Protecting memory from misinformation: Warnings modulate cortical reinstatement during memory retrieval

Jessica M. Karanian[a,1], Nathaniel Rabb[b], Alia N. Wulff[b], McKinzey G. Torrance[b], Ayanna K. Thomas[b], and Elizabeth Race[b]

[a]Department of Psychology, Fairfield University, Fairfield, CT 06824; and [b]Department of Psychology, Tufts University, Medford, MA 02155

Edited by Henry L. Roediger III, Washington University in St. Louis, St. Louis, MO, and approved July 24, 2020 (received for review May 2, 2020)

Exposure to even subtle forms of misleading information can significantly alter memory for past events. Memory distortion due to misinformation has been linked to faulty reconstructive processes during memory retrieval and the reactivation of brain regions involved in the initial encoding of misleading details (cortical reinstatement). The current study investigated whether warning participants about the threat of misinformation can modulate cortical reinstatement during memory retrieval and reduce misinformation errors. Participants watched a silent video depicting a crime (original event) and were given an initial test of memory for the crime details. Then, participants listened to an auditory narrative describing the crime in which some original details were altered (misinformation). Importantly, participants who received a warning about the reliability of the auditory narrative either before or after exposure to misinformation demonstrated less susceptibility to misinformation on a final test of memory compared to unwarned participants. Warned and unwarned participants also demonstrated striking differences in neural activity during the final memory test. Compared to participants who did not receive a warning, participants who received a warning (regardless of its timing) demonstrated increased activity in visual regions associated with the original source of information as well as decreased activity in auditory regions associated with the misleading source of information. Stronger visual reactivation was associated with reduced susceptibility to misinformation, whereas stronger auditory reactivation was associated with increased susceptibility to misinformation. Together, these results suggest that a simple warning can modulate reconstructive processes during memory retrieval and reduce memory errors due to misinformation.

misinformation | cortical reinstatement | eyewitness memory | fMRI

It is hard to imagine that our memories can deceive us. Yet when we remember, it is well established that we reconstruct fragments of stored information from the past, rendering our memories vulnerable to distortion and errors (1). One striking example of memory distortion occurs after exposure to misleading information about a past event, such as an inaccurate news report or suggestive questioning by a prosecutor. Decades of research indicate that exposure to even subtle forms of misinformation can significantly impair memory and lead to memory errors whereby misleading details are remembered as part of an original event (for a review, see ref. 2). Recent research has shown that this memory distortion, termed "the misinformation effect," can be further exacerbated when individuals recall details of an event before exposure to misinformation (3). For example, if participants are asked to remember details of an event immediately after witnessing it, they are more likely to incorporate misleading information about those details into later memory reports than if they did not receive an immediate test of memory (4–11). Thus, while engaging in repeated retrieval typically enhances memory retention (12), repeated retrieval can also enhance suggestibility in the context of misinformation. Given that instances of repeated memory retrieval are frequent in everyday life, ranging from the common retelling of a story to repeated eyewitness questioning during a trial, an important outstanding question is whether, and how, susceptibility to misinformation can be prevented in these contexts.

Warning individuals about the threat of misinformation may be one way to mitigate the effect of misinformation on memory. Prior studies have shown that susceptibility to misinformation can be significantly reduced when participants are warned that information encountered after an event (postevent information) may not be accurate (e.g., refs. 13–20). For example, when participants are told that the source of postevent information may not be credible or reliable, the misinformation effect can be reduced to half of its typical size (for a review, see ref. 21). The majority of prior studies demonstrating an effect of warning on the misinformation effect have provided warnings *after* exposure to misinformation (postwarning), which suggests that one way that warnings could enhance memory accuracy is by affecting retrieval processes. Indeed, prominent theoretical models attribute misinformation errors to faulty reconstructive processes during memory retrieval (1, 22). According to the source misattribution hypothesis, misinformation errors occur when participants retrieve misleading details from memory and misattribute these details to an

Significance

Exposure to misleading information can distort memory for past events (misinformation effect). Here, we show that providing individuals with a simple warning about the threat of misinformation significantly reduces the misinformation effect, regardless of whether warnings are provided proactively (before exposure to misinformation) or retroactively (after exposure to misinformation). In the brain, this protective effect of warning is associated with increased reactivation of sensory regions associated with the original event and decreased reactivation of sensory regions associated with the misleading information. These findings reveal that warnings can protect memory from misinformation by modulating reconstructive processes at the time of memory retrieval and have important practical implications for improving the accuracy of eyewitness testimony as well as everyday memory reports.

Author contributions: J.M.K., A.N.W., A.K.T., and E.R. designed research; J.M.K., N.R., A.N.W., M.G.T., A.K.T., and E.R. performed research; J.M.K., N.R., A.N.W., M.G.T., A.K.T., and E.R. analyzed data; and J.M.K., A.K.T., and E.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: jessica.karanian@fairfield.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2008595117/-/DCSupplemental.

First published August 31, 2020.

original source of information rather than a misleading source of information (e.g., "source confusion"; refs. 23 and 24). A related and complementary view is that the misinformation effect reflects a failure of source monitoring, whereby inaccurate details about an event are reactivated during retrieval and strategic monitoring functions are not engaged to the extent required to determine the correct source or encoding context in which the detail occurred (22).

Consistent with these faulty retrieval hypotheses of memory distortion, prior neuroimaging studies have demonstrated that cortical activity during memory retrieval can distinguish between true and false memories (e.g., refs. 25–30; for a review, see ref. 31). Of particular relevance is a prior study by Stark et al. which demonstrated that memory errors due to misinformation are associated with the reactivation of brain regions that initially represented the misleading details (32). In this study, participants encoded original event details in one modality (visual) and then were exposed to misleading postevent information in a different modality (auditory). During a later test of memory, accurate memory for the original event was associated with greater activity in brain regions associated with the original (visual) source of information (occipital cortex) whereas false memory for the misleading details was associated with greater activity in brain regions associated with the misleading (auditory) source of information (auditory cortex). These results support neural models of episodic memory which propose that memory retrieval, and particularly memory for the source or context of a memory, depends on the content-specific reinstatement of event features established at encoding (cortical reinstatement; refs. 33–37; for reviews, see refs. 38–40). Furthermore, these results extend these models by demonstrating that the reinstatement of misleading event features can lead to misremembering. An important outstanding question is whether warnings can bias this content-specific cortical reinstatement during memory retrieval and reduce misinformation errors.

The current study investigated this question in two experiments. We first conducted a behavioral experiment (experiment 1) to examine whether warning participants about the threat of misinformation can reduce memory errors due to misinformation, specifically in repeated retrieval contexts in which the misinformation effect is potentiated. Only one prior study has investigated the effect of warnings on misinformation errors in repeated retrieval contexts (4) and found that retrieval-enhanced suggestibility was reduced in participants who received a warning after exposure to misleading postevent information. This result aligns with the source monitoring framework and provides preliminary evidence that warnings can reduce susceptibility to misinformation by modulating memory retrieval, even in contexts in which susceptibility to misinformation is enhanced. Experiment 1 aimed to replicate and extend this result by testing whether this protective effect of warning generalizes to warnings provided before participants are exposed to misinformation (prewarning). In addition to influencing memory retrieval, prospective warnings could influence the initial encoding of misleading details into memory, for example by encouraging shallower processing or extra scrutiny of the postevent information (21). If this is the case, then prewarnings may be even more effective than postwarnings in reducing misinformation errors (17).

The experimental procedure used in experiments 1 and 2 is described in Fig. 1. Participants watched a silent video depicting a crime (witnessed event) and were then given an immediate test of recognition memory (initial memory test). Participants then listened to an audio narrative that described the crime (postevent information). The postevent narrative included consistent details (details about the crime that were described accurately), misleading details (details about the crime that were described inaccurately), and neutral details (details about the crime that were described neither accurately nor inaccurately). After the audio narrative, participants were given a final recognition memory test (final memory test) that assessed memory for the original witnessed event. The final memory test included questions about details that had been described accurately in the audio narrative (consistent trials), questions about details that had been described inaccurately in the audio narrative (misleading trials), and questions about details that had been described neutrally in the audio narrative (neutral trials). Importantly, participants were randomly assigned into one of three warning groups: no-warning, prewarning, and postwarning. Participants in the no-warning group did not receive a warning about the reliability of postevent information. Participants in the prewarning group received a warning about the reliability of postevent information prior to exposure to misinformation (before the audio narrative). Participants in the postwarning group received a warning about the reliability of the postevent information after exposure to misinformation (following the audio narrative). If both types of warning (prewarning and postwarning) bias memory retrieval such that details from the original event are retrieved in favor of details from the misleading source of information, memory performance should improve in both the prewarning and postwarning groups, as compared to the no-warning group. However, if prewarning also reduces the initial encoding of misleading details into memory (e.g., encoding of postevent information), prewarning may have a greater protective effect on memory compared to postwarning. An alternative possibility is that prewarnings do not impact memory accuracy in the context of repeated testing when the misinformation effect is potentiated. If this is the case, participants who receive a prewarning may perform similarly in the final test of memory compared to participants who do not receive a warning. Such a result would argue against warnings having a generalizable effect on memory retrieval.

