# Human relational memory requires time and sleep

Jeffrey M. Ellenbogen*†‡, Peter T. Hu*, Jessica D. Payne*, Debra Titone§, and Matthew P. Walker*‡

*Sleep and Neuroimaging Laboratory, Department of Psychiatry, Beth Israel Deaconess Medical Center, and †Departments of Neurology and Medicine (Sleep Division), Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115; and §Department of Psychology, McGill University, 1205 Doctor Penfield Avenue, Montreal, QC, Canada H3A 1B1

Edited by Edward E. Smith, Columbia University, New York, NY, and approved March 2, 2007 (received for review January 5, 2007)

**Relational memory, the flexible ability to generalize across existing stores of information, is a fundamental property of human cognition. Little is known, however, about how and when this inferential knowledge emerges. Here, we test the hypothesis that human relational memory develops during offline time periods. Fifty-six participants initially learned five "premise pairs" (A>B, B>C, C>D, D>E, and E>F). Unknown to subjects, the pairs contained an embedded hierarchy (A>B>C>D>E>F). Following an offline delay of either 20 min, 12 hr (wake or sleep), or 24 hr, knowledge of the hierarchy was tested by examining inferential judgments for novel "inference pairs" (B>D, C>E, and B>E). Despite all groups achieving near-identical premise pair retention after the offline delay (all groups, >85%; the building blocks of the hierarchy), a striking dissociation was evident in the ability to make relational inference judgments: the 20-min group showed no evidence of inferential ability (52%), whereas the 12- and 24-hr groups displayed highly significant relational memory developments (inference ability of both groups, >75%; $P < 0.001$). Moreover, if the 12-hr period contained sleep, an additional boost to relational memory was seen for the most distant inferential judgment (the B>E pair; sleep = 93%, wake = 69%, $P = 0.03$). Interestingly, despite this increase in performance, the sleep benefit was not associated with an increase in subjective confidence for these judgments. Together, these findings demonstrate that human relational memory develops during offline time delays. Furthermore, sleep appears to preferentially facilitate this process by enhancing hierarchical memory binding, thereby allowing superior performance for the more distant inferential judgments, a benefit that may operate below the level of conscious awareness.**

association | inference | learning | offline

The capacity to flexibly interrelate existing stores of knowledge is a fundamental property of higher learning and one that allows us to make innovative memory decisions in novel situations (1). For example, when studying the United States' highway system for the first time, if you learn that you can travel south on route 95 from Boston to New York and that you can also travel south on route 95 from New York to Washington, DC, you could interrelate these two facts and infer how to travel from Boston to Washington, DC, despite never having learned this information directly.

An established paradigm of such relational learning is transitive inference (refs. 2 and 3; also see ref. 4). One initially learns individual premises, such as A>B and B>C, and, without ever directly learning the relationship of A to C, infers that A>C. Thus, through a process of interrelating this information into a hierarchy (A>B>C), knowledge can be inferred beyond the individual component facts.

Undoubtedly, knowledge of the foundational building blocks (e.g., A>B, B>C, etc., the so-called "premise pairs"; Fig. 1*A*) is essential in offering the potential to make flexible inferences (e.g., A>C, the "inference pair"; Fig. 1*A*) (e.g., refs. 5–7). However, whereas knowledge of premise pairs is necessary, it does not mean the process of inference is assured. There are instances in which inference does not ensue, despite having efficiently learned the individual premise pairs (8). This leads to the testable hypothesis that following learning of the individual building blocks (premise pairs), the development of such relational memory binding evolves "offline" (e.g., over time and without further training).

During the last decade, numerous reports indicate that memory continues to improve during offline periods, most commonly during sleep (9–13). Indeed, offline processing can lead to improved performance in circumstances that might require relational learning. For example, complex motor patterns can initially be learned by "chunking" the entire sequence into smaller sequence strings (14). Subsequent offline periods have been shown to integrate these subunits into a complete, automated, motor-memory program (15). Likewise, several reports demonstrate that posttraining time delays, including sleep, may promote higher-order associations and the ability to generalize across motor-memory representations (16–18). Offline processing can also facilitate the extraction of relationships between recently learned, complex, acoustic patterns, allowing generalization of this knowledge to new linguistic sounds (19).

Here, we directly test the role of offline processing in the facilitation of human relational memory. Specifically, we used the transitive inference task to address two questions. (*i*) Following learning of the premise pair building blocks, is the capacity for relational memory immediately available, or does this inferential ability develop during subsequent offline periods? (*ii*) If there is offline development of inference, does the quantity or quality of this ability depend on the brain states of wake or sleep? Based on the growing body of evidence that favors offline learning, particularly during sleep (9–12), we hypothesized that, following proficient premise pair learning, inference would not automatically ensue but instead would evolve across offline time periods, especially sleep.

We therefore studied separate groups of participants, each having achieved the same level of premise pair learning, and tested them after varying offline time delays of 20 min, 12 hr, or 24 hr. Subjects studied six pairs of novel visual patterns [Fig. 1*A* and see supporting information (SI) Figs. 4 and 5]. Each pair was randomly assigned to a particular hierarchical order. Participants learned these individual premise pairs (represented schematically as A>B, B>C, C>D, D>E, and E>F) to a high degree of proficiency and were subsequently tested after the respective delay periods (Fig. 1*C*). Participants were instructed that they were learning individual comparisons (e.g., B>C) but were not informed of the hierarchical structure (A>B>C>D>E>F) from which inferences could be made (e.g., B>D, C>E, and B>E, Fig. 1*B*). After the time delay, premise pair performance was tested (e.g., B?C) together with novel item combinations never learned (e.g., B?E), thereby probing inferential ability.

