*******PAPER IN PRESS AT THE JOURNAL OF MEMORY AND LANGUAGE*******

Polarity and Attitude Effects in the Continued-Influence Paradigm

Andrew Gordon[1,2], Ullrich K. H. Ecker[3], Stephan Lewandowsky[2,3]

Authors' Affiliations:

[1]MIND Institute, University of California, Davis, Sacramento, USA.

[2]School of Psychological Science, University of Bristol, Bristol, UK.

[3]School of Psychological Science, University of Western Australia, Perth, Australia.

Declarations of interest: None

Corresponding author: Andrew Gordon
Department of Psychiatry and Behavioural Sciences
Solomon Lab, UC Davis M.I.N.D Institute, +1916-703-0347

Word count abstract: 244

Word count manuscript (excluding references, tables, and figure captions): 9,260

Abstract: Misinformation – information that is false or inaccurate – can continue to influence people's memory and reasoning even after it has been corrected. Researchers have termed this the continued influence effect (CIE). However, to date, research has focused exclusively on examining the CIE in a single polarity, namely the ongoing effect of initially affirmed material that is later negated. No research has yet examined how reliance on outdated information may be affected if this polarity is reversed, that is, if initially-negated information is reinstated. It also remains unclear how participants' pre-existing beliefs may impact the acceptance of a correction, with prior evidence showing conflicting results. To investigate these questions, across two experiments we presented participants scoring high versus low on measures of relevant attitudes with fictional news reports that contained a piece of critical attitude-relevant information. This information was either true throughout, false throughout, initially-affirmed then retracted, or initially-negated then reinstated. Participants' reliance on the critical information was subsequently measured with the use of inferential-reasoning items. Reinstatement of initially-negated information was insufficient to bring reliance on that information to a baseline level – that is, reliance on information presented as true throughout was greater than reliance on negated and then reinstated information. This result was symmetrical with the conventional CIE observed with a reversed polarity. The effect of participants' pre-existing attitudes on continued reliance was equivocal. The results therefore suggest that the CIE is not contingent on polarity, raising questions about the cognitive mechanisms underlying the effect.

Keywords: Misinformation; Worldview; Continued influence; Negation; Belief updating

Polarity and Attitude Effects in the Continued-Influence Paradigm

Decades of research have provided compelling evidence that misinformation – information that is initially held as accurate but subsequently corrected – can continue to influence people's judgements and decision making despite an acknowledged correction (e.g., Chan, Jones, Jamieson, & Albarracín, 2017; Ecker, Lewandowsky, Swire, & Chang, 2011; Johnson & Seifert, 1994; Lewandowsky, Ecker, & Cook, 2017; Rich & Zaragoza, 2016; Wilkes & Leatherbarrow, 1988; for reviews see Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Schwarz, Newman, & Leach, 2016). This phenomenon can be observed even when people demonstrably remember and understand a correction (e.g., Johnson & Seifert, 1994; Marsh, Meade, & Roediger, 2003), or are explicitly warned at the outset that they may be exposed to misinformation (Ecker, Lewandowsky, & Tang, 2010). Researchers have termed this pervasive finding the continued influence effect (CIE).

However, previous CIE research has focused exclusively on the impact of corrective information in a single direction: In a typical study, participants encounter an initial *affirmation* of a critical piece of information, followed by its subsequent *retraction*. No research has yet investigated how continued influence may be affected if the narrative polarity is reversed; that is, if participants receive an initial *negation* of a critical piece of information followed by a subsequent *reinstatement* of that same information. Indeed, in today's politically-charged climate of 'fake news' and 'post-truth', members of the public may be more likely than ever before to encounter not only misinformation, but also misinforming *negations* of correct information (for instance, people may be exposed to the false claim "there is no global warming"; Lewandowsky et al., 2015; van der Linden, Leiserowitz, Rosenthal, & Maibach, 2017). It is unclear how such an initial negation would affect people's subsequent level of belief for the initially-negated but subsequently

reinstated information – for example, if people encountered factual climate-change information in the media after having been exposed to climate-change denial. We propose two distinct possibilities: (a) a reinstatement of previously negated material may result in a 'reverse CIE', whereby belief is increased, but not to the level expected if information had been presented as true throughout; or (b) a reinstatement of initially-negated information may result in memory being updated to the level of the true-throughout presentation or even beyond (which would constitute an 'ironic rebound effect'). The two studies presented below were designed to adjudicate between these two possibilities (i.e., reverse-CIE or ironic rebound) by examining participants' reliance on information that was initially-negated and subsequently reinstated.