We next conducted a neuroimaging study (experiment 2) to investigate the mechanisms by which warnings influence memory accuracy in context of misinformation. Participants performed the same misinformation task used in experiment 1 in which misleading details were presented in a different modality (auditory) than original event details (visual) while undergoing functional magnetic resonance imaging (fMRI). Analysis focused on imaging data collected during the final memory test in order to test the hypothesis that warnings protect memory from misinformation by biasing reconstructive processes at the time of memory retrieval. We had two primary predictions based on the cortical reinstatement hypothesis and the prior neuroimaging results observed by Stark et al. (32). First, we predicted that warnings would reduce misinformation errors by encouraging the retrieval of details from the original source of information (silent video). If this is the case, then activity in visual regions should be greater during accurate memory decisions in participants who receive a warning compared to participants who do not receive a warning. Second, we hypothesized that warnings would reduce misinformation errors by reducing retrieval of details from the misleading source of information (audio narrative). If this is the case, then activity in auditory regions should be reduced during misleading trials in participants who received a warning compared to those who did not receive a warning. Finally, if changes in content-specific cortical reinstatement are related to memory performance, we predicted that the strength of sensory reactivation in visual and auditory cortex should predict susceptibility to misinformation on the final memory test. Specifically, we hypothesized that the magnitude of cortical reactivation in visual regions should positively scale with memory accuracy in the face of misinformation and the magnitude of cortical reactivation in auditory regions should negatively scale with memory accuracy in the face of misinformation.

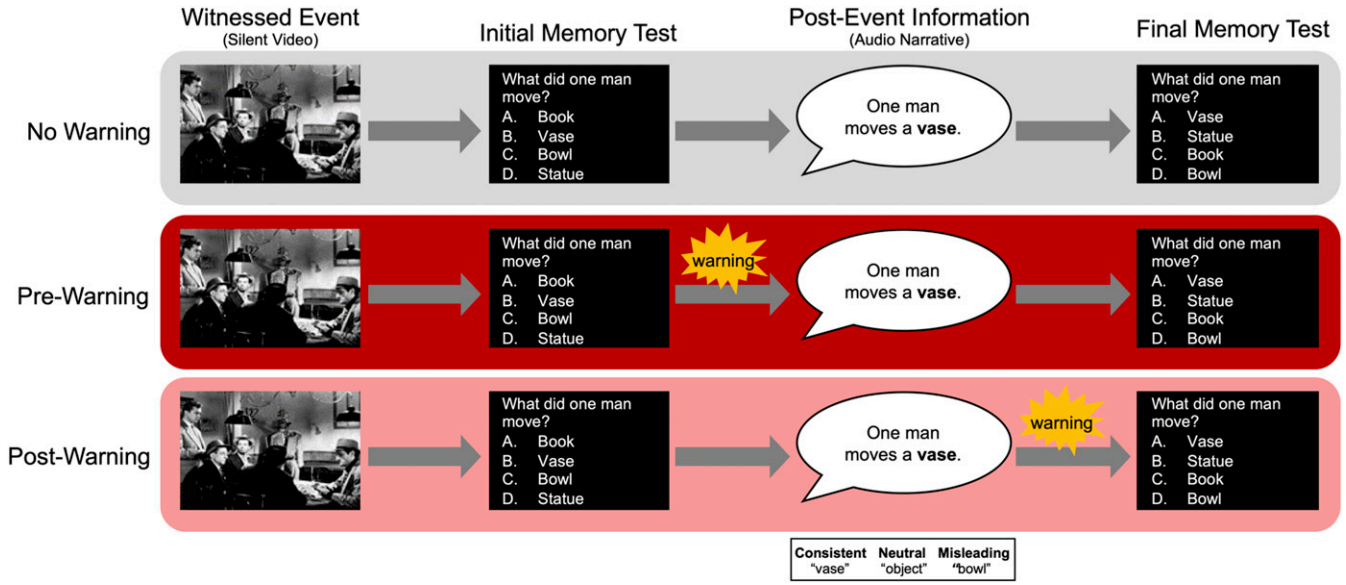

Fig. 1. Eyewitness memory paradigm used in experiments 1 and 2. Participants watched a silent video depicting a crime (witnessed event) and were then given an immediate test of recognition memory (initial memory test). Participants then listened to an audio narrative in which they were provided with postevent information that contained critical details that were either consistent, neutral, or misleading with respect to the original event. After the audio narrative, participants were given a final recognition memory test probing their memory for the original witnessed event. Participants in the no-warning group did not receive a warning about the veracity of postevent information. Participants in the prewarning group received a warning about the veracity of postevent information prior to the audio narrative. Participants in the postwarning group received a warning about the veracity of the postevent information after the audio narrative.

Results

Both Prewarning and Postwarning Reduce Misinformation Errors. In experiment 1, average accuracy on the initial memory test was similar to what has been observed in prior repeated testing paradigms ($M = 0.67$; refs. 3 and 4), with a spontaneous misinformation selection rate of 0.15. Of primary interest was memory accuracy during the final memory test (Fig. 2*A*). Consistent with prior demonstrations of the misinformation effect, recognition memory for original event details differed according to how these details had been described in the audio narrative (consistent/neutral/misleading; $F[2, 156] = 42.93$, $P < 0.001$, $n_p^2 = 0.35$). Pairwise comparisons revealed reduced accuracy for misleading trials ($M = 0.53$) compared to neutral trials ($M = 0.64$; t [80] = 3.39, $P < 0.005$, $d = 0.38$, 95% CI [0.15, 0.60]) and consistent trials ($M = 0.81$; t [80] = 8.34, $P < 0.001$, $d = 0.93$, 95% CI [0.66, 1.19]). Overall memory performance did not significantly differ according to warning group (no-warning, prewarning, postwarning; $F < 1$). Importantly, there was a significant interaction between trial type and warning group ($F[4, 156] = 3.26$, $P = 0.01$, $n_p^2 = 0.08$; *SI Appendix*, Table S1). Warnings affected memory for details that had been altered in the postevent narrative (misleading trials) ($F[2, 78] = 4.05$, $P = 0.02$, $n_p^2 = 0.09$) with no statistically significant cost to memory for details that had been described accurately (consistent trials) ($F[2, 78] = 1.89$, $P = 0.16$, $n_p^2 = 0.05$) or in a neutral manner (neutral trials) ($F < 1$). Interestingly, the increase in memory accuracy observed on misleading trials in participants who received a warning did not depend on whether the warning was given before or after exposure to misinformation. Memory accuracy on misleading trials was greater in both the prewarning group (t [51] = 2.48, $P < 0.05$, $d = 0.72$, 95% CI [0.01, 0.31]) and the postwarning group (t [53] = 2.45, $P < 0.05$, $d = 0.62$, 95% CI [0.004, 0.30]) compared to the no-warning group, and there was not a statistically significant difference in performance between the two warning groups ($t < 1$).

In addition to improving memory accuracy on misleading trials, both prewarning and postwarning reduced the likelihood of selecting the misleading detail during the final forced-choice recognition memory test ($F[2, 78] = 8.82$, $P < 0.001$, $n_p^2 = 0.19$; *SI Appendix*, Table S1). Specifically, pairwise comparisons revealed that warnings reduced misinformation selection in both the prewarning group (t [51] = 3.87, $P < 0.001$, $d = 1.13$, 95% CI [0.09, 0.37]) and the postwarning group (t [53] = 3.35, $P < 0.005$, $d = 0.83$, 95% CI [0.06, 0.34]) compared to the no-warning group. Importantly, there was not a statistically significant effect of warning on misinformation selection for consistent trials ($F[2, 78] = 1.75$, $P = 0.18$, $n_p^2 = 0.04$) or neutral trials ($F < 1$; for Bayesian analyses, see *SI Appendix*, *Results*). These results reveal that 1) warnings can improve memory accuracy in the face of misinformation without significantly reducing memory accuracy for details that were accurately described in the postevent information, and 2) the mnemonic benefits of warning occur regardless of whether warnings are given proactively (before exposure to misinformation) or retroactively (after exposure to misinformation).