Author contributions: J.M.E., P.T.H., J.D.P., D.T., and M.P.W. designed research; J.M.E., P.T.H., and M.P.W. performed research; J.M.E., P.T.H., J.D.P., D.T., and M.P.W. analyzed data; and J.M.E., P.T.H., J.D.P., D.T., and M.P.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

See Commentary on page 7317.

‡To whom correspondence may be addressed at: Sleep and Neuroimaging Laboratory, Department of Psychiatry, FD/Feldberg 862, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA 02115. E-mail: mwalker@hms.harvard.edu or jeffrey_ellenbogen@hms.harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0700094104/DC1.

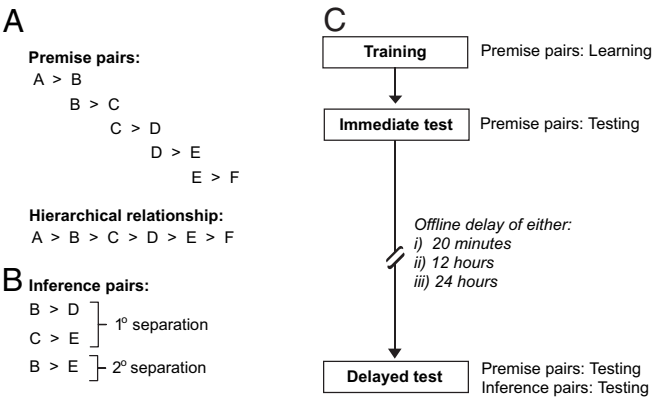


**Fig. 1.** The transitive inference task and experimental paradigm. (*A*) Six visual object stimuli (see SI Fig. 4, illustrated conceptually here as A–F) were combined as five premises (premise pairs), where ''>'' describes the relationship ''should be selected over.'' Unknown to subjects, these premise pairs formed an ordered hierarchy such that A>B>C>D>E>F. (*B*) To evaluate knowledge of the hierarchy, subjects were later tested using the premise pairs and novel ''inference'' pairs not previously learned. These inference pairs involved either one degree (B-D and C-E) or two degrees (B-E) of item separation. (*C*) All participants initially learned the premise pairs during a training session, involving reinforcement cues signifying which item was correct. Immediately following learning, the reinforcement cues were removed, and subjects were tested on the premise pairs to measure the extent of learning without feedback. Following a delayed offline time interval of 20 min, 12 hr, or 24 hr, subjects were again tested on the premise pairs but were also probed for hierarchical knowledge by using the novel inference pairs.

## Results

All groups underwent a three-step sequence of task performance (described in Fig. 1 and *Methods*). In short, subjects (*i*) initially trained on the premise pairs (A>B, B>C, C>D, D>E, and E>F), (*ii*) performed a test session on the premise pairs immediately after learning, and (*iii*) following an offline period of either 20 min, 12 hr, or 24 hr, performed a delayed test session that included the premise pairs together with intermixed testing on the novel, transitive-inference pairs (e.g., B>D, C>E, and B>E).

**Performance at Training and Immediate Testing.** Regardless of group assignment, subjects required similar amounts of initial training to learn the premise pairs to criterion [mean number of blocks: 20 min, 13.1 blocks; 12 hr, 12.5; 24 hr, 13.3; $F(2,51) = 0.49$; $P = 0.95$]. Immediately after training, all subjects were tested on their ability to discern the correct item of a pair. This immediate test procedure was identical to training except no feedback was provided. As described in Table 1, all three groups performed significantly better than chance (50%) on the premise pairs at this immediate test, with mean premise pair knowledge not being significantly different across the three groups (mean correct: 20 min, 84%; 12 hr, 82%; 24 hr, 85%; $F(2, 53) = 0.49$; $P = 0.61$). Therefore, all groups required equivalent numbers of training trials to learn the premise pairs, and all groups were equally able to retain and express knowledge of the premise pairs at the immediate test.

**Performance at Delayed Testing.** Following the offline delay of 20 min, 12 hr, or 24 hr (depending on random group assignment), all groups underwent a second test session, involving presentation of the original premise pairs intermixed with the novel inference pairs. Inference items were comprised of one degree of separation (B-D and C-E) or the more distant two degrees of separation (B-E).

**Delayed Test of Premise Pairs.** Just as all groups demonstrated high levels of premise pair performance at the immediate test, they similarly performed well at the delayed test (Fig. 2*A*; see Table 1 for individual premise pair scores). Moreover, the extent of premise pair retention at the delayed test, as with the immediate test, was nearly identical across the three groups [mean correct: 20 min, 90%; 12 hr, 89%; 24 hr, 89%; $F(2,53) = 0.74$; $P = 0.48$; Fig. 2*A*]. These data indicate that, after a delay period of 20 min, 12 hr, or 24 hr, all groups were similarly equipped with proficient knowledge of the necessary building blocks (premise pairs) required to generate transitive inference.

