in pigeons' learning of conditional discrimination tasks. Bull. Psychonomic Soc. *26*, 577–579. https://doi.org/10.3758/BF03330126.
104. Dugdale, N., and Lowe, C.F. (2000). Testing for symmetry in the conditional discriminations of language-trained chimpanzees. J. Exp. Anal. Behav. *73*, 5–22. https://doi.org/10.1901/jeab.2000.73-5.
105. Lionello-DeNolf, K.M., and Urcuioli, P.J. (2002). Stimulus control topographies and tests of symmetry in pigeons. J. Exp. Anal. Behav. *78*, 467–495. https://doi.org/10.1901/jeab.2002.78-467.
106. García, A., and Benjumea, S. (2006). The emergence of symmetry in a conditional discrimination task using different responses as propioceptive samples in pigeons. J. Exp. Anal. Behav. *86*, 65–80. https://doi.org/10.1901/jeab.2006.67-04.
107. Quian Quiroga, R. (2020). No Pattern Separation in the Human Hippocampus. Trends Cognit. Sci. *24*, 994–1007. https://doi.org/10.1016/j.tics.2020.09.012.
108. Redshaw, J. (2014). Does metarepresentation make human mental time travel unique? WIREs Cognitive Science *5*, 519–531. https://doi.org/10.1002/wcs.1308.
109. Turk-Browne, N.B., and Scholl, B.J. (2009). Flexible visual statistical learning: Transfer across space and time. J. Exp. Psychol. Hum. Percept. Perform. *35*, 195–202. https://doi.org/10.1037/0096-1523.35.1.195.
110. Park, S.H., Rogers, L.L., and Vickery, T.J. (2018). The roles of order, distance, and interstitial items in temporal visual statistical learning. Atten. Percept. Psychophys. *80*, 1409–1419. https://doi.org/10.3758/s13414-018-1556-1.
111. Chun, M.M., and Turk-Browne, N.B. (2008). Associative Learning Mechanisms in Vision. In Visual Memory (Oxford University Press), pp. 209–246. https://doi.org/10.1093/acprof:oso/9780195305487.003.0007.
112. Robertson, E.M. (2009). From Creation to Consolidation: A Novel Framework for Memory Processing. PLoS Biol. *7*, e1000019. https://doi.org/10.1371/journal.pbio.1000019.
113. Robertson, E.M. (2018). Memory instability as a gateway to generalization. PLoS Biol. *16*, e2004633. https://doi.org/10.1371/journal.pbio.2004633.
114. Mosha, N., and Robertson, E.M. (2016). Unstable Memories Create a High-Level Representation that Enables Learning Transfer. Curr. Biol. *26*, 100–105. https://doi.org/10.1016/j.cub.2015.11.035.
115. Robertson, E.M., Pascual-Leone, A., and Miall, R.C. (2004). Current concepts in procedural consolidation. Nat. Rev. Neurosci. *5*, 576–582. https://doi.org/10.1038/nrn1426.
116. Cheng, S., Werning, M., and Suddendorf, T. (2016). Dissociating memory traces and scenario construction in mental time travel. Neurosci. Biobehav. Rev. *60*, 82–89. https://doi.org/10.1016/j.neubiorev.2015.11.011.
117. Suddendorf, T., and Corballis, M.C. (1997). Mental time travel and the evolution of the human mind. Genet. Soc. Gen. Psychol. Monogr. *123*, 133–167. https://doi.org/10.1093/acprof:oso/9780195395518.003.0121.
118. Quiroga, R.Q. (2012). Concept cells: the building blocks of declarative memory functions. Nat. Rev. Neurosci. *13*, 587–597. https://doi.org/10.1038/nrn3251.
119. Brainard, D.H. (1997). The Psychophysics Toolbox. Spatial Vis. *10*, 433–436. https://doi.org/10.1163/156856897X00357.
120. Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spat. Vis. *10*, 437–442. https://doi.org/10.1163/156856897X00366.