Memory Enhancing Effects of Warning Occur Despite Reductions in Confidence. Prior research suggests that the introduction of postevent information that is consistent or inconsistent with a witnessed event can affect the confidence with which one makes memory decisions in addition to affecting memory accuracy (4). To investigate the effect of warning on confidence ratings in the context of misinformation, we performed an exploratory analysis of the average confidence ratings for each trial type (consistent/neutral/misleading; *SI Appendix*, Table S1). Confidence ratings differed across the three trial types ($F[2, 156] = 8.80$, $P < 0.001$, $n_p^2 = 0.10$) as well as across the three warning groups ($F[2, 78] = 3.19$, $P < 0.05$, $n_p^2 = 0.08$). There was also a significant interaction between trial type and warning group ($F[4, 156] = 3.30$, $P = 0.01$, $n_p^2 = 0.08$). Follow-up analyses revealed that warning had a significant effect on confidence ratings for consistent trials

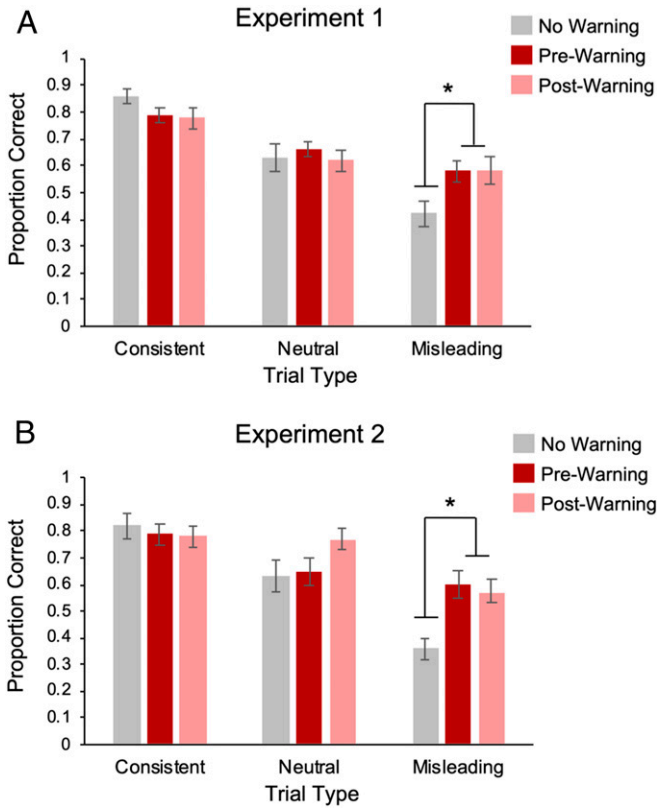

Fig. 2. Behavioral results from experiments 1 and 2. (*A*) Results from the final memory test during the behavioral experiment (experiment 1). (*B*) Results from the final memory test during the fMRI experiment (experiment 2). Proportion correct refers to the proportion of trials within each trial type (consistent, neutral, misleading) that were answered correctly (i.e., the number of trials in which participants selected the correct video detail divided by the total number of trials within that trial type). Error bars indicate between-participant SEs. *$P < 0.05$.

($F[2, 78] = 6.63$, $P < 0.005$, $n_p^2 = 0.15$) with pairwise comparisons revealing reduced confidence ratings in both the prewarning group (t [51] = 2.99, $P < 0.05$, $d = 0.91$, 95% CI [2.54, 22.73]) and the postwarning group (t [53] = 3.30, $P < 0.005$, $d = 0.89$, 95% CI [3.77, 23.59]) compared to the no-warning group. Warning also had a significant effect on confidence ratings for misleading trials ($F[2, 78] = 4.05$, $P = 0.02$, $n_p^2 = 0.09$) with pairwise comparisons revealing that confidence ratings were significantly reduced in the prewarning group (t [51] = 2.56, $P < 0.05$, $d = 0.74$, 95% CI [0.67, 19.51]) and numerically reduced in the postwarning group (t [53] = 2.35, $P = 0.06$, $d = 0.62$, 95% CI [−0.16, 18.33]) compared to the no-warning group. Warning did not have a statistically significant impact on the confidence ratings made during neutral trials ($F < 1$). These results suggest that increases in memory accuracy with warning are accompanied by reductions in participants' confidence in their memory decisions.

Behavioral Effects of Warning Replicate in an Independent Sample during fMRI. The results of experiment 1 revealed that warning participants about threat of misinformation can increase memory accuracy and reduce misinformation errors in the context of repeated memory retrieval. Importantly, this memory benefit did not depend on the timing of the warning: memory accuracy improved both when warnings were provided prior to misinformation exposure (prewarning) and when warnings were provided after exposure to misinformation (postwarning). Experiment 2 used fMRI to investigate the mechanisms by which warnings may confer such a protective effect on memory. Specifically, we tested the hypothesis that both prewarning and postwarning reduce misinformation errors by influencing cortical reinstatement at the time of memory retrieval.

The behavioral results of experiment 2 replicated those of experiment 1 (Fig. 2*B* and *SI Appendix*, Table S1). Average accuracy on the initial memory test (outside the scanner) mirrored that of experiment 1 ($M = 0.66$), with a spontaneous misinformation selection rate of 0.16. Of primary interest was memory accuracy on the final memory test (during scanning). Like experiment 1, a strong misinformation effect was observed on the final memory test ($F[2, 124] = 39.35$, $P < 0.001$, $n_p^2 = 0.39$) with pairwise comparisons revealing that memory accuracy was reduced for misleading trials ($M = 0.50$) compared to neutral trials ($M = 0.68$; t [64] = 4.85, $P < 0.001$, $d = 0.60$, 95% CI [0.34, 0.86]) and consistent trials ($M = 0.79$; t [64] = 7.68, $P < 0.001$, $d = 0.95$, 95% CI [0.66, 1.24]). In addition, there was not a statistically significant effect of warning on overall memory performance ($F < 1$). Importantly, there was a significant interaction between trial type and warning group ($F[4, 124] = 4.44$, $P < 0.001$, $n_p^2 = 0.13$), which reflects that warnings improved memory accuracy on misleading trials ($F[2, 62] = 5.61$, $P = 0.006$, $n_p^2 = 0.15$) at no statistically significant cost to performance on consistent trials ($F < 1$) or neutral trials ($F[2, 62] = 2.35$, $P = 0.10$, $n_p^2 = 0.07$; for Bayesian analyses, see *SI Appendix*, *Results*). Furthermore, the increase in memory accuracy on misleading trials in participants who received a warning did not depend on whether the warning was given before or after exposure to misinformation. That is, both the prewarning group (t [41] = 3.08, $P = 0.009$, $d = 0.95$, 95% CI [0.05, 0.39]) and the postwarning group (t [42] = 2.67, $P = 0.03$, $d = 0.84$, 95% CI [0.02, 0.36]) demonstrated greater memory accuracy on misleading trials compared to the no-warning group, and memory accuracy on misleading trials did not significantly differ between the two warning groups ($t < 1$).

In addition to improving memory accuracy on misleading trials, warnings also reduced the likelihood of selecting the misleading false detail during the final recognition memory test ($F[2, 124] = 5.59$, $P < 0.01$, $n_p^2 = 0.15$; *SI Appendix*, Table S1). Like experiment 1, warnings reduced the likelihood of selecting postevent misinformation on misleading trials ($F[2, 62] = 8.41$, $P < 0.001$, $n_p^2 = 0.21$). Misinformation selection was reduced in both the prewarning group (t [41] = 3.61, $P < 0.005$, $d = 1.17$, 95% CI [0.08, 0.38]) and the postwarning group (t [42] = 3.48, $P < 0.005$, $d = 1.03$, 95% CI [0.07, 0.37]) compared to the no-warning group. Warning did not have a statistically significant impact on the likelihood of choosing misleading details on consistent trials ($F[2, 62] = 1.86$, $P = 0.16$, $n_p^2 = 0.06$) or neutral trials ($F[2, 62] = 1.60$, $P = 0.21$, $n_p^2 = 0.05$). In contrast to experiment 1, a Friedman test revealed that warnings did not have a statistically significant effect on confidence ratings during the final memory test (χ^2 [2] = 4.28, $P = 0.12$; *SI Appendix*, Table S1). However, it is important to note that the confidence ratings during fMRI were made on a different and more restricted scale.