**Table 1. Mean group performance (percent score) on task pairs at immediate testing and delayed testing**

| | Group | | |
|---|---|---|---|
| Pair | 20-min | 12-hr | 24-hr |
| Immediate test premise pairs | | | |
| A-B | 78 (1.3) | 78 (1.1) | 78 (1.5) |
| B-C | 83 (3.5) | 83 (2.3) | 88 (1.7) |
| C-D | 82 (3.9) | 84 (2.1) | 82 (3.7) |
| D-E | 87 (1.4) | 81 (2.4) | 86 (2.9) |
| E-F | 89 (0.8) | 88 (1.4) | 88 (1.0) |
| Delayed test premise pairs | | | |
| A-B | 94 (2.3) | 95 (2.1) | 89 (7.5) |
| B-C | 88 (6.6) | 86 (4.0) | 88 (6.0) |
| C-D | 92 (4.6) | 85 (3.5) | 94 (2.7) |
| D-E | 90 (2.8) | 84 (4.7) | 85 (7.7) |
| E-F | 93 (4.0) | 94 (1.9) | 96 (2.1) |
| Delayed test inference pairs | | | |
| B-D | 63 (9.5) | 74 (7.0) | 68 (10.2) |
| C-E | 40 (11.8) | 72 (6.5) | 71 (9.4) |
| B-E | 54 (9.8) | 80 (5.6) | 86 (6.0) |
| Delayed test noninference pair | | | |
| A-F | 69 (8.5) | 90 (4.3) | 86 (9.4) |

Values in parentheses represent standard errors.

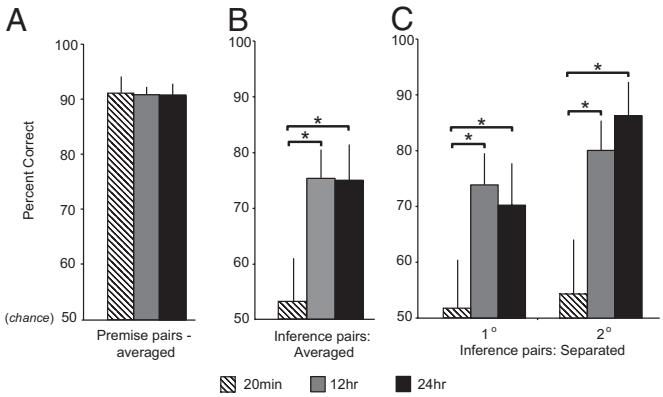


**Fig. 2.** Delayed test performance (percent correct) across groups. (*A*) Premise pair retention following the three delayed time intervals/groups (averaged across A-B, B-C, C-D, D-E, and E-F pairs; individual pair values provided in Table 1). Irrespective of the time delay group, all subjects expressed near-identical premise pair knowledge following the offline period. (*B*) Inference pair performance at the delayed test session across the three groups (averaged across all novel inference pairs, B-D, C-E, and B-E). There was no evidence for the development of hierarchical knowledge after the short (20-min) delay, with inference performance not significantly different than chance. In contrast, following time delays of either 12 or 24 hr, significantly above-chance performance was evident, with accuracy scores significantly different than the 20-min group. Performance between the 12- and 24-hr groups was not significantly different. (*C*) Inference pair performance also at the later delayed test session, across the three groups, but separated according to the distance of item separation, one degree of item separation (averaged B-D and C-E pairs) or two degrees (B-E pair). Asterisks represent significant performance difference ($P < 0.05$). Error bars represent standard errors.

**Delayed Test of Inference Pairs.** In marked contrast to the equality of premise pair performance across all groups, a clear dissociation was evident in the ability to make relational memory judgments on the novel inference pairs (Fig. 2*B*). Specifically, there was a significant effect of group assignment on average inference pair performance [20-min, 12-hr, or 24-hr groups; $F(2, 53) = 3.5$; $P = 0.04$; see Table 1 for individual scores]. Both the 12- and 24-hr conditions exhibited clear development of inference ability (both >75%), yet no inferential knowledge was evident in the 20-min condition. Comparisons among the three conditions demonstrated that averaged inference performance in the 12- and 24-hr groups was significantly greater than that of the 20-min group [two-tailed unpaired *t* test, $t(38) = 3.39$, $P = 0.001$ and $t(23) = 2.51$, $P = 0.02$, respectively; Fig. 2*B*]. Similar results were evident when separating inference on the basis of one- and two-degree distances (Fig. 2*C*; two-tailed unpaired *t* test, all $t \geq 2.06$, $P \leq 0.04$). No difference in averaged or one- and two-degree inference performance was evident between the 12- and 24-hr groups (two-tailed unpaired *t* tests, all $t \leq 0.52$, $P \geq 0.64$).

Indeed, inference performance in the 20-min group was not only lower than the 12- and 24-hr groups, but was not significantly different than chance [one-way *t* test for averaged inference performance relative to chance (50%); $t(11) = 1.10$, $P = 0.29$ (Fig. 2*B*), or when separated according to one- and two-degree inference distance; $t(11) < 0.42$, $P > 0.67$ (see Fig. 2*C* and Table 1 for individual scores)]. Conversely, inference performance in the 12- and 24-hr groups was consistently better than chance, either across the averaged inference scores, or when split across one- or two-degree distances (all $t > 2.69$, $P \leq 0.02$).

Therefore, despite all groups being equally and adequately equipped with premise pair knowledge immediately after training, and after the offline delay, only those experiencing a prolonged consolidation delay (12- and 24-hr) demonstrated relational binding of these elements, affording the ability for successful transitive inference judgments.