Considering the first possibility, given that the CIE reflects a primacy effect (i.e., *initially* presented misinformation appears to be difficult to overcome), just as a retraction is often unable to drive misinformation reliance down to a baseline, a reinstatement may be unable to fully neutralize an initial negation. This may occur either because the negation is encoded pragmatically as an assertion (i.e., simply as a single meaningful fact that asserts a certain state of the world), or because the constituent parts of the negated message—the 'core' feature ('X') and its negation marker ('not') — are incorporated into a single meaningful unit affirming the opposite, such that a statement 'not-X' may be encoded as 'therefore-Y' (see Mayo, Schul, & Burnstein, 2004). Importantly, in both cases the negation will be encoded as a single unit that is cognitively indistinct from the encoding of a piece of (initially-affirmed) misinformation in the typical CIE paradigm, plausibly implying that there should be no difference in the impact of a correction as a function of its polarity. That is, reinstatement would be followed by a reverse CIE that is the mirror image of the conventional CIE observed with a retraction of initially-affirmed information.

This concept is consistent with both previous CIE research as well as theoretical models of knowledge revision. One such model in particular, the Knowledge Revision Components framework (KReC - Kendeou & O'Brien, 2014, for a detailed review see Gerrig & O'Brien, 2005; Kendeou, Butterfuss, Kim & van Boekel, 2019; Kendeou & O'Brien, 2014; Kendeou, Walsh, Smith & O'Brien, 2014), assumes that new information encoded into memory (e.g., the correction) has the potential to activate related pieces of information in memory through resonance-based retrieval processes (O'Brien et al., 1998, 2010; Rapp & Kendeou, 2009). In this framework, for knowledge revision to succeed, a correction needs to resonate with the initial, corresponding misinformation such that co-activation occurs, triggering conflict detection and subsequent information integration. If knowledge revision is successful, subsequent cues (such as inference questions) will activate an updated, integrated representation and no CIE occurs. If knowledge revision is unsuccessful, however, subsequent cues have the potential to selectively activate the initial misinformation, leading to a CIE. To the extent that negations are encoded as single meaningful units, the KReC framework makes no specific predictions about polarity, and thus may provide a plausible account of both CIE and reverse-CIE effects.

Turning to the second possibility, in contrast to a retraction of initially-*affirmed* information, a reinstatement of initially-*negated* information may induce updating of memory to a level equivalent to, or even greater than, when that information is only-affirmed. Such an outcome might result from an initial negation enhancing the familiarity of the critical information (especially if the negated statement is initially not well represented in memory; Mckinstry, Dale, & Spivey, 2008; Reichle, Carpenter, & Just, 2000), which could potentially leave people with only a sense of the primary content and no strong memory for its negation. For example, Richard Nixon's famous line "I am not a crook" may have left

people with a sense of Nixon being associated with the term “crook” but a less clear sense of it being negated. Subsequent affirmative information may thus be particularly impactful as it is “sowed on fertile ground.”

This possibility relies on the ‘schema-plus-tag’ model of negation comprehension, which suggests that a negation is encoded as a core supposition (‘X’) that is then associated with a negation-tag (‘not’; see Gilbert et al., 1990; Johnson & Seifert, 1998; Mayo et al., 2004). Unlike under the previous possibility, the two constituent parts are not integrated into a single assertive representation. If negations are encoded in this manner, then the provision of a later reinstatement may combine with the initial core supposition, thereby resulting in its (ironic) strengthening. For example, suppose a person is provided with the information ‘not-X’, which creates two dissociated representations, the core (‘X’) and its negation-tag (‘not’; see Ecker et al., 2011). Upon later reinstatement, this critical piece of information ‘X’ is once again provided, this time as an affirmation. The initial negation therefore indirectly provides the kernel piece of information for its subsequent reinstatement, and the constituent ‘X’ in the initial ‘not-X’ may combine with the later ‘X’ to form a representation that is at least as strong as the one that would result from being exposed to ‘X’ on its own. Given that familiarity and repetition have been shown to be driving factors in people’s consequent level of belief in information (e.g., Dechêne, Stahl, Hansen, & Wänke, 2010; DiFonzo, Beckstead, Stupak, & Walders, 2016; Ecker et al., 2011; Putnam, Wahlheim, & Jacoby, 2014; Stadler, Scharrer, Brummernhenrich, & Bromme, 2013; Swire, Ecker, & Lewandowsky, 2017), a repetition of initially-negated information in a later reinstatement may thus wipe out the initial negation completely (also see Ecker, Hogan, & Lewandowsky, 2017; Kendeou et al., 2014).