121. Craik, F.I., and Lockhart, R.S. (1972). Levels of processing: A framework for memory research. J. Verb. Learn. Verb. Behav. *11*, 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X.
122. Tambini, A., and Davachi, L. (2019). Awake Reactivation of Prior Experiences Consolidates Memories and Biases Cognition. Trends Cognit. Sci. *23*, 876–890. https://doi.org/10.1016/j.tics.2019.07.008.
123. Ye, Z., Shi, L., Li, A., Chen, C., and Xue, G. (2020). Retrieval practice facilitates memory updating by enhancing and differentiating medial prefrontal cortex representations. Elife *9*, e57023. https://doi.org/10.7554/eLife.57023.
124. O'Neill, J., Pleydell-Bouverie, B., Dupret, D., and Csicsvari, J. (2010). Play it again: reactivation of waking experience and memory. Trends Neurosci. *33*, 220–229. https://doi.org/10.1016/j.tins.2010.01.006.
125. Nader, K. (2003). Memory traces unbound. Trends Neurosci. *26*, 65–72. https://doi.org/10.1016/S0166-2236(02)00042-5.
126. Brady, T.F., Konkle, T., Alvarez, G.A., and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. Proc. Natl. Acad. Sci. USA *105*, 14325–14329. https://doi.org/10.1073/pnas.0803390105.
127. Favila, S.E., Lee, H., and Kuhl, B.A. (2020). Transforming the Concept of Memory Reactivation. Trends Neurosci. *43*, 939–950. https://doi.org/10.1016/j.tins.2020.09.006.
128. Koolschijn, R.S., Shpektor, A., Clarke, W.T., Ip, I.B., Dupret, D., Emir, U.E., and Barron, H.C. (2021). Memory recall involves a transient break in excitatory-inhibitory balance. Elife *10*, e70071. https://doi.org/10.7554/eLife.70071.
129. Griswold, M.A., Jakob, P.M., Heidemann, R.M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., and Haase, A. (2002). Generalized autocalibrating partially parallel acquisitions (GRAPPA). Magn. Reson. Med. *47*, 1202–1210. https://doi.org/10.1002/mrm.10171.
130. Mugler, J.P., and Brookeman, J.R. (1990). Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). Magn. Reson. Med. *15*, 152–157. https://doi.org/10.1002/mrm.1910150117.
131. Cox, R.W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. Comput. Biomed. Res. *29*, 162–173. https://doi.org/10.1006/cbmr.1996.0014.
132. Glover, G.H., Li, T.Q., and Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn. Reson. Med. *44*, 162–167. https://doi.org/10.1002/1522-2594(200007)44:1<162::aid-mrm23>3.0.co;2-e.
133. Xie, L., Wisse, L.E.M., Das, S.R., Wang, H., Wolk, D.A., Manjón, J.V., and Yushkevich, P.A. (2016). Accounting for the confound of meninges in segmenting entorhinal and perirhinal cortices in T1-weighted MRI. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9901 LNCS, pp. 564–571. https://doi.org/10.1007/978-3-319-46723-8_65.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Anonymized fMRI data and summary behavioral data	This paper	CBS Data Sharing Plattform: https://doi.org/10.60178/cbs.20241223-001
Software and algorithms		
MATLAB 2016a	Mathworks Inc.	RRID: SCR_001622 https://www.mathworks.com/
Psychophysics Toolbox Version 3 (PTB-3)	Brainard (1997)[119]; Pelli (1997)[120]	RRID: SCR_002881 http://psychtoolbox.org/
R Version 3.5.3	R Development Core Team	RRID: SCR_001905 https://www.r-project.org/
Original code	This paper	Github: https://github.com/yokohiguchi/affirming_the_consequent