**Inference Across Wake and Sleep.** Emerging evidence indicates that different brain states (e.g., wake and sleep) differently impact offline consolidation (9, 10). Therefore, approximately half of the subjects in the 12-hr group were trained in the evening ($n = 14$) and, following a night of sleep, performed the delayed test the next morning (the "Sleep" group). The remaining 12-hr subjects performed the training in the morning and, following an intervening daytime period awake, completed the delayed test later that evening (hereafter called the "Wake" group, $n = 17$).

The Wake and Sleep groups exhibited nearly identical averaged premise pair performance at the immediate test [Sleep group, 83%; Wake group, 82%; two-tailed unpaired *t* test $t(29) = 0.38$, $P = 0.73$]. Likewise, at the 12-hr delayed test, both groups retained knowledge of the premise pairs to a similar extent [Sleep group, 88%; Wake group, 83%; two-tailed unpaired *t* test $t(29) = 1.17$, $P = 0.25$]. Therefore, both groups possessed highly proficient and similar premise pair knowledge for achieving inference judgments (see Table 2).

Although transitive inference performance averaged across all novel pairs (B-D, C-E, and B-E) was numerically higher in the Sleep group (79%) compared with the Wake group (72%), this difference was not significant [two-tailed unpaired *t* test $t(28) = 0.67$, $P = 0.51$]. However, when separating inference performance based on the degree of distance separation, one degree (B-D and C-E) vs. two degrees (B-E), a remarkable qualitative difference emerged (Fig. 3).

By using a two-way mixed ANOVA, comparing Sleep vs. Wake and one degree vs. two degrees of distance as factors, a significant sleep-by-degree interaction was evident [$F(1, 27) = 7.3$, $P = 0.01$; Fig. 3*A*]. There was no difference between the Wake and Sleep conditions in their inference ability across the one degree of item separation [average of B-D and C-E pairs: Sleep, 72%; Wake, 74%; two-tailed unpaired *t* test: $t(29) = 0.18$; $P = 0.86$; Fig. 3*A*]. However, a clear difference was apparent in inference ability across the more distant two degrees of separation, with the Sleep group expressing

**Table 2. Mean performance (percent score) on task pairs at immediate testing and delayed testing in the 12-hr Wake and Sleep groups**

| | Group | |
|---|---|---|
| Pairs | 12-hr Wake | 12-hr Sleep |
| Immediate test premise pairs | | |
| A-B | 78 (1.8) | 79 (1.4) |
| B-C | 79 (3.8) | 87 (1.3) |
| C-D | 86 (2.7) | 82 (3.3) |
| D-E | 81 (3.3) | 80 (3.5) |
| E-F | 87 (2.4) | 89 (1.4) |
| Delayed test premise pairs | | |
| A-B | 93 (3.6) | 98 (1.1) |
| B-C | 83 (6.4) | 91 (4.4) |
| C-D | 84 (4.9) | 86 (5.1) |
| D-E | 81 (6.7) | 87 (6.5) |
| E-F | 92 (3.0) | 97 (1.9) |
| Delayed test inference pairs | | |
| B-D | 75 (9.6) | 73 (10.5) |
| C-E | 73 (8.4) | 71 (10.7) |
| B-E | 69 (8.8) | 93 (4.6) |
| Delayed test noninference pair | | |
| A-F | 88 (5.8) | 94 (6.4) |

Values in parentheses represent standard errors.

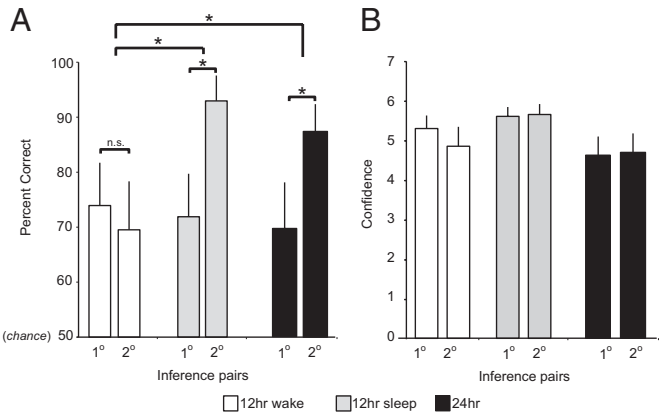


**Fig. 3.** Delayed inference performance and corresponding confidence ratings. (*A*) Delayed inference pair performance (percent correct) across the 12- and 24-hr delay. Twelve-hr groups are separated according to the Wake and Sleep subgroup assignment, and inference performance is separated according to the distance of item separation, one degree of item separation (averaged B-D and C-E pairs) or two (B-E pair; individual pair values provided in Table 1). In the Wake group, performance was not different across the one- and two-degree inference judgments. However, in both groups that experienced an intervening night of sleep (e.g., 12-hr Sleep and 24-hr groups), significantly better performance was expressed on the more distant two-degree inference judgment compared with the one-degree judgment. (*B*) Corresponding confidence ratings (from 1 to 7, with 7 representing the highest confidence) for the one- and two-degree inference pairs in the 12-hr Wake and Sleep subgroups, as well as the 24-hr group. In contrast to the marked performance differences in performance for the one- and two-degree inference in the 12-hr Sleep and 24-hr groups, no corresponding increase in subjective confidence was apparent, in any group. Box-and-whisker plots describing individual subjects' distributions are provided in SI Fig. 5. Mean subjective confidence ratings for the one- and two-degree inference pairs in the 20-min group were 3.7 and 3.4, respectively. Asterisks represents significant performance difference ($P < 0.05$). Error bars represent standard errors.