In addition to this, and in contrast to an initial-*affirmation*, it is conceivable that the provision of an initial-*negation* creates an implicit assumption that there must be a degree of tacit belief, or at least plausibility, in the negated information. That is, people may think that a negation would not be provided unless it negated an unarticulated assumption. In the context of the example provided earlier, Richard Nixon's "I am not a crook" speech may have made some people feel sceptical and therefore particularly ready to accept subsequent events that revealed the contrary to be true. On this account, a reinstatement would be highly effective, potentially resulting in an ironic rebound beyond when the information is true-throughout. Such an effect of scepticism requires that the initial negation is not integrated into a single meaningful unit, but instead as two dissociated representations, enabling a subsequent reinstatement to confirm the earlier scepticism and uniquely amplify the core ('X') message but not its negation-tag ('not'). The notion of scepticism facilitating efficient updating (and potentially an ironic rebound) is consistent with research demonstrating that scepticism facilitates recognition of correct information and rejection of retracted falsehoods (Lewandowsky, Stritzke, Oberauer, & Morales, 2005).

However, any discussion of the effects of misinforming information in the real world must also take into account the role of the recipient's worldview, that is, their basic beliefs about how the world should operate (see Lewandowsky et al., 2012; Nyhan & Reifler, 2010). While typical CIE effects are readily demonstrable with fictitious, neutral materials, worldview factors (i.e., political partisanship, religion, etc.) have been shown to exert an influence on what information a person will accept (Ecker, Lewandowsky, Fenton, & Martin, 2014; Kahne & Bowyer, 2017; Redlawsk, 2002; Schaffner & Roche, 2016; Sharot, Korn, & Dolan, 2011). While some research has found evidence to suggest that the ongoing influence of initially-acquired information is heightened if the subsequent retraction is

worldview-incongruent (see Ecker & Ang, 2019; Lewandowsky et al., 2005, 2012), other studies have shown that a retraction can be equally effective regardless of its congruence with a person's worldview, suggesting that differences in post-retraction belief may simply reflect pre-retraction differences (Ecker et al., 2014; also see Wood & Porter, 2019). The role of worldview in the processing of retractions or corrections therefore remains somewhat unresolved. Moreover, the effects of worldview on the processing of initial *negations* and subsequent *reinstatements* are completely unknown. One possibility is that the potential scepticism effects discussed previously emerge as a function of worldview. That is, scepticism effects may be particularly apparent in situations where a person has a prior motivation to be sceptical; for instance, a staunch Democrat may be more sceptical about (initial-negation) claims made by Republican politicians (e.g. Richard Nixon) than a Republican, and a subsequent reinstatement may thus be *more* effective.

Our studies therefore also examined the role of participants' worldview. Across two experiments, we presented participants scoring higher versus lower on a prejudice measure with fictional news reports concerning a crime. For both experiments, a piece of critical worldview-relevant information was either provided, negated, reinstated, or retracted, and participants' subsequent reliance on that critical information was examined.

Experiment 1

In Experiment 1, participants were presented with a fictitious news report about a violent robbery that took place in a convenience store in south-western Australia. In total, four versions of the report were created. In two versions, the robber was initially described as being of Aboriginal descent, and in one of these two versions this information was later retracted (*affirmed-then-retracted* condition), whereas in the other it was not (*only-affirmed* condition). Two other versions of the report both initially described the robber as *not* being

of Aboriginal descent. In one of these two versions, this initial denial was retracted (*negated-then-reinstated* condition), whereas in the other version it was not (*only-negated* condition). Participants were also split by their level of racial prejudice, measured by the Attitudes Toward Indigenous Australians scale (ATIA; Pedersen, Beven, Walker, & Griffiths, 2004). Experiment 1 therefore used a 2 (racial prejudice: high vs. low) × 4 (condition: only-negated, negated-then-reinstated, only-affirmed, affirmed-then-retracted) between-subjects design.