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The number of participants was determined based on recruitment numbers in recent implicit memory research, which typically involves around 30 participants.[37,110] For each experiment, a total of 30 participants were included in the analysis. Participants were randomly assigned to each experiment.

Experiments 1a, b, and c had 31 (17 women and 14 men, age range 18–23), 30 (14 women and 16 men, age range 18–23), and 31 (13 women and 18 men, age range 18–24) participants, respectively. One participant from Experiment 1a and the other participant from Experiment 1c were excluded from the analysis due to their performance on a cover task in the exposure phase falling below the 95% accuracy criterion, with accuracies of 92.0% and 89.0%, respectively. Experiment 2 had 30 participants (12 women and 18 men, age range 18–22). Experiment 3 had 30 participants (12 women and 18 men, age range 20–25). Experiment 4 had 34 participants (13 women and 21 men, age range 20–31), but four participants were excluded from the analysis for the following reasons: one participant had an arachnoid cyst, two participants encountered an MRI system error and a program error, respectively, and one participant self-reported experiencing headaches and numbness in the arm during the scan, as well as misunderstanding the instructions. Personal details, including ancestry, race, and ethnicity, were not collected. No gender-related effects were observed in the results.

All participants in the current study had normal or corrected-to-normal vision and were not aware of the purpose of the experiment. Written informed consent was obtained from all the participants before enrollment in the study. Experimental protocols were approved by the institutional review boards (Wako3 2019-25).

METHOD DETAILS

Experimental paradigm

In this study, we combined an established paradigm used for the investigation of the affirming-the-consequent fallacy[4,12,27–30] with a protocol used in implicit memory research. This paradigm aimed to investigate the roles of memory in the affirming-the-consequent fallacy while minimizing the effects of explicit reasoning during the recognition of learned transitive relations and controlling for the potential influence of strategic or intentional rehearsal during encoding and post-encoding periods. In this paradigm, fixed transitive relations are repeatedly presented unbeknownst to participants. Indeed, participants were kept unaware of the presentations of the relations while implicit memory of the relations was formed,[42,45–51] thereby minimizing the involvement of explicit reasoning in the affirming-the-consequent fallacy. Furthermore, this paradigm prevents rehearsal by participants during the post-encoding period. Previous studies have provided evidence for the beneficial effects of rehearsal on memory consolidation.[121] Other studies have demonstrated that rehearsal facilitates memory modification, reconstruction, and the integration of old and new memories.[65,122–125] Thus, it is important to avoid the contamination of rehearsal effects so that observed memory changes over time can be attributed to the passage of time. An implicit memory paradigm is well-suited for ensuring this due to its incidental and implicit nature.[39,46] The detailed design and procedures are described below in each experiment's section.

Experiment 1a

Apparatus

Stimulus presentation was controlled by MATLAB (The MathWorks, Inc., Natick, MA, USA) with Psychophysics Toolbox.[119,120] The visual stimuli were presented on an LED monitor (ViewSonic VX2263SMHL, 20 × 14 in.) with a resolution of 1280 × 768 pixels and a

refresh rate of 60Hz. The position of the participant's head was fixed with a chin rest at a visual distance of 57cm from the monitor. Responses were made on a standard computer keyboard.

Stimuli

Forty-eight sequences, each consisting of three objects (e.g., ABC, DEF, GHI), were generated for each participant by combining images of manmade and natural objects. Each participant had a unique subset of 144 objects (72 objects per category) that were randomly selected from the 392 objects (196 objects per category). These images were collected from the database,[126] as well as from Google image search, and photographs taken by the experimenter and friends. Objects were cut out from the images and had no background. Object sequences were generated from the subset to counterbalance the different combinations of object categories. For example, there were 6 unique sequences for the combination of "manmade-manmade-natural," and the number of sequences was equal for each of the 8 possible manmade/natural combinations of three objects. This resulted in 48 object sequences for each participant, with no overlap in objects. The sequences were presented at the center of the display, within 8 degrees of visual angle.

Design and procedure

Exposure phase. Participants were presented with a series of objects, and their task was to judge whether the current object was manmade or natural (i.e., cover task) as accurately as possible by pressing a key on the keyboard. Participants were instructed to maintain a 95% accuracy in the cover task. For incorrect or no responses, participants heard a low-pitched sound. Stimulus-onset asynchrony and interstimulus interval of each object were 1.5 s and 0.75 s, respectively, resulting in each image being presented for 0.75 s.

The experiment began with a practice phase with 30 objects that were randomly generated from the 392 objects, followed by a 10-block exposure phase. Unbeknownst to the participants, each of the 48 object sequences appeared once in every block during the exposure phase. Presenting each sequence 10 times in the exposure phase was intended to induce incidental learning in the participants. The sequences were presented in a random order within each block (e.g., ABCXYZGHILMNDEF …).

Test phase. Before starting the test phase, participants were informed that some sequential patterns had been presented repeatedly during the exposure phase. On each trial in the test phase, participants were presented with a forward and a new sequence (e.g., ABC vs. AEI). They were asked to report which of the two sequences they had already seen during the exposure phase. The order of the two sequences was randomized for each trial. Stimulus-onset asynchrony and interstimulus interval of each object were identical to those in the exposure phase. There was no time limit on the RT.