far greater performance across this larger degree of inference [B-E: Sleep, 93%; Wake, 70%; two-tailed unpaired $t$ test: $t(29) = 2.21$; $P = 0.03$; Fig. 3*A*]. Moreover, there was a significant difference between one- and two-degree inference performance within the Sleep group [two-tailed paired $t$ test: $t(13) = 2.54$; $P = 0.03$] but no such difference within the Wake group [two-tailed paired $t$ test: $t(13) = 0.80$; $P$, 0.43].

One possible explanation for this performance dissociation between the Wake and Sleep groups might be the difference in time of day when the delayed testing took place, evening for the Wake group and morning for the Sleep group. To determine whether inference ability was similarly enhanced following sleep, but when the delayed test took place in the evening, we examined performance in a 24-hr group. Like the Sleep group, the 24-hr subjects were tested following an overnight period containing sleep. Yet, like the Wake group, testing occurred in the evening. As described in Fig. 3*A*, inference performance in the 24-hr group was remarkably similar to the Sleep group, with two-degree inference performance being significantly different than one-degree inference performance [86% and 70%, respectively, two-tailed paired $t$ tests: $t(13) = 3.18$; $P < 0.005$]. Furthermore, as with the Sleep group, comparing the interaction of inference degree and group (24-hr compared with Wake in this instance) by using a two-way mixed ANOVA, a significant group-by-degree interaction was evident [$F(1,27) = 7.8$, $P = 0.01$; Fig. 3*A*]. Thus, a qualitatively different benefit to inference performance was observed across offline periods containing sleep: preferentially facilitating the ability to make more distant inferential judgments, irrespective of circadian test time.

Given this sleep enhancement for the more distant pairs (e.g., B-E), we examined the posttest questionnaire reports, evaluating whether participants were aware of this benefit by probing the confidence of their responses. Subjects provided confidence ratings for the inference pairs on a scale of one to seven (seven was most confident). Surprisingly, despite transitive inference performance for the two-degree distance being better for the Sleep relative to the Wake group (Fig. 3*A*), these performance benefits were not commensurate with a significant increase in confidence for those answers, which were similar across both groups (confidence of B-E pair: Sleep, 5.6; Wake, 4.8; $P = 0.16$; Fig. 3*B*). Thus, although the groups that experienced sleep expressed superior performance for the two-degree inference judgment, this benefit did not appear to be explicitly evident to the subjects, based on confidence ratings.

## Discussion

The ability to interrelate existing stores of information is a fundamental property of human memory. This flexible process of association allows for the realization of novel relationships within previously learned sets of information. Here, we explored the role of offline processing, including time awake and time containing sleep, in the development of relational memory. In the case of transitive inference, learning of the individual components (the so-called premise pairs, e.g., A>B, B>C, C>D, etc.) is known to be critical for the capability of inference (e.g., B>D) (5–7). However, here we demonstrate that the acquisition of premise pairs is necessary, but not sufficient, for inferential knowledge to develop. Instead, offline periods following learning appear to be necessary to trigger the evolution of relational knowledge, a capacity that was not present shortly after learning. These findings center around three related topics: (*i*) the time course for developing inference, (*ii*) the brain state dependency of inference (e.g., wake and sleep), and (*iii*) the dissociation between performance and awareness of inference, each of which we now discuss in turn.

**Time Course.** Immediately following training, all subjects demonstrated highly proficient knowledge of the premise pair building blocks of inference. Likewise, at the subsequent delayed test session, subjects maintained knowledge of these premise pairs, levels that were near equivalent across groups. Therefore, regardless of the length of the offline delay (20 min, 12 hr, or 24 hr), all groups had the same potential for generating transitive inference as measured by knowledge of the premise pairs. Despite this equality, however, a striking dissociation was evident in subsequent inference ability: subjects tested shortly after training showed no evidence of inferential ability, displaying chance performance. Yet those tested following delay periods spanning 12 and 24 hr demonstrated a highly significant facilitation of inference. Therefore, somewhere between 20 min and 12 hr, a process of offline binding of the premise pairs developed, resulting in the increased capability for relational judgments.

If, as we claim, inference can take time to develop, then why have previous studies described the ability for inference immediately after training, without the need for such offline delays (5–7, 20, 21)? Common among these past studies, and distinct from our paradigm, is that participants were trained to ceiling levels on the premise pairs. Notably, in one of the few studies where premise pair knowledge was not at a ceiling, using a similar training criterion to the one we report here, healthy participants were not able to express inference ability when tested immediately after training (8). A further difference also pertains to the structure of training that most paradigms use, often presenting the premise pairs in ascending or descending order during initial learning, a method that may facilitate hierarchical knowledge more quickly, relative to the randomized order used here. Collectively, these prior findings suggest that saturating subjects with premise pair training, or training the pairs in order, can lead to inference without the requisite offline processing. However, in the more ecologically valid circumstance, where individual item knowledge (here, the premise pairs) is not learned in an ordered manner nor overtrained to perfection, time, and especially sleep, appear capable of establishing this relational ability.