Warnings Increase Activity in the Visual Regions during Memory Retrieval. We next investigated the neural mechanisms by which warnings reduce susceptibility to misinformation by comparing neural activity during the final memory test in the warning (prewarning and postwarning) and no-warning groups. If warnings improve memory accuracy by encouraging retrieval of details from the original source of information (crime video), participants who received a warning should demonstrate greater visual activity during accurate memory decisions compared to participants who did not receive a warning. To test this hypothesis, we analyzed neural activity in bilateral visual regions during accurate memory decisions (correct > incorrect) using

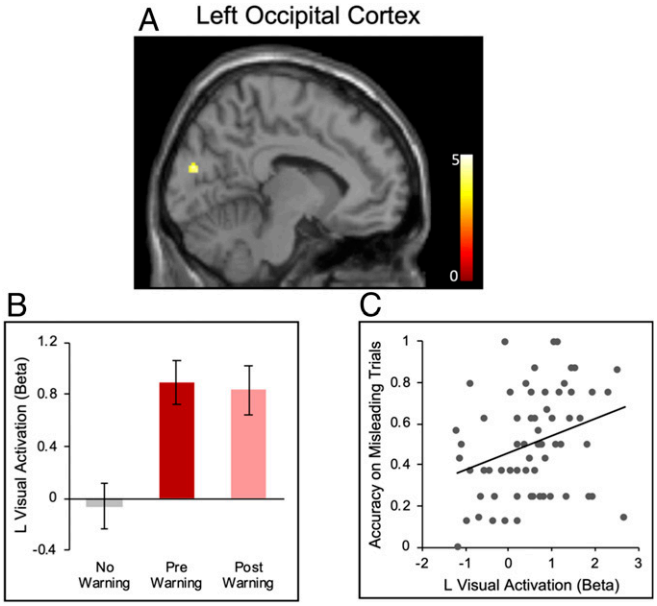

Fig. 3. Warnings increase sensory reactivation in visual processing regions during accurate memory decisions. (*A*) Activity in left occipital cortex (BA 18) was greater during accurate memory decisions (correct > incorrect) in participants who received a warning compared to participants who did not receive a warning. (*B*) Bar graph depicting mean activation (beta weights) within the occipital region of interest as a function of warning group (no-warning, prewarning, postwarning). Error bars indicate between-participant SEs. (*C*) Activity in the occipital cortex during accurate memory decisions was positively associated with memory performance on misleading trials (reduced misinformation effect).

small-volume family-wise error correction. Participants who received a warning demonstrated greater activity in the left occipital cortex (Brodmann area [BA] 18; Montreal Neurological Institute [MNI]: −10, 86, 18) compared to participants who did not receive a warning (Fig. 3*A*). When activity in this region was analyzed separately by group (no-warning, prewarning, and postwarning), there was a main effect of warning, as expected (F [2, 62] = 9.08, $P < 0.001$, $n_p^2 = 0.23$) (Fig. 3*B*). Importantly, post hoc t tests revealed that activity was greater in both the prewarning group (t [41] = 3.78, $P < 0.005$, $d = 1.21$, 95% CI [0.35, 1.56]) and the postwarning group (t [42] = 3.58, $P < 0.005$, $d = 1.05$, 95% CI [0.29, 1.50]) compared to the no-warning group. In addition, there was not a statistically significant difference in activity between the two warning groups (t [41] = 0.25, $P = 0.97$, $d = 0.07$, 95% CI [−0.67, 0.55]). We also conducted an exploratory whole-brain analysis to investigate potential effects of warning outside of this visual region (uncorrected, $P < 0.001$). A strikingly similar effect of warning was observed in the right occipital cortex (BA 18; MNI: 10, −84, 20) and right parahippocampal cortex (BA 36; MNI: 18, −38, −12) (*SI Appendix*, Table S2 and Fig. S1).

The pattern of results in visual regions supports the hypothesis that warnings reduce misinformation errors by reinstating cortical activity associated with the original (visual) source of information. Further support for this proposal comes from analysis of the relationship between the strength of sensory reactivation in visual regions during accurate memory decisions and memory performance across individuals (Fig. 3*C*). The magnitude of occipital cortex activity positively correlated with memory accuracy on misleading trials (r [63] = 0.31, $P = 0.01$, 95% CI [0.07, 0.52]), revealing that participants who demonstrate stronger visual reactivation during memory retrieval commit fewer misinformation errors. A similar relationship between neural activity and behavior was also observed in the right occipital and right parahippocampal regions identified in the whole-brain analysis (*SI Appendix*, Fig. S1).

Warnings Decrease Activity in the Auditory Cortex during Memory Retrieval. The above results support the hypothesis that warnings reduce susceptibility to misinformation by encouraging retrieval of original event details. We next investigated whether warnings also protect from misinformation by reducing the retrieval of information from the misleading (auditory) source. If this is the case, participants who received a warning should demonstrate reduced neural activity in the auditory cortex during misleading trials compared to participants who did not receive a warning. To test this hypothesis, we analyzed neural activity in the bilateral auditory cortex using small-volume family-wise error correction. Activity in the right primary auditory cortex (BA 41; MNI: 54, −30, 10) was reduced in participants who received a warning compared to participants who did not receive a warning during misleading trials (misleading > baseline) (Fig. 4*A*). When activity in this region was analyzed separately by group (no-warning, prewarning, and postwarning), there was a main effect of warning, as expected (F[2, 62] = 11.34, $P < 0.001$, $n_p^2 = 0.27$). Importantly, post hoc t tests revealed significantly reduced activity in both the prewarning group (t [41] = 4.41, $P < 0.001$, $d = 1.30$, 95% CI [0.55, 1.86]) and the postwarning group (t [42] = 3.75, $P < 0.005$, $d = 1.04$, 95% CI [0.36, 1.66]) compared to the no-warning group, and no statistically significant difference in activity between the two warning groups (t [41] = 0.71, $P = 0.76$, $d = 0.25$, 95% CI [−0.46, 0.85]) (Fig. 4*B*). These results align with the hypothesis that both types of warning improve memory accuracy by reducing retrieval from the misleading source of information. Further support for this proposal comes from the analysis of the relationship between auditory reactivation and

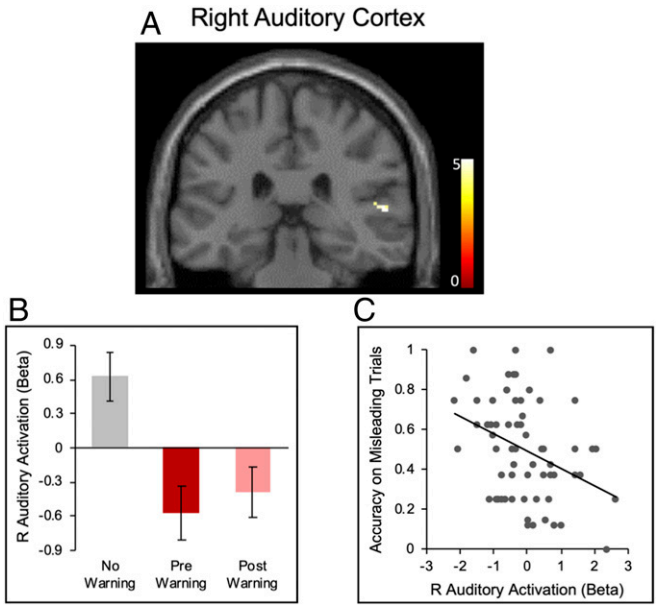

Fig. 4. Warnings decrease sensory reactivation in the auditory cortex on misleading trials. (*A*) Activity in the right primary auditory cortex (BA 41) was greater in participants who did not receive a warning compared to participants who did receive a warning during misleading trials (misleading > baseline). (*B*) Bar graph depicting mean activation (beta weights) within the auditory cortex region of interest as a function of warning group (no-warning, prewarning, postwarning). Error bars indicate between-participant SEs. (*C*) Activity in the auditory cortex during misleading trials was negatively associated with memory performance on misleading trials (increased misinformation effect).

memory performance on misleading trials (Fig. 4*C*). A significant negative correlation was observed (r [63] = −0.35, $P < 0.005$, 95% CI [−0.12, −0.55]), indicating that participants who demonstrate weaker auditory reactivation on misleading trials commit fewer misinformation errors. An exploratory whole-brain analysis (uncorrected, $P < 0.001$) revealed a similar effect of warning in regions outside the auditory cortex (*SI Appendix*, Table S3), including the right anterior prefrontal cortex (BA 10; MNI: 40, 52, −4) and right supramarginal gyrus (BA 40; MNI: 52, −26, 18) (*SI Appendix*, Fig. S2).

