Reducing the Perceived Reliability of an External Memory Store Reduces Susceptibility to its Manipulation

April E. Pereira*, Megan O. Kelly, Xinyi Lu, and Evan F. Risko

Department of Psychology, University of Waterloo, Waterloo, Ontario

For correspondence, please contact April Pereira at april.pereira@uwaterloo.ca

Abstract

Offloading memory to external stores (e.g., a saved file) allows us to evade the limitations of our internal memory. One cost of this strategy is that the external memory store used may be accessible to others, and thus, manipulated. Here we examine how reducing the perceived reliability of an external memory store may impact participants' susceptibility to its manipulation (i.e., endorsing manipulated information as authentic). Across two pre-registered experiments, participants were able to store to-be-remembered information in an external store and on two critical trials, we surreptitiously manipulated the information in that store. Results demonstrate that an explicit notification of a previous manipulation, a reduction in perceived reliability, can decrease susceptibility to manipulation of the external memory store.

Individuals are often presented with to-be-remembered information that is critical for accomplishing their future goals. Given that the ability to store and retrieve accurate information from memory is limited, this can cause problems with respect to what we wish to remember versus what we may only be capable of remembering (Cowan, 2010). As such, it is often easier to rely on external storage devices rather than to rely on one's internal/biological memory (Eskritt & Ma, 2014; Hutchins, 1995; Kelly & Risko, 2019; Lu et al, 2020; Sparrow et al, 2011). In this digital age, the amount of information that can be stored externally (e.g., in cyberspace) is virtually limitless and is often readily accessible. The use of external memory storage in place of internal memory storage can be thought of as a form of *cognitive offloading* (Risko & Gilbert, 2016). While offloading memory demands in this manner allow individuals the benefit of having an extended memory system with a vast capacity, there are costs associated with offloading memory to these external stores (Ferguson et al., 2015; Kelly & Risko, 2019; Lu et al, 2020; Sparrow et al., 2011). One such cost is that the external store, because of its location outside our mind/brain, is often readily subject to explicit forms of manipulation (Clark, 2010b; Sterelny, 2004). This is particularly problematic when our external memory stores are in places accessible via the Internet (e.g., personal information stored "in the cloud") and thus, in principle, accessible by others.

Endorsement

Upon retrieval from an external memory store (e.g., the cloud, a notebook), one must decide whether to endorse the information in the external store as that which had been stored there originally (i.e., *the endorsement problem*; Arango-Muñoz, 2013). One approach would involve simply endorsing any information in one's external memory store as legitimate. Indeed, anecdotally this is how many individuals describe their use of external memory stores (e.g.,

calendars, contact lists). This kind of stance (i.e., complete reliance) might reflect the fact that the information was poorly encoded into internal/biological memory initially because the individual anticipated that they would have access to the external memory store (Kelly & Risko, 2019; Lu et al., 2020) and/or a strong belief in the reliability of the external memory store. If one were to take this stance, then this would lead one to be susceptible to the manipulation of that store, a prediction confirmed in recent research (Risko et al., 2019).

In a series of experiments, Risko et al. (2019) presented participants with to-be-remembered words and instructed them to save the presented information to a computer file that they could access during a subsequent recall test. Doing so allowed the participants the opportunity to offload the memory demands to the external store. Unsurprisingly, this allowed near-perfect "recall" of the stored information at test. After repeating this procedure across multiple trials, on the final (critical) trial, researchers manipulated the information in the participant's external memory store by inserting a novel word into it. Individuals often failed to notice the manipulation, as most recalled the inserted information as if it had been initially presented. Importantly, endorsement was not absolute; that is, individuals did not appear to merely trust their external store uncritically. How, then, do individuals decide to endorse the information in their external memory stores? In the present investigation, we pursue this broad question through an examination of whether the perceived reliability of an external memory store modulates individuals' sensitivity to manipulation of its contents, and if so, how.

Reliability

One factor that may play an important role in whether an individual endorses information in an external memory store is how *reliable* they consider that external memory store to be (Storm & Stone, 2015; Weis & Weise, 2019). Indirect evidence consistent with this idea is

available in research on the human use of automated systems (Lewandowsky et al., 2000; Muir & Moray, 1996). Weis and Wiese (2019) examined the effect of actual and believed reliability on an individual's decision to offload task demands in a mental rotation task. In this task, participants had the option to rotate the stimuli either internally (mentally) or externally, with a rotation knob that rotated the object on a computer screen. The knob's actual reliability and an instruction altering participants' beliefs about the knob's reliability (believed reliability) were manipulated, and the frequency of cognitive offloading (i.e., the use of the knob) and perceived knob utility was measured. They found that participants adjusted their offloading based on the actual and believed reliability of the knob. When participants experienced a decrease in the knob's actual reliability or were led to believe that the knob's reliability was lower than it was, participants reduced their use of the external rotation option. Thus, the extent to which people offload their cognition is based on both actual, and possibly erroneous, believed reliability.

In the context of offloading memory demands, Storm and Stone (2015) provided evidence that the reliability of an external memory store modulated the benefit of offloading. Across three experiments, Storm and Stone (2015) demonstrated that when participants believed that a file containing a list of to-be-remembered words would be saved and accessible at test, there was a benefit to the recall of an intervening list that was not saved. The authors proposed that offloading the initial list reduced its proactive interference on the subsequent list. Importantly, for the present effort, this beneficial effect of offloading was not observed when the external memory store was considered unreliable. Unreliability in this case was manipulated by participants experiencing an ineffective saving process. Storm and Stone (2015) suggested that when the external memory store was perceived as unreliable, individuals were less likely to

offload their memory to that store (despite it being available), thus reducing the benefit to the subsequent list.