New sequences were constructed by combining the first, second, and third items from different forward sequences to match the object category (manmade/natural) orders within that trial. For example, the new sequence, AEI, was generated by combining three forward sequences, ABC, DEF, and GHI. Additionally, if the object category of the forward sequence was manmade-manmade-manmade, then the new sequence in that trial was also manmade-manmade-manmade.

The test phase was conducted on two separate days: the first test phase was conducted immediately after the exposure phase, and the second test phase was conducted 24 h after the start of the exposure phase. The test phase on each day consisted of 24 trials. If the same sequences were used in both the first and second test phases, performance in the second test phase could be affected by the first test phase due to memory reconstruction through reactivation mechanisms.[122,127,128] One possibility is that learning could occur with the sequences used in the first test phase, potentially improving performance in the second test phase. Another possibility is that the first test phase might result in integration of memories, leading to a decline in performance during the second test phase. To rule out these possibilities of learning between the two test phases, we used only half of the 48 sequences from the exposure phase for each test phase. In other words, each sequence was used either in the first or second test phase and never appeared on two consecutive days. For example, if the sequence "teapot→wheel→flower" was used on the first day, this sequence was not used again on the second day.

Reaction time analysis

Trials with RTs over 10 s were excluded as outliers. On average, 0.2% of trials were excluded across participants in Experiment 1.

Experiment 1b

The apparatus, stimuli, design, and procedure were identical to those in Experiment 1a, with the exception that in the test phase, participants were presented with backward and new sequences (e.g., CBA vs. IEA) on each trial. These sequences were constructed as follows. First, forward and new sequences were generated using the same method as in Experiment 1a. Second, the generated sequences were reversed in their sequential orders.

Experiment 1c

The apparatus, stimuli, design, and procedure were identical to those in Experiments 1a and b, with the exception that in the test phase, participants were presented with forward and backward sequences (e.g., ABC vs. FED) on each trial. On each trial, the original sequence of the backward sequence (e.g., DEF) was randomly selected from those that matched the object category order (manmade/natural) of the forward sequence.

Experiment 2

The apparatus, stimuli, design, and procedure were identical to those in Experiments 1a, b, and c, with the exception that the test phase on each day consisted of one block of 24 trials containing the forward and backward sequences (e.g., ABC vs. FED), and

another block of 24 trials containing the forward and new sequences (e.g., ABC vs. AEI). The backward and new sequences were generated using the same method as in Experiments 1a and 1b. The order of the two blocks was counterbalanced across participants. In the RT analysis, trials with RTs over 10 s were excluded as outliers. On average, 0.5% of trials were excluded across participants.

Experiment 3

The apparatus, stimuli, design, and procedure were identical to those in Experiments 1a, b, and c, with the exception that in the test phase, participants were asked about recognizing two-object pairs instead of three-object sequences. The aim of this experiment is to test whether each sequence is learned based on temporal proximity among items in the sequence, rather than temporal order or neighboring elements. For this aim, it is necessary to eliminate transitive and neighboring elements from the stimuli used in the test phases of this experiment. Thus, we avoided using three-object sequences such as BCA or CAB because transitive elements like BC or AB remain. Similarly, sequences such as BAC or ACB were not used because neighboring elements like BA or CB are still present. Even with pairs, we avoided combinations such as AB or BC, as they include transitive elements. We specifically used CA as a stimulus to test the effects of temporal proximity as this stimulus does not have either transitive or neighboring elements. Each trial consisted of the temporal-proximity pair and a new pair (e.g., CA vs. IA) in randomized pair order. Objects in each temporal-proximity pair were the third and first items from a forward sequence (i.e., backward sequence without middle item), while the new pair objects were random third and first items of different sequences. The object category (manmade/natural) order was matched between the pairs in each trial.