Cortical Reinstatement and Hippocampal Activity. Thus far, we have provided evidence that warnings modulate cortical activity during memory retrieval in regions associated with accurate (visual) and inaccurate (auditory) sources of information. We next performed an exploratory analysis to investigate whether this sensory-specific reinstatement was related to activity in the hippocampus using anatomically defined hippocampal regions of interest (ROIs) based on the Automated Anatomical Labeling (AAL) atlas. According to cortical reinstatement theories, the hippocampus coordinates the reactivation of cortical representations during episodic memory retrieval (38, 39, 41–44). If the hippocampus supports the reactivation of both accurate and misleading details (32), activity in the hippocampus should scale with our measures of cortical reinstatement in visual and auditory regions. Indeed, a positive correlation was observed between activity in the left occipital (BA 18) cortex and activity in both the left hippocampus (r [63] = 0.44, $P < 0.001$, 95% CI [0.22, 0.62]) and the right hippocampus (r [63] = 0.47, $P < 0.001$, 95% CI [0.26, 0.64]) during accurate memory decisions (correct > incorrect) (*SI Appendix*, Fig. S3). A positive correlation was also observed during misleading trials (misleading > baseline) between activity in the right auditory cortex (BA 40) and activity in the left hippocampus (r [63] = 0.30, $P = 0.02$, 95% CI [0.06, 0.50]), though not in the right hippocampus (r [63] = 0.16, $P = 0.19$, 95% CI [−0.08, 0.39]) (*SI Appendix*, Fig. S3). Although warnings did not have a significant effect on activity in the left and right hippocampus during memory retrieval, the patterns of effects were consistent across hemispheres and were similar to the effects of warning observed in sensory regions during accurate memory decisions and misleading trials (*SI Appendix*, Fig. S4).

Discussion

Memory is notoriously fallible and susceptible to misinformation. In two experiments, we demonstrated that a simple warning about the threat of misinformation can significantly reduce the negative impact of misinformation on memory. In experiment 1, we found that both prospective warnings (provided before exposure to misinformation) and retrospective warnings (provided after exposure to misinformation) reduce memory errors due to misinformation, even in the context of a repeated testing paradigm in which the negative impact of misinformation on memory is potentiated (3, 4). In experiment 2, we replicated this behavioral effect of warning in the context of fMRI and identified two mechanisms by which warnings influence reconstructive processes in the brain at the time of memory retrieval. First, warnings increased reinstatement of sensory activity associated with accurate event details. Second, warnings decreased reinstatement of sensory activity associated with misleading postevent information. Furthermore, we found that the strength of the content-specific cortical reactivation in visual and auditory regions predicted behavioral performance and susceptibility to misinformation. Together, these results provide insight into the nature of memory distortions due to misinformation and the mechanisms by which misinformation errors can be prevented.

The finding that both prospective and retrospective warnings can mitigate the negative effect of misinformation on memory aligns with numerous prior studies that have demonstrated beneficial effects of warning on memory accuracy (e.g., refs. 14, 16, 18, and 20; for a review, see ref. 21). Importantly, our results extend the beneficial effect of warning to more naturalistic repeated retrieval contexts in which individuals are more susceptible to misleading postevent information (3). Only one prior study has shown that retroactive warnings can mitigate the misinformation effect in the context of repeated memory retrieval (4). The current study provides an important replication of this result and demonstrates that prospective warnings can also reduce misinformation errors in the context of repeated testing. The finding that warnings enhance memory accuracy regardless of whether they are provided before or after exposure to misinformation has important theoretical implications given that repeated testing is thought to enhance suggestibility by enhancing both the encoding of misleading details and the fluency by which these details are later retrieved (4). Although prior work has suggested that prospective warnings may be more effective than retrospective warnings in protecting memory from misinformation, given that prospective warnings could potentially reduce both the initial encoding and ultimate retrieval of misleading details (17, 21), a direct comparison of prospective and retrospective warnings in the current study revealed that both types of warning enhance memory accuracy to a similar extent. Importantly, both types of warning reduced susceptibility to misinformation without significantly reducing memory accuracy for details that were accurately described in the postevent narrative (consistent and neutral details). The selective effect of warning on memory for details that had been altered in the postevent narrative (misleading details) suggests that warnings do not simply cause participants to adopt a more conservative response criterion during memory decisions. Instead, these results are consistent with the hypothesis that both prospective and retrospective warnings can reduce misinformation errors by modulating reconstructive processes at the time of memory retrieval.

Neural measures of sensory reinstatement during the final memory test further support the proposal that warnings reduce misinformation errors by affecting memory reconstruction. Warnings influenced neural activity during retrieval in two distinct ways. First, warnings increased content-specific neural activity during accurate memory decisions in brain regions associated with the encoding of original event details (visual cortical regions). Second, warnings reduced content-specific neural activity in brain regions associated with the encoding of misleading event details (primary auditory cortex). These results align with the proposal that warnings encourage retrieval of details from the original source of information while reducing retrieval from the misleading source of information (e.g., refs. 22–24, 40, and 45) and suggest a neural mechanism by which this is achieved. Specifically, these results suggest that warnings influence memory retrieval by biasing cortical reinstatement.

According to cortical reinstatement theories, memory retrieval involves reactivation of neural regions engaged during encoding (e.g., refs. 33, 34, 36, and 40; for a review, see ref. 38). Our findings are consistent with recent fMRI evidence that cortical reinstatement can lead to both successful remembering as well as misremembering depending on the nature of the reactivated content, and suggest that misinformation errors may be driven in part by the reactivation of representations linked to the misleading source (refs. 25–30, 32, and 46; for a review, see ref. 31). Importantly, we demonstrate that external factors, such as warning, can bias sensory activity in regions that represent accurate as well as misleading details. This provides neural evidence that the retrieval of misleading details is not obligatory and can be biased in favor of the retrieval of accurate details (4). Our finding that the strength of neural responses in both visual

and auditory regions predicts memory performance highlights the behavioral relevance of these sensory-specific responses in the brain: whereas greater visual activity was associated with greater protection from misinformation, greater auditory activity was associated with greater susceptibility to misinformation. While it is difficult to determine the precise nature of the underlying processes driving these sensory effects (e.g., retrieval success versus retrieval attempts), the relationship between activity in these regions and behavioral performance suggests that they are related to the accuracy of memory retrieval. More broadly, these results suggest that the reinstatement of sensory-specific representations during retrieval influences memory-based decisions in the face of misinformation and have real-world applications for improving eyewitness memory. For example, these results suggest that the accuracy of eyewitness memory reports could be improved by eyewitness interview techniques that encourage the mental reinstatement of an original event's context or source over the context or source of postevent information (46).

Guided by theoretical models of episodic memory which propose that the hippocampus coordinates the reinstatement of cortical representations during episodic memory retrieval (38, 39, 41–44), we also examined the relationship between sensory reinstatement and hippocampal activity during the final memory test. The strength of sensory reinstatement in visual processing regions (occipital cortex) positively correlated with the magnitude of hippocampal activity during accurate memory decisions. In addition, the strength of sensory reinstatement in auditory processing regions (auditory cortex) positively correlated with the magnitude of hippocampal activity during misleading trials. These results provide evidence linking cortical reinstatement to hippocampal activity during memory retrieval (e.g., refs. 37, 47, and 48) and are consistent with the proposal that hippocampally mediated retrieval may contribute to both accurate and inaccurate memory decisions (25, 32, 49). In addition, the effects of warning on hippocampal activity mirrored those observed in sensory-specific cortex. Though these effects were weak and should be interpreted with caution, they raise the intriguing possibility that hippocampally mediated reinstatement is not obligatory and can be shaped by external factors such as warning.