Present Investigation

In the present investigation, we extended previous work which examined individuals' susceptibility to endorsing information that has been surreptitiously inserted into their external memory store (Risko et al., 2019). In particular, we compared this susceptibility across a condition wherein individuals were "naïve" to the insertion, as in the work by Risko et al. (2019), and a condition in which individuals had been informed that we had previously manipulated their external memory store.

The reported experiments followed the same general procedure as that of Risko et al. (2019) but required that participants perform five trials instead of four. On each trial, participants were shown a list of to-be-remembered words, one a time, and had to type them into a computer file that they were instructed would be available during test (which was always the case). Participants then completed an arithmetic distractor task. During the recognition test in Experiment 1 or recall test in Experiment 2, participants were given access to their saved file to consult if desired. The procedure was the same for the first three trials, to develop a sense of trust in and familiarity with the external memory store. On the fourth trial, the researcher surreptitiously inserted a word into the participant's saved list in the time between the encoding task and retrieval (i.e., while participants completed the distractor task). Participants then completed their recognition/recall test on the fourth trial, and diverging from Risko et al. (2019) in which the task ends after this fourth trial, we explicitly notified participants that this manipulation of their external memory store had taken place. Critically, participants then completed a fifth trial, similar to the fourth trial. That is, we again inserted an item into the

participants' external memory store while they performed the distractor task between the encoding task and recognition/recall test. Thus, this fifth trial took place when participants knew that the reliability of their external store was compromised.

The critical question on both Trials 4 and 5 is whether individuals endorse the inserted item as having been presented during encoding and further, whether the likelihood of this endorsement changes following being apprised of the external memory store's vulnerability to manipulation. Based on previous research (Storm & Stone, 2015; Weis & Weise, 2019), we predicted that when participants were told that their external memory store could be manipulated, they would be less susceptible to a subsequent manipulation of it. We were also interested in the form that this putative decrease in susceptibility might take. For example, this reduced susceptibility might emerge as a decrease in endorsement for all items (e.g., a kind of general skepticism or bias against the external store) or a more specific increase in the likelihood that the inserted item is detected as such (i.e., increased sensitivity). In addition to endorsement, we also assessed participants' ability to pick out the inserted item on Trial 5, and self-reports of strategies (from internal memory reliance to external store reliance) employed.

Experiment 1

In Experiment 1 (preregistered at https://osf.io/xzw4t/), participants performed the tasks described above. The retrieval test was a modified recognition test wherein participants were presented with each study word (i.e., originally presented during encoding), and, on Trials 4 and 5, the inserted item as well—the only foil. After Trial 4, participants were told about the insertion of the item into their external store (i.e., their typed list) and asked if they noticed. After Trial 5, participants were first asked about the offloading strategy they employed. Participants rated on a scale from 1–5 the extent to which they relied on their typed list (i.e., the external

store) versus their internal memory. Participants were then asked if they noticed if a word was inserted on Trial 5 and finally, asked to select a word from their external memory store (i.e., list) that they thought was most likely to have been inserted. Data and materials for Experiment 1 are available at https://osf.io/xzw4t/.

Method

Participants

Data from 32 participants were collected based on an a priori power analysis with the desired power of .80 ($\alpha = .05$, two-tailed) to detect a 32% difference in participants' confidence in the inserted word from Trial 4 to Trial 5 (see *confidence during recognition test* below for details). Participants were undergraduate psychology students at the University of Waterloo participating for course credit. Data from two participants were replaced due to incomplete data.

Apparatus

Both the participant and researcher were seated in the same room with a divider separating their workstations. At the participant's workstation were two computers and two monitors, one of which displayed the instructions and task (display monitor), and the other used to create and save their typed lists (workspace monitor). These monitors were connected to the computers and monitors at the researcher's workstation to remotely control them and observe the participant's progression through the experiment (this was not made explicit to the participants, however). At the researcher's workstation were three computers with three corresponding monitors displaying each of the two monitors from the participant's workstation, and one was used to covertly access the participant's list and insert a word when required.

Stimuli

Five lists were created using the SenticNet 4 word corpus (Cambria, Poria, Bajpai & Schuller, 2016). The lists were counterbalanced across trial position. The word lists varied in lengths (i.e., 15, 17, 19, 21, 23) so that when participants progressed to the insertion trials (Trials 4 and 5), a one-word insertion would not be easily detectable. The inserted words were yoked to specific word lists, such that each list had the same designated word as the inserted word. Whenever a list was presented on Trials 4 or 5, the designated inserted word for that list was inserted in the middle position of the list. The word lists presented for the non-manipulated trials (1-3) did not include its yoked inserted word, and thus, list lengths were 14, 16, 18, 20, and 22. In the analysis, the inserted item was compared to a control item, which was the word presented directly before it, or in cases where that item was not encoded, the control item was directly preceding that item. Within and between each word list (including the inserted words), words were not meaningfully different in length or frequency, with median word lengths ranging from 6 to 7 and median list frequencies ranging from 258-764 (using frequency count from SUBTLEX-US; Brysbaert & New, 2009). At encoding, words were presented visually in the center of the screen in Arial font and each word was presented for 5 s with a 1 s interstimulus interval.

Post-Trial 4 notification question

After completing the recognition test for the first word insertion trial (Trial 4), participants responded to Question 1, which asked, "During the arithmetic task, we typed *"[inserted word]"* into your text file. Did you notice?"