One potential candidate mechanism of this offline phenomenon is the process of consolidation (22). Numerous studies investigating the consolidation of both declarative and nondeclarative memory have demonstrated the continued modulation of recently acquired information offline (9–12, 17, 18). Here, we extend the knowledge that consolidation processes benefit individual memory items by demonstrating that similar offline delays can also lead to the associative integration of these item elements into a "metamemory representation," from which can emerge flexible behavioral repertoires, such as inferential judgments.

**Brain State.** When testing inference, two qualitatively different judgments were examined: inference involving one degree of separation (e.g., B-D and C-E; Figs. 1*B* and 3) and inference involving two degrees of separation (e.g., B-E; Figs. 1*B* and 3). Analyzing performance separately for these two measures in the 12-hr Wake and Sleep groups resulted in another marked dissociation. Specifically, both groups expressed a similar ability to make inferences across one degree of separation (B-D and C-E), yet a significant 35% advantage was seen following sleep for inferences across the more distant, two-degree judgment (B-E), relative to the wake group. Thus, a qualitative distinction in relational memory was apparent across identical time delays, determined by whether that offline period contained sleep.

An alternative explanation for such a sleep-specific effect, however, might be the diurnal time of testing. The 12-hr Sleep and Wake groups differ not only by the intervening brain state experienced (asleep or awake) but also by the time of day that inference was tested (9 a.m. or 9 p.m.). It might be that the potential for achieving a specific kind of inference (B-E) is higher in the morning than the evening. As a consequence, differences between the Wake group (tested in the evening) and Sleep group (tested in the morning) might simply reflect circadian influences, rather than sleep/wake effects.

This explanation is inadequate, however, when considering in-

ference performance in the 24-hr group. Like the Wake group, those in the 24-hr group were tested in the evening, although, like the Sleep group, they experienced a night of sleep between training and testing. Subjects in the 24-hr group performed similarly to subjects in the Sleep group, showing a disproportionate enhancement of inference ability for the two-degree pair (B-E; Fig. 3*A*). Because the 24-hr group was tested at the same circadian time as the Wake group but had a relational memory profile matching that of the Sleep group, time-of-day differences alone cannot explain the superior inference performance in the Sleep group. Furthermore, initial learning of the premise pairs and test performance on the immediate and delayed tests of the premise pairs were similar across all groups, regardless of whether they were conducted in the evening or the morning. Given the growing corpus of data regarding sleep and learning (9–13), we feel that the most plausible biological candidate modulating such memory enhancements is sleep itself. However, it is possible that an unidentified factor occurring in the night portion of the diurnal cycle may also regulate these effects.

Another alternative explanation for the offline performance improvements seen in this study may be that subjects were consciously ruminating on the premise pairs following learning, and that conscious deliberation led to enhanced inference ability. However, subjects were unaware that the later session would involve testing. Rather, subjects were simply told that they would return for a second session, thereby reducing the likelihood of conscious deliberation, or at least the impetus to deliberate. Moreover, in the event that subjects did ruminate during the extended 12- and 24-hr offline periods, simply as a consequence of initial training, one would expect their later premise pair performance to be similarly improved, which it was not. We therefore consider an offline consolidation process to be the more likely mechanism regulating such inference enhancements, rather than conscious rumination.

The profile of performance in the 24-hr group also indicates that the overnight sleep benefits are not temporary; rather, they persist at least throughout the following day. (Subjects in this group were trained in the evening and tested the following evening.) As such, sleep, even if followed by a day of potentially interfering waking activities, leads to enhanced performance on the most distant inference judgment. It should also be noted that one-degree inference ability (B-D and C-E) was not different across all three groups; hence, any interference or circadian explanation would have to apply only to a very selective form of inference performance. Therefore, the superior postsleep performance cannot be explained by time of day, differences in training or subsequent premise pair knowledge, or daytime interference. Instead, it appears that this relational memory benefit for the most distant inferential judgment in the Sleep and 24-hr groups is a benefit derived from the brain state of sleep itself.

It is interesting to note the similarity between this finding and recent evidence implicating sleep in the enhancement of memory associations (23), the development of flexible, creative information processing (24, 25), and the relational building of component motor-sequence memories (15, 17, 18). Together, these data provide a new and emerging role for sleep in facilitating associative integration of information, a form of memory binding or extracting experience generalities. A potential candidate structure orchestrating these associative effects might be the hippocampus. Numerous studies have emphasized the dependence of transitive inference on the hippocampal integrity (1). Considering that the hippocampus has consistently been implicated in offline memory reprocessing, manifest in neuronal "replay" following learning (e.g., see ref. 26), a speculative hypothesis is that similar neural reactivation during offline periods of wake and (especially) sleep facilitates relational mapping between learned items. Therefore, such offline hippocampal reprocessing may underlie not only the strengthening of individual item memory, but the binding, and hence subsequent flexible use and expression, of acquired declarative memories.

**Awareness.** Offline time periods containing sleep lead to improvements in inference, particularly the most distant relational judgment (B-E). Yet there was no corresponding increase in subjects' confidence for these answers, indicating a potential dissociation between relational memory performance and subjective awareness of this performance. As a consequence, the additional offline sleep benefits were not reflected in participants' certainty about their judgments, despite accuracy levels of >80%.