Experiment 4

In the exposure phase, the apparatus, stimuli, design, and procedure were identical to those in the above experiments. Both the first and second test phases were conducted in an MRI scanner. The visual stimuli were projected onto a frosted glass screen placed above the subject's head with a resolution of 1920 × 1080 pixels and a refresh rate of 60 Hz. The visual angle from the object size was identical to those in the exposure phase. The test phase on each day was divided into 4 fMRI runs ("run" of 194 s, corresponding to a total of 194 echo-planar imaging (EPI) volumes, 176 volumes for the task and 8 volumes for blanks before and after the task), each with 6 trials (24 trials in total per each day). Throughout the runs, participants were presented with forward and backward sequences (e.g., ABC vs. FED) as in Experiment 1c, and were asked to report which of the two sequences they had seen during the exposure phase. Participants saw this question on the screen and delivered an answer within 1.5 s using an MRI-compatible button box. Stimulus-onset asynchrony and interstimulus interval of each image were identical to those in the exposure phase (1.5 s and 0.75 s), but the inter-sequence intervals (ISIs) and intertrial intervals (ITIs) were varied with a range of 2–12 s and 2–15 s, respectively, to jitter the onsets of each sequence for deconvolving event-related fMRI activity.

MRI data were acquired by using a 3T MRI scanner (Prisma, Siemens Medical System) using a 64-channel head coil at the RIKEN Center for Brain Science. Functional images were collected by using an EPI sequence with an echo time (TE) of 30 ms, a repetition time (TR) of 1,000 ms, a flip angle (FA) of 64°, a field of view of 192 × 192 mm^2, a matrix size of 64 × 64, a slice thickness of 3 mm, with 51 contiguous slices oriented parallel to the AC-PC line, an acceleration factor of 2 for the GRAPPA parallel imaging technique,[129] and a multi-band factor of 3. Throughout the entire fMRI session, we confirmed that participants were awake by monitoring their eyes. In addition, their respiratory and cardiac signals were collected by using a pressure sensor and a pulse oximeter, respectively. Prior to the fMRI session, whole-brain anatomical scans were acquired using a 3D T1-weighted MPRAGE[130] with a resolution of 1 × 1 × 1 mm^3 (TR = 1,820 ms, TE = 2.74 ms, inversion time = 917 ms, FA = 8°).

Functional images were corrected for head motion using the AFNI command 3dvolreg.[131] Physiological noise was removed based on the participants' respiratory and cardiac signals collected during the test phase.[132] We conducted the removal of slow drift components and spike noise related to head motion and slice scan time correction with the first slice as a reference using in-house software. No spatial smoothing was applied. Subsequently, the functional images were co-registered to each participant's T1 anatomical images using SPM12.

A standard GLM analysis using SPM12 was conducted. Six events (forward sequence, backward sequence, ISI, key press response, ITI) were modeled as regressors separately for the first and second days. For each voxel, the signal timecourse was regressed against a timecourse of predicted fMRI signal responses created by convolving a canonical haemodynamic response function with the regressors. The response intensity for each ROI was then calculated by averaging beta values for the events across the voxels of the ROI. The fMRI activities for the forward and backward sequences were defined as the contrasts between these two events and ITI as a baseline.

ROIs were functionally defined with NeuroSynth, a meta-analytic tool for identifying loci of activation from published fMRI studies.[72] Using a term-based meta-analysis in NeuroSynth database (https://neurosynth.org), We defined regions found with the search term "recognition memory" with a higher threshold of $p < 0.0001$ ($z > 4.0$) and a spatial extent cluster size of more than 30 voxels as the memory network in this study. This memory network contained 14 regions. These 14 regions were used as ROIs in this study.

The segmentation of the hippocampus was performed using ITK-SNAP[83] and the ASHS toolbox.[84] ASHS was run with the ASHS-PMC-T1 atlas,[133] and the anterior and posterior parts of the hippocampus were anatomically defined on each participant's T1-weighted image.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were performed using MATLAB and R. One-sample t tests and paired t tests were used to analyze differences in participants' choices across conditions in all experiments. A two-way ANOVA was conducted to examine interactions between trial type and day in Experiment 2, while in Experiment 4, two-way ANOVAs were performed to analyze interactions of fMRI activity by sequence type and day for each ROI. Bonferroni correction was applied for all multiple comparisons. The sample size (n), defined as the number of participants included in the analysis, was 30 for each experiment after exclusions. Exclusion criteria are detailed in the experimental model and study participant details section. The statistical details of the experiments can be found in the results section. In the results figures (Figures 3, 4, 5, and 6, S2, and S3), data are presented as horizontal lines representing group means, with dots indicating individual participant values. Error bars represent standard errors of the mean. A nominal significance threshold of $\alpha < 0.05$ was used. No methods were used to assess whether the data met the assumptions of the statistical approach.