Post-task questionnaire

Upon the completion of the second word insertion trial, Trial 5, participants were asked three questions specific to that final trial. Question 1 asked, "Please select the option that best describes your recognition strategy during the final (fifth) trial of this study." Participants had six options to choose from, including: (a) I relied exclusively on my typed list during the recognition test, (b) I relied mostly on my typed list during the recognition test, (c) I relied about equally on both my list and my internal memory during the recognition test, (d) I relied mostly on my internal memory during the recognition test, (e) I relied exclusively on my internal memory during the recognition test, and (f) None of the above. Question 2 asked, "On the last trial we may have added a word to your typed list that was not presented originally. Please respond yes or no as to whether you believe we inserted a word into your list on your final trial." Question 3 stated, "Please open up your last list. Please review this list and type out a word you think was inserted. Even if you don't think something was added, please guess." Participants were shown their final manipulated list to refer to for Question 3.

Procedure

Participants were seated at their workstation, approximately 50 cm in front of two adjacent monitors (display and workspace monitors). Each trial began with an encoding task, in which one word at a time was presented in white on a grey background. As each word was presented on the right display monitor, participants simultaneously typed each word into a text file on the left workspace monitor. If on the rare occasion participants missed writing a word, then they would not have the opportunity for it to be presented again. After the encoding task was complete, participants were asked to save their '.txt' file on the left workspace monitor. With their saved list now closed, participants completed a 30-s arithmetic distractor task on the

right display monitor, which asked them to answer ‘true’ or ‘false’ to simple arithmetic equations. After the distractor task, participants were instructed to open up their ‘.txt’ file on the workspace monitor and complete a recognition test on the display monitor, using the list as an aid if they chose to. During each recognition test, participants were asked to provide a confidence rating for each word one at a time, corresponding to if they believed each word in the recognition test was presented during the encoding task. For each word, participants provided a confidence rating of (1) definitely not presented during encoding, (2) probably not presented during encoding, (3) probably presented during encoding, or (4) definitely presented during encoding. There was no time limit. Three trials were completed in this manner. No items were inserted on Trials 1-3, thus, all the items presented in the recognition test were targets.

On the fourth trial, while participants were completing the distractor task, the researcher used one of the monitors at their workstations to covertly access the participant’s saved, closed list, and insert a word into the middle position of that list. This would take place undisclosed to participants, and their display monitor did not change while the researcher altered the contents of the file it held. When opening their saved list for the recognition test, participants unknowingly accessed this now manipulated list. Participants then performed the recognition test for Trial 4, on which the inserted item was presented as a foil. After the recognition test, participants answered the Post-Trial 4 notification question, the wording of which informed them that their external memory store was vulnerable to manipulation. Participants then completed the fifth and final trial, including the same manipulation as Trial 4. Participants subsequently answered Questions 1, 2, and 3 from the post-task questionnaire and to conclude, the researchers debriefed the participants about the true purpose of the study and reason for deception.

Results

Descriptive data from Experiment 1 is available in Table 1. All mixed-effects models reported throughout were conducted using the lme4 package (Bates et al., 2014). Interactions among the fixed factors were also included in the model when appropriate—as indicated in the preregistered analyses. We included intercepts for participant as a random effect unless otherwise specified. In the possible cases that models resulted in singular fits this factor was removed. When an interaction term is not significant, we report results with and without it in the model. It is important to note that in the reported analyses, responses to the inserted item was compared to a control item (actually presented item), which was the word presented directly before it, or in cases where that item was not encoded, the control item was directly preceding that item. Lastly, when a non-pre-registered analysis is conducted we refer to it in text as exploratory.

Endorsement

Endorsement was calculated by dichotomizing confidence responses. If participants responded "1" or "2" (definitely or probably not presented during encoding), this was considered a "no" response (i.e., not endorsed), whereas if they responded "3" or "4" (probably or definitely presented during encoding), this was considered a "yes" response (i.e., endorsed).

As seen in Figure 1, mean endorsement for the control items presented on Trials 4 and 5 were both 1.00 and for inserted items, were .94 and .72 respectively. We analyzed the effect of notifying participants of the unreliability of their external memory store by comparing responses on Trial 4 and Trial 5 on endorsement for each item type (inserted vs. control) using separate McNemar's Chi-squared tests with a continuity correction. There was a statistically significant difference in the endorsement of the inserted item across Trials 4 and 5, $\chi 2(1) = 4.00$, $p = .046$,

such that the inserted item was endorsed more on Trial 4 than Trial 5. Because participants endorsed the control item 100% of the time on both Trials 4 and 5, no statistical analysis is reported. We also analyzed the effect of item type (inserted vs. control) on endorsement separately for each trial (Trial 4 vs. Trial 5) using the same statistical test. There was no statistically significant difference in the endorsement of control and inserted items on Trial 4, $\chi2(1) = 0.50$, $p = .480$, but there was on Trial 5, such that inserted items were endorsed significantly less often than the control items, $\chi2(1) = 7.00$, $p = .008$. We preregistered a mixed-effects logistic regression to test the interaction between the effects of item type (inserted vs. control) and trial (Trial 4 vs. Trial 5) on endorsement with random intercepts for participant, however, this model failed to converge and as such no results are reported.