An important outstanding question is how warnings bias reinstatement-related activity in the face of misinformation. One possibility is that warnings induce an internal mental state, or retrieval mode, that prioritizes access to relevant information in memory (original event details) over irrelevant information in memory (misleading event details) (50). Indeed, prior behavioral work has suggested that warnings trigger control or monitoring mechanisms that enable memory to be accessed strategically (4, 51). Although we did not observe increased neural activity in brain regions typically associated with memory monitoring when participants were warned, such as frontal or parietal cortex, it is possible that warnings influence more temporally extended control states that could be better detected in fMRI analyses that target more sustained neural responses (52, 53). Warnings could also increase scrutiny of retrieved information, for example by monitoring the source of information retrieved from memory (e.g., postretrieval processing) (17). Interestingly, warned participants demonstrated reduced activity on misleading trials compared to unwarned participants in prefrontal regions (e.g., BA 9/46 and 10) typically associated with effortful retrieval monitoring and reduced memory errors (54). Furthermore, this frontal activity was negatively related to memory performance across participants. Though speculative, this raises the possibility that by reducing the reactivation of inaccurate details, warnings reduce demands on frontal control processes that evaluate or select between competing representations in memory (54, 55). Although the slow temporal resolution of fMRI makes it difficult to determine the onset of memory-related activity in relation to the timing of memory decisions, future research using methods with higher temporal resolution, such as electroencephalography or magnetoencephalography, could explore these possibilities (56).

A second outstanding question is the degree to which the effects of warning on memory retrieval depend on the initial encoding of accurate and misleading details into memory. Although the current study focused on neural activity during memory retrieval, it is well established that processes supporting the encoding and retrieval of episodic information are strongly interdependent (39) and that interactions between representations during both encoding and retrieval contribute to memory distortion (30). Indeed, several prior studies have shown that neural activity during exposure to misinformation predicts subsequent misinformation errors (57, 58). Future research should investigate whether warnings, particularly prewarnings, can reduce susceptibility to misinformation by modulating this encoding-related activity.

Materials and Methods

Experiment 1.

Participants. Eighty-one undergraduate students from Tufts University (25 women) participated in the study (M_{age} = 19 y, SD = 0.98) and were awarded course credit for their participation. Participants were randomly assigned to one of three groups: no-warning (n = 27), prewarning (n = 26), and postwarning (n = 28). A power analysis was conducted to determine our sample size based on prior literature demonstrating retrieval-enhanced suggestibility using a similar paradigm (59). This study was approved by the Institutional Review Board at Tufts University. All participants provided written informed consent.

Stimuli.

Video of witnessed event. During the encoding period, participants viewed a 22-min video clip from the black and white silent film Rififi (60). The video depicts the events surrounding a burglary of a jewelry store and contains no dialogue.

Audio narrative. A narrative synopsis of the witnessed event was recorded by a female speaker and consisted of 115 sentences spoken at a rate of 135 to 160 words per minute. Twenty-four sentences contained critical details that would be probed during the memory tests. These sentences either 1) accurately described a detail (underlined) from the original event (consistent) (e.g., "Revealed at the bottom of the case is a rope."), 2) inaccurately described a detail from the original event (misleading) (e.g., "Revealed at the bottom of the case is a towel."), or 3) provided an alternative (neither consistent nor inconsistent) detail from the original event (neutral) (e.g., "Revealed at the bottom of the case is a useful object."). An equal number of critical sentences contained consistent, misleading, or neutral details. Each critical detail appeared only once during the narrative and the assignment of each detail to the consistent, neutral, or misleading condition was counterbalanced across participants. Critical sentences were separated by at least three filler sentences that were not probed in the memory tests. Following the procedures used in prior misinformation experiments (e.g., ref. 32), all filler sentences were consistent with the content of the video such that the majority of sentences in the audio narrative described accurate information.

Memory tests. Participants were given two memory tests: the initial memory test, which was given immediately after participants viewed the video of the witnessed event, and the final memory test, which was given after participants listened to the audio narrative. In both, participants' memory for the 24 critical details (consistent, misleading, or neutral) from the witnessed event (video) was assessed with a four alternative forced-choice recognition memory test. Both tests consisted of 24 questions that appeared on a computer monitor, one at a time, and asked about a critical detail from the witnessed event. All questions were probed in chronological order (e.g., the same order that they appeared in the video). Four alternative answers were displayed below each question and consisted of the correct detail shown in the video, a misleading detail, and two highly plausible lures (as determined by pilot testing). The order of the four alternative answers was randomized across tests and participants. For each question, participants indicated their response with a button press on a computer keyboard. Participants then indicated their confidence in their response by entering a numerical rating on a sliding scale from 0 (guess) to 100 (very confident) on the computer keyboard. All questions and confidence ratings were self-paced and participants could not return to a question once they had indicated their answer.

For experiment 1, two test questions were excluded from all behavioral analyses due to problematic lures.

Procedure. The experiment consisted of four stages: a video of a witnessed event, an initial memory test, an audio narrative recounting the witnessed event that included misleading details, and a final memory test. Participants first watched the video clip of the witnessed event on a computer monitor. Immediately following the video, participants were administered the initial memory test. Then, participants completed a 10-min filler task in which they completed paper-and-pencil Sudoku puzzles, after which they heard the audio narrative through over-ear headphones. Participants were then given the final memory test. The primary manipulation of interest was whether participants received a warning about the veracity of the postevent information (audio narrative) and the timing of this warning (before or after the audio narrative). Participants were randomly assigned to no-warning, prewarning (warned before the audio narrative), or postwarning (warned after the audio narrative) conditions. Participants in the no-warning group were not warned about the threat of misinformation. Critically, for the warning groups, the instructions either before (prewarning) or after (postwarning) the audio narrative included a warning about the reliability of the audio narrative which informed participants that the accuracy of the narrative could not be verified. The exact wording of the warnings can be found in *SI Appendix, Appendix A*.

Behavioral analysis. For both experiments 1 and 2, recognition memory performance (proportion correct) on the initial and final memory tests was calculated by dividing the number of test trials in which participants selected the correct video detail within each trial type (consistent, neutral, control) by the total number of trials for that given trial type (consistent, neutral, misleading). Misinformation selection was calculated as the proportion of misleading trials in which participants selected the misleading detail that had been inaccurately described in the auditory narrative. As a baseline comparison, we also calculated the proportion of consistent and neutral trials in which participants spontaneously selected a misleading detail (that had not been mentioned in the auditory narrative). Tukey's correction for multiple comparisons was applied to all post hoc *t* tests.

Experiment 2.

Participants. Eighty adult participants aged 18 to 35 were recruited from the Boston area and were compensated $20/h for participation. All participants were right-handed, native English speakers, had normal or corrected-to-normal vision, and reported no history of traumatic head injury. Nineteen participants were excluded prior to fMRI analysis due to technical problems during scanning, noncompliance during the scanning session (e.g., falling asleep), or for being outside the target age range (>35 y old). This yielded a final sample of 65 (no-warning: $n = 22$, prewarning: $n = 21$, postwarning, $n = 22$; $M_{age} = 24$ y, SD = 4; 57% female). In addition, analysis of the initial memory test (outside the scanner) did not include two participants as their data files did not save. This study was approved by the Institutional Review Board at Tufts University. All participants provided written informed consent.

Stimuli.

Video of the witnessed event. The same video of the witnessed event used in experiment 1 was used in experiment 2.

Audio narrative. The audio narrative used in experiment 2 was the same as that used in experiment 1, with the following modifications to accommodate fMRI analysis. First, a jittered interstimulus interval with an average of 6 s (range: 4 to 8 s) was inserted between each sentence of the audio narrative during which a series of arrows pointing left or right was presented. Participants made a left/right button press on a button box indicating the direction of the arrows. Second, some longer sentences presented in the audio narrative in experiment 1 were split into two sentences in the audio narrative in experiment 2 to ensure sentence length was consistent across all trial types (consistent, misleading, neutral). This resulted in a total of 130 sentences in experiment 2 (24 critical, 106 filler). Finally, 5 of the 24 critical sentences were reorganized to ensure that the critical detail was always presented at the end of the sentence.

Memory tests. The same memory tests used in experiment 1 were used in experiment 2, with the following modifications to accommodate for fMRI analysis. For both the initial and final memory test during experiment 2, each question was displayed for 7 s, during which participants made their memory decisions, followed by a 500-ms blank screen. Then, participants were given 3 s to indicate their level of confidence on an ordinal scale that ranged from 1 to 4 with 1 representing guess/low confidence and 4 representing high confidence. During the initial test (outside the scanner), all responses were made on a laptop keyboard. During the final test (inside the scanner), all responses were made on a MRI-compatible button box. During a jittered interstimulus interval which occurred between each memory question (M = 8 s), participants completed the same arrows task described above for the audio narrative. In addition, due to a coding error one sentence near the end of the narrative appeared three times (once in each condition) for a subset of participants ($n = 24$). Behavioral analyses showed no statistically significant differences in response patterns between participants who did or did not hear this erroneous item. Thus, we have no reason to believe that it had a meaningful effect on the present data. Nonetheless, this item was omitted from the behavioral and imaging data.