These findings mirror other examples of relational learning in the absence of awareness. For example, Greene *et al.* (7) demonstrated no correlation between transitive inference performance following learning of a five-pair hierarchy and participants' awareness of this relational knowledge. Likewise, Frank *et al.* (8) and Smith *et al.* (5) have described the expression of transitive inference in subjects unaware of a hierarchical relationship. Although relational memory can be significantly better when subjects are consciously aware of the hierarchy than when they are not (5, 8), it would appear that the manifestation of relational memory is not necessarily dependent on subjective awareness of this knowledge. It is interesting to speculate whether a state of limited awareness is more or less preferential in favoring automatic modes of memory processing. A nonconscious operation might ultimately lead to more efficient use of this information (27).

In summary, here we demonstrate that the process of human relational learning, as indexed by transitive inference, develops during offline delays. Furthermore, the more distant aspects of inference appear to be selectively enhanced following offline time periods that contain sleep. Intriguingly, however, these overnight gains do not appear to be consciously apparent to the individual, suggesting that such benefits operate below the level of awareness. Collectively, these results provide new insights into how and when the process of human relational memory develops, findings that may have important implications for understanding how these memory processes are facilitated, as well as how they deteriorate with age (28) and fail in specific disease states (20).

## Materials and Methods

**Participants.** Potential participants between 18 and 30 years of age completed a screening questionnaire before selection. Study enrollment was precluded on the basis of prescription or psychoactive medication, illicit drug use, or a past or current history of neurological, psychiatric, or sleep disorders. All participants agreed to abstain from alcohol or caffeine during the study and for 24-hr before it. Those participants in groups that spanned overnight periods reported the amount of sleep obtained by way of sleep logs [mean 7.3 hr (SD 0.9 hr)]. The study was approved by the local human studies committee, and all subjects provided written informed consent.

The computerized task was presented in a quiet testing room on a 12.1-inch computer screen using E-prime software (Psychology Software Tools, Inc., Pittsburgh, PA). Visual items were composed of six abstract color patterns; each was normalized for color and luminance, and readily distinguishable from the others (see SI Fig. 4). The patterns were fully counterbalanced in their assignment to subjects within and across groups.

**Procedure.** Fifty-six healthy participants [mean age, 23 years (SD 4.1); 31 females] performed an initial training session, followed by an immediate test session. Depending on random assignment, individuals then experienced a 20-min, 12-hr (wake or sleep), or 24-hr offline delay. Following this time interval, all subjects performed a delayed test session (Fig. 1*C*). Sample sizes for individual groups were as follows: 20-min, $n = 12$; 12-hr, $n = 31$ (divided into a 12-hr Sleep group, $n = 14$, and a 12-hr Wake group, $n = 17$), and 24-hr, $n = 13$.

In the 20-min group, training and all testing took place at 9 p.m. (±30 min). The 12-hr group was trained and immediately tested either in the morning or the evening [9 a.m./p.m. (±30 min),

depending on Wake/Sleep subgroup assignment], and performed the 12-hr delayed test session the following evening or morning [9 p.m./a.m. (±30 min), respectively]. Subjects assigned to the 24-hr group performed the training and immediate test session in the evening [9 p.m. (±30 min)], and returned at the same time the following evening to complete the delayed test session.

**Training.** Training involved the presentation and learning of five object pairs, referred to here as the "premise pairs" and represented schematically as A>B, B>C, C>D, D>E, and E>F, where ">" describes the reinforced relationship "should be selected over" (Fig. 1*A*). Participants were instructed that two visual objects would appear side by side on the screen, one at a time. On each trial, participants saw one of the five premise pairs (either A-B, B-C, C-D, D-E, or E-F). Subjects were instructed to select the correct item, at first by trial and error, but that with practice, they may be able to learn which of the two object items was correct, based on cued feedback. If subjects selected the correct item of the pair, the selected item would move to the upper portion of the screen, revealing a smiling-face reinforcement stimulus underneath it. If they selected the incorrect object, the item would move to reveal no reinforcement stimuli, indicating an incorrect answer. Left or right positions of individual patterns for each pair were counterbalanced, and participants indicated their response by pressing 1 (left index finger) for the left-side stimulus and 0 (right index finger) for the right-side stimulus on a standard keyboard.

Items were organized into blocks, each containing 10 trials. Therefore, each block presented each of the five items twice, once forward and once backward, randomly (e.g., A>B and B<A, where A is the correct selection in both instances), thereby negating any screen location bias. Items within each block were arranged in pseudorandom order to avoid revealing the hierarchy (e.g., items were never allowed to be presented in the chains longer than two pairs of the hierarchy order (e.g., A>B would not be followed by B>C so that A>B>C would not be overtly obvious). Each participant saw a given block no more than once.

All subjects trained to a set criterion, the measure of which was dynamically evaluated throughout the training process. Specifically, all participants first underwent two blocks of training. After completing the third training block, and every block thereafter, performance was automatically scored. Another training block was presented only if the participant did not reach a criterion of >75% on all of the "middle pairs" (e.g., B-D, D-E, E-F). The middle pairs were used for criterion, rather than all pairs, because the middle pairs were the building blocks of inference (e.g., one must learn B>C and C>D to answer the inference question: B?D).