Confidence

We also analyzed confidence ratings as a continuous variable (using the entire 1-4 scale). An exploratory within-subjects Analysis of Variance (ANOVA) was conducted to test the effects of trial (Trial 4 vs. Trial 5) and item type (inserted vs. control) on confidence ratings. The results revealed a main effect of trial, $F(1, 31) = 9.59$, $p = .004$, $\eta_G^2 = .052$, and item type, $F(1, 31) = 12.40$, $p = .001$, $\eta_G^2 = .103$, and a significant interaction between trial and item type, $F(1, 31) = 5.07$, $p = .003$, $\eta_G^2 = .035$. Using pre-registered paired-samples t-tests, confidence ratings for inserted items were significantly higher on Trial 4 ($M = 3.78$, $SD = 0.75$) than on Trial 5 ($M = 3.16$, $SD = 1.27$), $t(31) = 2.69$, $p = .011$, $d = 0.48$. For control items, there was no significant difference in the confidence ratings between Trial 4 ($M = 4.00$, $SD = 0$) and Trial 5 ($M = 3.94$, $SD = .25$), $t(31) = 1.44$, $p = .161$, $d = 0.25$. When analyzing the effect of item type separately for Trials 4 and 5, there was no significant difference in the confidence ratings for control ($M = 4.00$, $SD = 0$) and inserted ($M = 3.78$, $SD = 0.75$) items on Trial 4; $t(31) = 1.65$, $p = .109$, $d = 0.29$. On

Trial 5, confidence was significantly lower for inserted items (M = 3.16, SD = 1.27) than for control items (M = 3.94, SD = 0.25), $t(31) = 3.37$, $p = .002$, $d = 0.59$. A mixed effects regression was conducted to test the interaction between effects of trial (Trial 4 vs. Trial 5) and item type (inserted vs. control) on confidence ratings with random intercepts for participant and revealed the interaction to be significant, $b = -0.56$, $SE = 0.25$, $t = -2.22$, $p = .029$.

Post-task questionnaire

For means of responses to the notification question and post-task Questions 1 (strategy; 0: completely external - 5: completely internal), 2 (think inserted; 0: no; 1: yes), and 3 (guess accuracy; 0: incorrect guess; 1: correct guess), see Table 1. In a series of regressions, we used individuals' reported strategy on Trial 5 as a predictor of whether they endorsed the inserted item on Trial 5 (logistic regression), confidence (1-4) for the inserted item on Trial 5 (linear regression), whether participants thought a word had been inserted on Trial 5 (logistic regression), and whether they correctly selected the inserted word on Trial 5 when asked (logistic regression). The overall mean self-reported strategy was rated 3.16 (SD = 1.02) on a scale from 1: (exclusive reliance on the external list) to 5 (exclusive reliance on internal memory). Strategy was not a significant predictor of endorsement of the inserted item, $b = -0.92$, $SE = 0.54$, $z = -1.68$, $p = .092$, but did predict confidence, $b = -0.46$, $SE = 0.21$, $t = -2.16$, $p = .039$, such that the more external the recognition strategy reported, the higher the confidence rating for the inserted item. Strategy did not predict whether participants thought a word was inserted on Trial 5, $b = -0.51$, $SE = 0.38$, $z = 1.34$, $p = .180$, or if they accurately guessed the inserted word, $b = 0.69$, $SE = 0.46$, $z = 1.52$, $p = .129$. These relations should be considered in light of the small sample size in Experiment 1.

Discussion

Consistent with previous research, participants often failed to notice a word inserted into their external memory stores. Indeed, on Trial 4, 94% of participants responded "yes" (i.e., a 3 or 4 on the confidence scale) that the inserted item had been presented and they were highly confident in their endorsement (3.78 on a 4-point scale). Critically, both endorsement and confidence decreased on Trial 5, after participants were told that we had previously manipulated their external memory store, though both endorsement rate (72%) and confidence rating (3.16) remained high. The notice in between Trials 4 and 5 appeared to have no substantive impact on the control item (i.e., the item that was actually presented). This is consistent with the notion that any effect of the notice primarily led to increased ability to discern the inserted item (i.e., foil) from actual target (control) items, rather than a general skepticism of the external store contents. The strategy report results were mixed, but there was some limited evidence that a self-reported reliance on the external memory store was related to a higher confidence rating of the inserted item.

Experiment 2

In Experiment 2 (pre-registered at https://osf.io/3v7j2), we sought a conceptual replication using a modified recall test at retrieval rather than a recognition test. One potential issue with using a recognition test is that participants can respond "yes" to the inserted item for reasons other than the presence of the inserted item in the external memory store. For example, individuals might have simply got into the habit of responding with a confidence rating of "4" (definitely presented during encoding) to all of the items, provided that a large majority of items (all except the inserted item) were presented during the encoding phase. A free recall test does not suffer from this limitation. For this recall test, participants were provided with a text box in

which they typed all of the words that had been presented on that trial. As in Experiment 1, during the recall test, participants could consult their saved lists (i.e., their external memory stores). Thus, the act of "recalling" the inserted word (i.e., typing it into the response box) would be unlikely, unless participants were actively endorsing the information in the external memory store. In Experiment 2, we continued to collect confidence ratings but given the change in test, these ratings took on a different meaning. That is, participants were asked to provide confidence ratings for all of the items they recalled. We again used a four-point scale, but here, each point corresponded to a percentage range of confidence the item had been presented starting at above 50% (1: 51-60%; 2: 61-75%; 3: 76-94%; 4:95-100%). In addition to switching to a recall test, we also included a no-notice condition wherein participants did *not* receive notice of the insertion after Trial 4. Lastly, participants completed the study online and we collected a much larger sample than in Experiment 1 to increase power. Data and materials for Experiment 2 are available at https://osf.io/xzw4t/.