Procedure. The experiment consisted of the same four stages as experiment 1: a video of a witnessed event, an initial memory test, an audio narrative recounting the witnessed event that included some misleading details, and a final memory test. Participants first watched the video clip of the witnessed event on a computer monitor outside the scanner. Immediately following the video, participants were administered the first memory test on a laptop computer before entering the MRI scanner. Then, participants entered the scanner and both the audio narrative and the final memory test occurred while brain images were acquired. Participants listened to the audio narrative through MRI-compatible earphones and viewed the final memory test questions on an overhead mirror that contained the image of a screen onto which the questions were projected. Each of the questions in the memory test was displayed for 7 s and participants were encouraged to respond within that time. Participants were instructed to minimize movement as much as possible during scanning and indicated their responses on a button box.

Behavioral analysis. The behavioral analyses used in experiment 2 were identical to those used in experiment 1.

MRI data collection and preprocessing. Structural and functional images were acquired on a Siemens 3T Magnetom Prisma Fit scanner (Siemens Medical) with a 32-channel head coil at the Massachusetts Institute of Technology Athinoula A. Martinos Imaging Center. Functional data were acquired using a T2*-weighted echo-planar imaging sequence (TR = 2,000 ms, TE = 30 ms, flip angle = 90°, field of view = 210 × 210 mm, matrix = 64 × 64, slice thickness = 3.0 mm). Forty axial slices parallel to the anterior commissure–posterior commissure (AC-PC) line were obtained. High-resolution structural images of the whole brain were acquired using a T1-weighted, rapid gradient echo-pulse sequence (MPRAGE; TR = 1,800 ms, TE = 2.36 ms, flip angle = 8°, field of view = 250 × 250 mm, slice thickness = 0.87 mm; 208 slices, 0.9 × 0.9 × 0.9 mm resolution).

Image preprocessing and data analysis were performed using SPM12 (Wellcome Department of Cognitive Neurology, London, UK). Functional volumes for each participant were slice-time corrected with the middle slice in time used as a reference and corrected for head motion. The T1-weighted anatomical volume was coregistered to the functional data and segmented into gray and white matter. Segmented images were used to calculate spatial normalization parameters to MNI space, which were then applied to the functional data. As part of spatial normalization, data were resampled to 2 × 2 × 2 mm. Functional images were then spatially smoothed with a 4-mm full width at half maximum (FWHM) Gaussian kernel.

fMRI analysis. Functional data from the second memory test were analyzed using two separate general linear models (GLMs). The first model included separate regressors for accurate memory trials (correct) and inaccurate memory trials (incorrect), collapsed over confidence to increase power. The second model included separate regressors for each trial type (consistent, misleading, neutral). Trials were modeled as epochs defined by the onsets and duration (7 s) of the memory probe and were convolved with the canonical hemodynamic response function. Reaction times did not differ across warning groups or trial types (Fs < 1). Both GLMs also included a single regressor for arrows trials (baseline) and six nuisance regressors for each of the motion correction parameters. A high-pass filter of 1/128 Hz was applied to remove low-frequency noise. Contrasts of interest were computed for each participant at the single-subject level and subjected to a random effects (second level) GLM analysis to investigate the effect of warning (warning vs. no-warning). Participants who received the warning before or after the audio narrative were included in the warning group. Post hoc analyses investigated potential activation differences between the prewarning and postwarning groups within regions demonstrating main effects of warning. Given our a priori hypotheses of activation differences within specific anatomical regions of interest (e.g., visual cortex and auditory cortex), we report peaks whose statistic exceeded $P < 0.05$ corrected for family-wise error using small volume correction (SVC) and a minimum cluster size of eight contiguous voxels. The visual cortex region of interest was defined by combining the left and right cuneus and lingual gyrus regions of the AAL anatomical atlas (61). The auditory cortex region of interest was defined by combining BAs 41 and 42 bilaterally from the WFU_PickAtlas (62, 63). For the hippocampus, anatomical ROIs were drawn from the AAL atlas (61). For completeness, we also performed exploratory whole-brain analyses to

investigate potential activation differences outside of the visual cortex and auditory cortex ($P < 0.001$, cluster extent = 8 voxels).

Data Availability. The data from experiments 1 and 2 have been deposited in the OSF database (DOI: 10.17605/https://osf.io/WGN83/).

ACKNOWLEDGMENTS. This material is based upon work supported by the National Science Foundation under Grant 1728764 and NIH under Shared Instrumentation Grant S10OD021569. We thank Camille Carlisle, Jessica Jimbo, and the staff at the Martinos Imaging Center at the Massachusetts Institute of Technology for assistance with fMRI data collection.