**Immediate Testing.** After training, all subjects were given a 5-min rest and then began the immediate test session to determine their initial retention level of the premise pairs (e.g., A-B, B-C, C-D, D-E, and E-F, tested in random order; Fig. 1*C*). Testing was similar to training, except that reinforcement cues were removed, allowing evaluation of learning extent without feedback. Subjects were informed that they would see the same pairs of visual patterns from training, and informed of the removal of performance feedback. Participants were asked to select the "correct" visual pattern based on the training experience. Premise pairs were presented by using the same block protocol to that of training, with the exception that the test session ended after five blocks without any criterion in place.

**Delayed Testing.** Following the offline time delay of either 20 min, 12 hr, or 24 hr, subjects returned for this second test. At the beginning of the test session, subjects were again informed they would be presented with the item pairs, as with immediate testing, but were also informed that some of the patterns might be combined in novel ways, and, if that happened, to make their "best guess" on that trial. The novel trials include the transitive inference pairs (e.g., B-D and C-E, involving one-degree of separation, and B-E, involving two-degrees of separation; Fig. 1*C*), together with the noninference pair A-F. The noninference pair can be evaluated without consideration of hierarchical relations, because "A" is always reinforced and "F" is never reinforced Thus, the determination of A over F can be achieved noninferentially simply because A is always correct and F is always incorrect. In contrast, the transitive inference pairs (e.g., B-D) can only be evaluated hierarchically because both items (B and D) are equally reinforced during training. The nine pairs (five premise, three inference, and one noninference) were organized into five blocks randomized in presentation order across subjects. Within a block, the pairs were also presented in a randomized order, and each pair was tested twice, allowing a balanced left–right screen assignment (e.g., A?B and B?A).

Following the delayed test, participants completed an automated questionnaire involving presentation of each of the nine pairs, one at a time, in random order, again without feedback. Subjects were asked to (*i*) select the correct picture from each pair by using the keyboard response, as with training and testing, and, after this choice, (*ii*) make a subjective judgment about how confident they were in this answer by responding to a seven-point confidence rating scale (where 7 was the most confident), using a top-row numeric keyboard response.

We thank Edwin Robertson, Robert Stickgold, and Marina Bedny for their thoughtful insights and constructive contributions to the manuscript. This work was supported by National Institute of Health Grants MH48800, MH69935 (to M.P.W.), K30-4095 (to J.M.E.), and T32-7901 (to J.M.E.) and the American Academy of Sleep Medicine (M.P.W.).

1. Eichenbaum H (2004) *Neuron* 44:109–120.
2. Bryant PE, Trabasso T (1971) *Nature* 232:456–458.
3. Dusek JA, Eichenbaum H (1997) *Proc Natl Acad Sci USA* 94:7109–7114.
4. Van Elzakker M, O'Reilly RC, Rudy JW (2003) *Hippocampus* 13:334–340.
5. Smith C, Squire LR (2005) *J Neurosci* 25:10138–10146.
6. Greene AJ, Gross WL, Elsinger CL, Rao SM (2006) *J Cognit Neurosci* 18:1156–1173.
7. Greene AJ, Spellman BA, Dusek JA, Eichenbaum HB, Levy WB (2001) *Mem Cognit* 29:893–902.
8. Frank MJ, O'Reilly RC, Curran T (2006) *Psychol Sci* 17:700–707.
9. Robertson EM, Pascual-Leone A, Miall RC (2004) *Nat Rev Neurosci* 5:576–582.
10. Walker MP, Stickgold R (2006) *Annu Rev Psychol* 10:139–166.
11. Stickgold R, Walker MP (2005) *Trends Neurosci* 28:408–415.
12. Ellenbogen JM, Payne JD, Stickgold R (2006) *Curr Opin Neurobiol* 16:716–722.
13. Norman KA, Newman EL, Perotte AJ (2005) *Neural Netw* 18:1212–1228.
14. Sakai K, Kitaguchi K, Hikosaka O (2003) *Exp Brain Res* 152:229–242.
15. Kuriyama K, Stickgold R, Walker MP (2004) *Learn Mem* 11:705–713.
16. Keele SW, Ivry R, Mayr U, Hazeltine E, Heuer H (2003) *Psychol Rev* 110:316–339.
17. Spencer RM, Sunm M, Ivry RB (2006) *Curr Biol* 16:1001–1005.
18. Cohen DA, Pascual-Leone A, Press DZ, Robertson EM (2005) *Proc Natl Acad Sci USA* 102:18237–18241.
19. Fenn KM, Nusbaum HC, Margoliash D (2003) *Nature* 425:614–616.
20. Titone D, Ditman T, Holzman PS, Eichenbaum H, Levy DL (2004) *Schizophr Res* 68:235–247.
21. Moses SN, Villate C, Ryan JD (2006) *Neuropsychologia* 44:1370–1387.
22. McGaugh JL (2000) *Science* 287:248–251.
23. Stickgold R, Scott L, Rittenhouse C, Hobson JA (1999) *J Cognit Neurosci* 11:182–193.
24. Walker MP, Liston C, Hobson JA, Stickgold R (2002) *Brain Res Cognit Brain Res* 14:317–324.
25. Wagner U, Gais S, Haider H, Verleger R, Born J (2004) *Nature* 427:352–355.
26. Ji D, Wilson MA (2007) *Nat Neurosci* 10:100–107.
27. Venturino M (1991) *J Exp Psychol Hum Percept Perform* 17:677–695.
28. Rapp PR, Kansky MT, Eichenbaum H (1996) *Behav Neurosci* 110:887–897.