Method

Participants

160 participants were included in the study and recruited online using Amazon's Mechanical Turk and completed the study within one hour for $9.00 USD. All participants were over the age of eighteen. One participant was replaced due to incomplete data and sixty participants were replaced based on exclusion criteria (see below for details). The number of usable participants collected was based on increasing power from an unpublished recall experiment (https://osf.io/wk62f) to better detect critical interactions between notice and item type. We roughly doubled our sample size for each condition present in Experiment 2.

Materials

The *Stimuli, Post-Trial 4 notification question,* and *Post-task questionnaire* used were the same as in Experiment 1.

Confidence measure

Beside each word that they typed ("recalled"), participants were asked to provide a confidence rating corresponding to how much they believed it was presented to them in the encoding task. For each word they recalled, participants provided a confidence rating of (1) possibly presented originally (i.e., between 51% and 60% chance it was presented), (2) moderately likely presented originally (i.e., between 61% and 75% chance it was presented), (3) very likely presented originally (i.e., between 76% and 94% chance it was presented), or (4) definitely presented originally (i.e., between 95% and 100% chance it was presented). There was a 5-minute time limit for the recall test before the program automatically proceeded to the next task.

Debriefing questionnaire

At the end of the experiment, participants were asked three questions that we used to exclude participants. Question 1 asked, "Did you take any notes or write anything down while completing the task?" Question 2 asked, "Were you doing anything else while completing this task? (e.g. Netflix)." For Questions 1 and 2, the options of *yes* or *no* were provided in multiple-choice format. Question 3 asked, "Is there any reason we should or should not use your data? (It's okay if you think you weren't able to give it your best, just let us know)." The options of "feel free to use my data" and "don't use my data" were provided in multiple-choice format.

Procedure

Each trial began with an encoding task, in which one word at a time was presented in blue on a white background. Participants were told to type each word as it appeared in exactly the way it was presented. As each word was presented in the middle of the screen, participants had 6 seconds to type the word in a text box below it to "save it" on a list (counterbalanced to populate on the left or right side of the screen, at the participant level). After 6 seconds, participants were presented with the next word and their previously typed word was added to the list. This list was presented on the right or left side of the screen under the title "saved list." No special characters, numbers, or capitalizations that the participant typed would be translated to their saved list. If on the rare occasion, participants missed writing a word, then they would not have the opportunity for it to be presented again and it would not be added to their saved list. After the encoding task was complete, participants had an opportunity to view their list for 10 s before it disappeared, and they moved on to the 30-s arithmetic distractor task, which had a time limit of 10 s per question. After the distractor task, participants completed the recall test, during which they were presented with their saved list on the same side of the screen as it had been presented during encoding. In the middle of the screen, there was a text box and participants were instructed to only type ("recall") words that they thought were presented during the encoding task along with a confidence rating, using their saved list as an aid if they chose to. Participants were advised that if they thought a word was not presented to them, they should not type ("recall") it in the text box. Three trials were completed in this way. On the fourth trial, when presented with their saved list at recall, it was presented with the inserted word halfway into their typed list, undisclosed to participants. Once participants completed the recall test for Trial 4, those in the notice condition were asked the Post-Trial 4 notification question, the

wording of which informed them that their external memory store was vulnerable to manipulation. Those in the no-notice condition moved on to Trial 5 without any notice. Afterwards, all participants completed the fifth and final trial, including the same manipulation as Trial 4. Participants subsequently answered Questions 1, 2, and 3 from the post-task questionnaire, completed the debriefing questionnaire, and were debriefed on the true purpose of the study and reason for deception.

Results

Descriptive data from Experiment 2 is available in Table 2. Average confidence ratings reported consist of only items that were recalled. The single control item to be compared to the single inserted item was decided in the same manner as Experiment 1. Sixty participants were excluded from all analyses based on failing any of the following criteria: (1) typing the word before the inserted word (used as the control) or the word before that, (2) typing at least 90% of the words they were supposed to on Trials 4 and 5 (the instruction was to write down 100% of the words), (3) accurately answering over 70% on the simple math problems during the arithmetic distractor task, (4) providing a confidence rating of 1-4, as instructed, to any recalled word (since our DVs are looking at the confidence of that recalled item, but we are not able to infer it). In a debriefing questionnaire at the end of the experiment, participants were excluded from all analyses if they answered yes to (1) doing something other than the task, (2) writing/screenshotting any words down during the encoding task, or (3) asking that we not use their data. All mixed-effects models reported throughout were conducted in the same manner as outlined in Experiment 1.

Recall

The mean proportions of items recalled as a function of condition (no-notice and notice) and item type (control vs. inserted) is presented in Figure 2. A mixed effects logistic regression with the predictors notice condition (no-notice vs. notice), trial (Trial 4 and 5) and item type (inserted vs. control) revealed a three-way interaction between condition, trial, and item type, b = -3.03, SE = 1.22, z = -2.49, p = .013. Two separate regressions revealed a significant interaction between trial and item type in the notice condition, b = -2.25, SE = 0.86, z = -2.62, p = .009, but not the no-notice condition, b = 0.55, SE = 0.86, z = 0.64, p = .520. When the interaction term in the latter model was removed, participants were significantly more likely to recall items on Trial 4 than Trial 5, b = -0.82, SE = 0.40, z = -2.06, p = .039, and significantly more likely to recall the control item than the inserted item, b = -3.50, SE = 0.60, z = -5.78, p < .001. We next performed separate regressions on the inserted and control items in the notice condition. In the notice condition, for inserted items, recall was significantly higher on Trial 4 compared with Trial 5, b = -2.17, SE = 0.55, z = -3.92, p < .001. No significant difference was revealed for control items, b = -1.17, SE = 1.16, z = -1.01, p = .310.