1. D. L. Schacter, *The Seven Sins of Memory: How the Mind Forgets and Remembers*, (Houghton Mifflin, 2001).
2. E. F. Loftus, Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learn. Mem.* **12**, 361–366 (2005).
3. J. C. Chan, A. K. Thomas, J. B. Bulevich, Recalling a witnessed event increases eyewitness suggestibility: The reversed testing effect. *Psychol. Sci.* **20**, 66–73 (2009).
4. A. K. Thomas, J. B. Bulevich, J. C. K. Chan, Testing promotes eyewitness accuracy with a warning: Implications for retrieval enhanced suggestibility. *J. Mem. Lang.* **63**, 149–157 (2010).
5. J. C. Chan, J. A. Lapaglia, The dark side of testing memory: Repeated retrieval can enhance eyewitness suggestibility. *J. Exp. Psychol. Appl.* **17**, 418–432 (2011).
6. J. A. LaPaglia, J. C. Chan, Testing increases suggestibility for narrative-based misinformation but reduces suggestibility for question-based misinformation. *Behav. Sci. Law* **31**, 593–606 (2013).
7. M. M. Wilford, J. C. Chan, S. J. Tuhn, Retrieval enhances eyewitness suggestibility to misinformation in free and cued recall. *J. Exp. Psychol. Appl.* **20**, 81–93 (2014).
8. L. T. Gordon, A. K. Thomas, Testing potentiates new learning in the misinformation paradigm. *Mem. Cognit.* **42**, 186–197 (2014).
9. L. Gordon, J. B. Bulevich, A. K. Thomas, Looking for answers in all the wrong places: How testing facilitates learning of misinformation. *J. Mem. Lang.* **83**, 140–151 (2015).
10. E. J. Rindal, R. M. DeFranco, P. R. Rich, M. S. Zaragoza, Does reactivating a witnessed memory increase its susceptibility to impairment by subsequent misinformation? *J. Exp. Psychol. Learn. Mem. Cogn.* **42**, 1544–1558 (2016).
11. B. J. Butler, E. F. Loftus, Discrepancy detection in the retrieval-enhanced suggestibility paradigm. *Memory* **26**, 483–492 (2018).
12. H. L. Roediger 3rd, A. C. Butler, The critical role of retrieval practice in long-term retention. *Trends Cogn. Sci. (Regul. Ed.)* **15**, 20–27 (2011).
13. H. Blank, Memory states and memory tasks: An integrative framework for eyewitness memory and suggestibility. *Memory* **6**, 481–529 (1998).
14. K. L. Chambers, M. S. Zaragoza, Intended and unintended effects of explicit warnings on eyewitness suggestibility: Evidence from source identification tests. *Mem. Cognit.* **29**, 1120–1129 (2001).
15. D. K. Eakin, T. A. Schreiber, S. Sergent-Marshall, Misinformation effects in eyewitness memory: The presence and absence of memory impairment as a function of warning and misinformation accessibility. *J. Exp. Psychol. Learn. Mem. Cogn.* **29**, 813–825 (2003).
16. G. Echterhoff, W. Hirst, W. Hussy, How eyewitnesses resist misinformation: Social postwarnings and the monitoring of memory characteristics. *Mem. Cognit.* **33**, 770–782 (2005).
17. E. Greene, M. S. Flynn, E. F. Loftus, Inducing resistance to misleading information. *J. Verb. Learn. Verb. Be.* **21**, 207–219 (1982).
18. D. S. Lindsay, Misleading suggestions can impair eyewitnesses' ability to remember event details. *J. Exp. Psychol. Learn.* **16**, 1077–1083 (1990).
19. J. Underwood, K. Pezdek, Memory suggestibility as an example of the sleeper effect. *Psychon. Bull. Rev.* **5**, 449–452 (1998).
20. D. B. Wright, Misinformation and warnings in eyewitness testimony: A new testing procedure to differentiate explanations. *Memory* **1**, 153–166 (1993).
21. H. Blank, C. Launay, How to protect eyewitness memory against the misinformation effect: A meta-analysis of post-warning studies. *J. Appl. Res. Mem. Cogn.* **3**, 77–88 (2014).
22. M. S. Ayers, L. M. Reder, A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychon. Bull. Rev.* **5**, 1–21 (1998).
23. D. S. Lindsay, M. K. Johnson, The reversed eyewitness suggestibility effect. *Bull. Psychon. Soc.* **27**, 111–139 (1989).
24. D. S. Lindsay, M. K. Johnson, The eyewitness suggestibility effect and memory for source. *Mem. Cognit.* **17**, 349–358 (1989).
25. S. D. Slotnick, D. L. Schacter, A sensory signature that distinguishes true from false memories. *Nat. Neurosci.* **7**, 664–672 (2004).
26. D. L. Schacter *et al.*, Neuroanatomical correlates of veridical and illusory recognition memory: Evidence from positron emission tomography. *Neuron* **17**, 267–274 (1996).
27. K. A. Kurkela, N. A. Dennis, Event-related fMRI studies of false memory: An Activation Likelihood Estimation meta-analysis. *Neuropsychologia* **81**, 149–167 (2016).
28. H. Kim, R. Cabeza, Differential contributions of prefrontal, medial temporal, and sensory-perceptual regions to true and false memory formation. *Cereb. Cortex* **17**, 2143–2150 (2007).
29. J. M. Karanian, S. D. Slotnick, Confident false memories for spatial location are mediated by V1. *Cogn. Neurosci.* **9**, 139–150 (2018).
30. B. Zhu *et al.*, Multiple interactive memory representations underlie the induction of false memory. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 3466–3475 (2019).
31. D. L. Schacter, S. D. Slotnick, The cognitive neuroscience of memory distortion. *Neuron* **44**, 149–160 (2004).
32. C. E. Stark, Y. Okado, E. F. Loftus, Imaging the reconstruction of true and false memories using sensory reactivation and the misinformation paradigms. *Learn. Mem.* **17**, 485–488 (2010).
33. L. Nyberg, R. Habib, A. R. McIntosh, E. Tulving, Reactivation of encoding-related brain activity during memory retrieval. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11120–11124 (2000).
34. M. E. Wheeler, S. E. Petersen, R. L. Buckner, Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11125–11129 (2000).
35. C. J. Vaidya, M. Zhao, J. E. Desmond, J. D. Gabrieli, Evidence for cortical encoding specificity in episodic memory: Memory-induced re-activation of picture processing areas. *Neuropsychologia* **40**, 2136–2143 (2002).
36. I. Kahn, L. Davachi, A. D. Wagner, Functional-neuroanatomic correlates of recollection: Implications for models of recognition memory. *J. Neurosci.* **24**, 4172–4180 (2004).
37. A. M. Gordon, J. Rissman, R. Kiani, A. D. Wagner, Cortical reinstatement mediates the relationship between content-specific encoding activity and subsequent recollection decisions. *Cereb. Cortex* **24**, 3350–3364 (2014).
38. J. F. Danker, J. R. Anderson, The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding. *Psychol. Bull.* **136**, 87–102 (2010).
39. M. D. Rugg, J. D. Johnson, H. Park, M. R. Uncapher, Encoding-retrieval overlap in human episodic memory: A functional neuroimaging perspective. *Prog. Brain Res.* **169**, 339–352 (2008).
40. K. J. Mitchell, M. K. Johnson, Source monitoring 15 years later: What have we learned from fMRI about the neural mechanisms of source memory? *Psychol. Bull.* **135**, 638–677 (2009).
41. J. L. McClelland, B. L. McNaughton, R. C. O'Reilly, Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
42. K. A. Norman, R. C. O'Reilly, Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychol. Rev.* **110**, 611–646 (2003).
43. L. Davachi, J. F. Danker, "Cognitive neuroscience of episodic memory" in *Oxford Handbook of Cognitive Neuroscience*, K. N. Ochsner, S. M. Kosslyn, Eds. (Oxford University Press, 2013), pp. 375–388.
44. M. Moscovitch *et al.*, Functional neuroanatomy of remote episodic, semantic and spatial memory: A unified account based on multiple trace theory. *J. Anat.* **207**, 35–66 (2005).
45. M. S. Zaragoza, S. M. Lane, J. K. Ackil, K. L. Chambers, "Confusing real and suggested memories: Source monitoring and eyewitness suggestibility" in *Memory for Everyday and Emotional Events*, N. L. Stein, P. A. Ornstein, B. Tversky, C. Brainerd, Eds. (Erlbaum, 1997), pp. 401–425.
46. M. K. Doss, J. K. Picart, D. A. Gallo, The dark side of context: Context reinstatement can distort memory. *Psychol. Sci.* **29**, 914–925 (2018).
47. M. L. Mack, A. R. Preston, Decisions about the past are guided by reinstatement of specific memories in the hippocampus and perirhinal cortex. *Neuroimage* **127**, 144–157 (2016).
48. A. J. Horner, J. A. Bisby, D. Bush, W. J. Lin, N. Burgess, Evidence for holistic episodic recollection via hippocampal pattern completion. *Nat. Commun.* **6**, 7462 (2015).
49. R. Cabeza, S. M. Rao, A. D. Wagner, A. R. Mayer, D. L. Schacter, Can medial temporal lobe regions distinguish true from false? An event-related functional MRI study of veridical and illusory recognition memory. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4805–4810 (2001).
50. E. Tulving, *Elements of Episodic Memory*, (Oxford University Press, 1983).
51. M. K. Johnson, S. Hashtroudi, D. S. Lindsay, Source monitoring. *Psychol. Bull.* **114**, 3–28 (1993).
52. K. Velanova *et al.*, Functional-anatomic correlates of sustained and transient processing components engaged during controlled retrieval. *J. Neurosci.* **23**, 8460–8470 (2003).
53. C. C. Woodruff, M. R. Uncapher, M. D. Rugg, Neural correlates of differential retrieval orientation: Sustained and item-related components. *Neuropsychologia* **44**, 3000–3010 (2006).
54. D. A. Gallo, I. M. McDonough, J. Scimeca, Dissociating source memory decisions in the prefrontal cortex: fMRI of diagnostic and disqualifying monitoring. *J. Cogn. Neurosci.* **22**, 955–969 (2010).
55. B. A. Kuhl, J. Rissman, M. M. Chun, A. D. Wagner, Fidelity of neural reactivation reveals competition between memories. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5903–5908 (2011).
56. B. P. Staresina, M. Wimber, A neural chronometry of memory recall. *Trends Cogn. Sci. (Regul. Ed.)* **23**, 1071–1085 (2019).
57. Y. Okado, C. E. L. Stark, Neural activity during encoding predicts false memories created by misinformation. *Learn. Mem.* **12**, 3–11 (2005).
58. C. L. Baym, B. D. Gonsalves, Comparison of neural activity that leads to true memories, false memories, and forgetting: An fMRI study of the misinformation effect. *Cogn. Affect. Behav. Neurosci.* **10**, 339–348 (2010).
59. L. M. Gordon, A. K. Thomas, The forward effects of testing on eyewitness memory: The tension between suggestibility and learning. *J. Mem. Lang.* **95**, 190–199 (2017).
60. H. Bezard, R. Bérard, P. Cabaud, J. Dassin, *Rififi*, J. Dassin, director (motion picture, Pathé-Consortium Cinéma, 1955).
61. N. Tzourio-Mazoyer *et al.*, Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15**, 273–289 (2002).
62. J. A. Maldjian, P. J. Laurienti, R. A. Kraft, J. H. Burdette, An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage* **19**, 1233–1239 (2003).
63. J. A. Maldjian, P. J. Laurienti, J. H. Burdette, Precentral gyrus discrepancy in electronic versions of the Talairach atlas. *Neuroimage* **21**, 450–455 (2004).