Confidence

A mixed effects regression including notice condition (no-notice vs. notice), trial (Trial 4 and 5) and item type (inserted vs. control), revealed a three-way interaction between condition, trial, and item type, b = -0.67, SE = 0.23, t = -2.93, p = .004. Two separate regressions for the notice and no-notice conditions revealed a significant two-way interaction between trial and item type in the notice condition, b = -0.44, SE = .18, t = -2.46, p = .015, but no such interaction in the no-notice condition, b = 0.21, SE = 0.14, t = 1.50, p = .135. When the interaction term in the model was removed for the no-notice condition, participants were not significantly more likely to

report higher confidences on a given trial, $b = 0.01$, $SE = 0.07$, $t = 0.11$, $p = .911$, but were significantly more likely to report lower confidence in the inserted than control item, $b = -0.25$, $SE = 0.07$, $t = -3.39$, $p < .001$. We next performed separate regressions for the effect of trial on inserted and control items each, in the notice condition. For inserted items, confidence was significantly lower on Trial 5 than Trial 4, $b = -0.56$, $SE = 0.18$, $t = -3.18$, $p = .003$. No significant difference was revealed for control items, $b = -0.07$, $SE = 0.08$, $t = -0.94$, $p = .350$.

Post-task questionnaire

For means across the notification question and post-task Questions 1 (strategy; 0: completely external - 5: completely internal), 2 (think inserted; 0: no; 1: yes), and 3 (guess accuracy; 0: incorrect guess; 1: correct guess), see Table 2.

To assess if the notification manipulation influenced self-reported strategy, an exploratory t-test was conducted. Recognition strategy did not differ across the no-notification condition ($M = 3.84$, $SD = 1.12$) and the notification condition ($M = 3.77$, $SD = .0.83$), $t(158) = 0.40$, $p = .700$. Given both normality and the homogeneity of variance assumptions were violated (p's $< .05$), a non-parametric test (Mann-Whitney-Wilcoxon Test) was also conducted and revealed similar results, $W = 3450$, $p = .400$. We also compared (again exploratory) responses across the no-notice and notice conditions for whether participants believed we had inserted an item on Trial 5, and their accuracy at guessing the inserted item on Trial 5, using separate Chi-squared tests with a continuity correction. There was a statistically significant difference in the belief of insertion across conditions, $\chi2(1) = 8.08$, $p = .004$, such that those in the notice condition more often reported believing that a word was inserted on Trial 5. There was no difference in the accuracy of guessing the inserted item on Trial 5 across conditions, $\chi2(1) = 3.60$, $p = .058$.

Next, in a series of regressions, we used individuals' reported strategy on Trial 5 as a predictor of whether they recalled the inserted item on Trial 5 (using logistic regression), confidence for the inserted item on Trial 5 (using linear regression), whether participants thought a word had been inserted on Trial 5 (using logistic regression), and whether they correctly selected the inserted word for Trial 5 when asked (using logistic regression).

Participants reporting a more external strategy were (i) more likely to recall the inserted item than those reporting a more internal strategy, $b = 0.75$, $SE = 0.19$, $z = 3.89$, $p < .001$, (ii) more likely to have a higher confidence rating for the inserted item, $b = 0.35$, $SE = 0.12$, $t = 2.88$, $p = .005$, (iii) were less likely to report that a word was inserted, $b = -0.57$, $SE = 0.20$, $z = -2.91$, $p = .003$, and (iv) more likely to have a lower accuracy in guessing the inserted word, $b = -0.43$, SE = 0.17, $z = -2.49$, $p = .013$. It is important to note that within Trial 5, only those that recalled the inserted item *and* had a subsequent confidence rating (46/80 participants in the no-notice condition, and 33/80 participants in the notice condition) were included in the linear regression for confidence rating. An exploratory logistic regression analysis, with both condition and a condition by strategy interaction as predictors, revealed no interaction for any of the four regressions listed above (recall of inserted item, confidence in inserted item, reporting a word was inserted, accurately guessing the inserted word).

Discussion

Experiment 2 extends the main result of Experiment 1 to a modified recall test. That is, receiving notice that an external memory store was not reliable, reduced individuals' susceptibility to the acceptance of manipulated information in their external store. Consistent with Experiment 1, participants often failed to notice a word inserted into their external memory stores. Indeed, across conditions on Trial 4, a majority of participants (65% in the no-notice

condition, 78% in the notice condition) recalled the inserted word and were confident that it had been previously presented (3.46/4 in the no-notice condition, 3.52/4 in notice condition). Critically, once given notice of the previous manipulation, recall and confidence decreased significantly on Trial 5, such that only 41% of participants recalled the inserted item and with less confidence (3.03/4).

Consistent with Experiment 1, the notice in between Trials 4 and 5 appeared to have no substantive impact on the control item (i.e., an actually presented item). The same results were found for confidence ratings. This suggests that any effect of the notice was primarily to increase individuals' abilities to discern actually presented target/control items from the inserted item (i.e., the foil). In the no-notice condition, there was a small general reduction in items recalled from Trial 4 to Trial 5 and participants were generally more likely to recall and have higher confidence in control than inserted items. This suggests that when given no notice of the manipulation, participants do not subsequently show such an increased ability to discriminate between control and inserted items, but instead show evidence of overall reduced trust (or general skepticism) in the store. It is unclear, at this point, what the cause of that effect might be, though it is important to note that while individuals in the no-notice condition were never told of the manipulation on Trial 4, a word was inserted, nevertheless. The strategy report results demonstrated that self-reported reliance on the external memory store was related to more recall of the inserted item, higher confidence in the inserted item, lower likelihood of thinking a word was inserted, and lower accuracy in guessing the inserted item.

General Discussion

Using external aids to offload cognitive demands has long been a memorial strategy allowing us to evade the limitations of our internal/biological memory (Clark, 2010a; Donald,

1991; Nestojko et al., 2013; Risko & Gilbert, 2016). However, there are costs to allocating memory demands to external locations, such as the potential susceptibility of these external memory stores to being accessed and manipulated. In the present investigation, we examined the role of the perceived reliability of one's external store in their susceptibility to manipulations of that store. As found by Risko et al. (2019), the majority of participants did not initially notice a manipulation of their external memory store. However, once given explicit notification of the manipulation, participants became better able to detect a subsequently inserted item without compromising their endorsement of the original contents. Still, even with an explicit notification of an insertion, many participants were unable to discriminate inserted from target words in their external memory stores. These results are consistent with expectations derived from prior work (Lewandowsky et al., 2000; Muir & Moray, 1996; Storm & Stone 2015, Weis & Wiese, 2019). Below we discuss these results and participants' self-reported reliance.

Understanding Endorsement in Distributed Memory Systems

Regardless of whether a memory is stored internally or externally, upon retrieval of that memory, participants must decide to endorse it or not (Arango-Muñoz, 2013). In the retrieval phase of the present experiments, participants had to decide if the items in their external store were the items that they, themselves, originally typed. The present results are consistent with the idea that whether an individual endorses the information in their external memory store depends on how *reliable* they consider it to be (Storm & Stone, 2015; Weis & Weise, 2019). As noted in the introduction, Weis and Wiese (2019) and Storm and Stone (2015) provided evidence consistent with the idea that perceived reliability decreases an individual's reliance on external resources. Thus, one potential explanation of the present results is that after being notified of their stores' potential lack of reliability, individuals invested more effort in encoding the items

(i.e., despite storing information, they did not forgo all effort to remember them internally). This could lead to greater differences (e.g., in the feeling of familiarity) between actually presented and inserted items, at retrieval. A different (but not mutually exclusive) explanation might be that the notification of manipulation shifts some individuals strategically, such that they move from a kind of "blind trust" to a more "memory-based" decision. The latter, of course, would allow them to discriminate between targets and inserted items effectively. Arguably, the observation that individuals in the notice and no-notice conditions did not differ in their self-reported strategies might be inconsistent with this latter view. That said, more research is needed to better understand how external store reliability modulates susceptibility to external store manipulation in distributed memory contexts.

Individual Differences in Self-Reported Reliance

After the second insertion in their external store (on Trial 5), participants provided a self-report rating of their reliance on their internal versus external memory store. When an individual had an external store available, participants did not *have to* use it to offload memory demands. If participants were to rely on their internal memory, then one could imagine that they would be better equipped to not endorse the inserted item. Overall, the relations between self-reported strategy and the various measures of one's susceptibility to the manipulation of their external store reported here seems consistent with this idea. That is, in Experiment 2, reported strategy was a significant predictor of endorsement of the inserted item, confidence in the inserted item, reporting a word was inserted, and accurately guessing the inserted word (these effects were in the same direction but not significant in Experiment 1 which had a smaller sample and a retrieval task of recognition instead of recall).

While the self-reported strategy data is interesting, it is important to note that individuals may not be able to accurately assess the extent to which they relied on their internal vs. external stores. In addition, given that the self-report questions followed the retrieval phase, participants' retrieval performance (e.g., Trial 5) could have influenced their answers. For example, participants may have successfully detected the inserted item and because of this, reported relying on their internal memory. An alternative approach to indexing individual differences in reliance on an external store by self-report could involve more indirect methods (e.g., pupil dilation during encoding).

Conclusion

Offloading memory to external stores is a critical strategy allowing us to evade the limitations of our internal memory. One cost of this approach is that it potentially exposes our "memories" to manipulation, provided that they reside out in the proverbial open. The presented research reinforces this idea, as most participants failed to notice a manipulation of their external store, and also demonstrates that an explicit notification of a previous manipulation (i.e., a notice of external store unreliability) can decrease this susceptibility. In a technologically advanced age, in which a large amount of to-be-remembered information can be externally stored, understanding the associated risks is crucial to utilizing our distributed memory systems efficiently.

Acknowledgements

This work was supported by a Discovery Grant (#04091) from the Natural Sciences and Engineering Research Council of Canada (NSERC), an Early Researcher Award from the Province of Ontario (#ER14-10-258), funding from the Canada Foundation for Innovation and Ontario Research Fund (#37872) and from the Canada Research Chairs (#950-232147) program to E.F.R.

Declaration of Interest Statement

The authors report no conflict of interest.

Open Practices Statement

The preregistration for this research is available at https://osf.io/xzw4t/, https://osf.io/3v7j2. The data and materials for this research is available at https://osf.io/xzw4t/.

References

Arango-Muñoz, S. (2013). Scaffolded memory and metacognitive feelings. *Review of Philosophy and Psychology*, *4*, 135–152. https://doi.org/10.1007/s13164-012-0124-1

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved frequency measure for American English. *Behavior Research Methods, 41*, 977–990.

Cambria, E., Poria, S., Bajpai, R., & Schuller, B. (2016). SenticNet 4: A semantic resource for sentiment analyses based on conceptual primitives. Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers, pp. 2666–2677.

Clark, A. (2010a). *Supersizing the mind*. Oxford, UK: Oxford University Press.

Clark, A. (2010b). Memento's revenge: The extended mind, extended. In R. Menary (Ed.). *The extended mind* (pp. 43–66). Cambridge, MA: MIT press.

Cowan, N. (2010) The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science.* 19, 51–579.

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Cambridge, MA: Harvard University Press.

Eskritt, M., & Ma, S. (2014). Intentional forgetting: Note-taking as a naturalistic example. *Memory and Cognition*, *42*(2), 237–246. https://doi.org/10.3758/s13421-013-0362-1

Ferguson, A. M., McLean, D., & Risko, E. F. (2015). Answers at your fingertips: Access to the Internet influences willingness to answer questions. *Consciousness and Cognition, 37*, 91–102.

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science, 19*(3), 265–288. https://doi.org/10.1207/s15516709cog1903_1

Kelly, M. O., & Risko, E. F. (2019). The isolation effect when offloading memory. *Journal of Applied Research in Memory and Cognition, 8*(4). https://doi.org/10.1016/j.jarmac.2019.10.001

Lewandowsky, S., Mundy, M., & Tan, G. P. A. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied*, *6*(2), 104–123. https://doi.org/10.1037//1076-898x.6.2.104

Lu, X., Kelly, M. O., & Risko, E. F. (2020). Offloading information to an external store increases false recall. *Cognition*, 104428. Advance online publication. https://doi.org/10.1016/j.cognition.2020.104428

Muir, B., M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*, 429-460.

Nestojko, J. F., Finley, J. R., & Roediger, H. L. (2013). Extending cognition to external agents. *Psychological Inquiry, 24*, 321–325.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Risko, E. F., Kelly, M. O., Patel, P., & Gaspar, C. (2019). Offloading memory leaves us vulnerable to memory manipulation. *Cognition*, *191*. https://doi.org/10.1016/j.cognition.2019.04.023

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, *333*(6043), 776–778. https://doi.org/10.1126/science.1207745

Sterelny, K. (2004). Externalism, epistemic artefacts, and the extended mind. In R. Schantz (Ed.). *The externalist challenge* (pp. 239–254). Berlin, DE: Walter de Gruyter.

Storm, B. C., & Stone, S. M. (2015). Saving-enhanced memory: The benefits of saving on the learning and remembering of new information. *Psychological Science*, *26*(2), 182–188. https://doi.org/10.1177/0956797614559285

Weis, P. P., & Wiese, E. (2019). Using tools to help us think: Actual but also believed reliability modulates cognitive offloading. *Human Factors*, *61*(2), 243–254. https://doi.org/10.1177/0018720818797553

Table 1

Means (*SD*s) of all dependent variables (Confidence, Endorsement, Post-Trial 4 notification question, Post-task questions 1-3; Strategy, Think inserted, Guess accuracy) are reported across the various conditions. For Trials 1-3, the control confidence and endorsement are mean values for all encoded items. For Trials 4-5, the control confidence and endorsement are means of the one control item.

	Trial 1	**Trial 2**	**Trial 3**	**Trial 4 (pre-notification)**	**Trial 5 (post-notification)**
Control confidence	3.97 (0.26)	3.95 (0.38)	3.99 (0.16)	4.00 (0.00)	3.94 (0.25)
Inserted confidence	-	-	-	3.78 (0.75)	3.16 (1.27)
Control endorsement	.99 (0.08)	.98 (0.13)	.99 (0.06)	1.00 (0.00)	1.00 (0.00)
Inserted endorsement	-	-	-	.94 (0.25)	.72 (0.46)
Notification question	-	-	-	.19	-
Strategy	-	-	-	-	3.16 (1.02)
Think inserted	-	-	-	-	.66
Guess accuracy	-	-	-	-	.34

Table 2

Means (SDs) of all dependent variables (Confidence, Recall, Post-Trial 4 notification, Post-task question answers 1-3; Strategy, Think inserted, Guess accuracy) are reported across the various conditions. For Trials 1-3, the control confidence and recall are mean values for all encoded items. For Trials 4-5, the control confidence and recall are means of the one control item.

Condition		Trial 1	Trial 2	Trial 3	Trial 4 (pre-notification)	Trial 5 (post-notification)
Notice	Control confidence	3.28 (1.12)	3.95 (0.23)	3.85 (0.39)	3.85 (0.46)	3.78 (0.56)
	Inserted confidence	-	-	-	3.52 (0.97)	3.03 (1.24)
	Control recall	0.95 (0.22)	0.99 (0.11)	0.99 (0.07)	0.93 (0.27)	0.90 (0.30)
	Inserted recall	-	-	-	0.78 (0.42)	0.41 (0.50)
	Notification question	-	-	-	0.40	-
	Strategy	-	-	-	-	3.77
	Think inserted	-	-	-	-	0.78
	Guess accuracy	-	-	-	-	0.58
No-notice	Control confidence	3.75 (0.79)	3.80 (0.66)	3.72 (0.82)	3.81 (0.63)	3.77 (0.71)
	Inserted confidence	-	-	-	3.46 (1.06)	3.65 (0.85)
	Control recall	0.78 (0.41)	0.84 (0.37)	0.97 (0.18)	0.94 (0.24)	0.88 (0.33)
	Inserted recall	-	-	-	0.65 (0.48)	0.58 (0.50)
	Notification question	-	-	-	-	-
	Strategy	-	-	-	-	3.84
	Think inserted	-	-	-	-	0.55
	Guess accuracy	-	-	-	-	0.41