Method

Participants. Based upon an a-priori power analysis, Ecker and colleagues (2014) tested 144 participants (with a minimum sample size of 90 sufficient to detect a medium-sized effect, $\eta_p^2 = .1$; $\alpha = .05$; $1 - \beta = .80$). Given the similarity between this previous work and the current design we aimed to recruit a minimum of 192 participants (based on the inclusion of an extra condition beyond that used by Ecker and colleagues). Our final sample consisted of 207 participants, all undergraduate students from the University of Western Australia (140 females, 67 males; age range 17-47, *M* = 20.2 years, *SD* = 4.71). Three further participants were recruited but were excluded based on non-completion of the answer booklet. Participants were randomly allocated to the different script conditions (with the constraint of roughly equal cell sizes). A subset of *n* = 125 were selected from the upper and lower thirds of a population of students pre-screened with the ATIA (total *N* = 1,059). The remaining *n* = 82 participants were not pre-screened due to pragmatic reasons (i.e., lack of access to a pre-screened population), and were recruited as a random opportunity sample through advertisements.

The ATIA scale (Pedersen et al., 2004) consists of 18 items, answered on 7-point Likert scales. Scores on the six reverse-coded items were transformed so that all items

shared the same polarity (with higher scores indicating higher prejudice). Participants in the opportunity sample completed the ATIA scale during the experimental session, whereas pre-screened participants completed the ATIA only at pre-screening. Scores on the ATIA scale ranged from 18 to 106 (the minimum possible score is 18 and the maximum is 126). Participants were split into high and low racial-prejudice groups with a median split on the entire sample (*Md* = 40). The mean racial-prejudice score for the low prejudice group (*n* = 104) was *M* = 26.68 (*SD* = 4.78), and for the high prejudice group (*n* = 103) it was *M* = 63.17 (*SD* = 13.06).

Stimuli (all stimuli are available at https://osf.io/jfsqm/). Stimuli were loosely based upon the news reports used in the study by Ecker and colleagues (2014). The fictional news reports consisted of either 12 (only-negated and only-affirmed conditions) or 13 (negated-then-reinstated and affirmed-then-retracted conditions) messages concerning a robbery at a convenience store in the town of Margaret River, Western Australia. Across conditions, news reports differed in sentence 4, where the critical information concerning the race of the suspects was introduced (in all but the only-affirmed condition). The reports also differed in sentence 11 (sentence 10 in the only-affirmed condition), where depending on condition the information concerning the race of the suspects was initially-stated, retracted, or reinstated (or not mentioned in the only-negated condition; for detail see Table 1).

Participants' understanding of the report and their level of reliance on the critical race-related information was tested using an open-ended questionnaire booklet that was administered once a distractor anagram task had been completed. The booklet contained an event-summary question, eight fact-recall questions, and eleven inference questions. In the

convenience sample, the ATIA scale was given at the end to avoid responses to the ATIA biasing responses to the inference questions.

The event-summary question asked participants to supply a short summary of the event. The fact-recall questions targeted arbitrary event details unrelated to the critical race-related information (e.g., “In which town did the robbery take place”); these questions were included to allow for the potential exclusion of participants who had not adequately encoded the news report. The inference questions were designed to evoke references to the critical race-related information, while allowing participants to provide responses that did not relate to the race of the suspects. For example, the question “Why might the other two patrons of the store have suggested that the attackers may have been intoxicated?” could be answered by using race-related information (e.g., by referring to the stereotype that alcoholism is common in Aboriginal people), or by not referring to race-related information (e.g., by assuming that the perpetrators showed signs of intoxication such as slurred speech). The final question asking directly about the race of the suspects (“Were any Aboriginals involved in the robbery?”) was answered on a scale from 0 (strongly disagree) to 7 (strongly agree).

Table 1. Illustration of presentation order for critical race-related information across all conditions.

	Only-negated	**Negated-then-reinstated**	**Only-affirmed**	**Affirmed-then-retracted**
Message 4	"[…] the suspects are local, male residents and none of them are of Aboriginal descent"	"[…] the suspects are local, male residents and none of them are of Aboriginal descent"	*NO CRITICAL INFORMATION*	"A spokesperson stated … they now have three Aboriginal male suspects in custody"
Message 11	*NO CRITICAL INFORMATION*	"A spokesperson stated […] they now have three Aboriginal male suspects in custody"	"A spokesperson stated […] they now have three Aboriginal male suspects in custody"*****	"[…] the suspects are local, male residents and none of them are of Aboriginal descent"

****** *This sentence appeared in message 10*

Procedure. Participants were first given an ethics-approved information sheet before providing written consent (in the pre-screened sample, participants received two information sheets and consented separately to the ATIA screening and the main experiment). Participants completed the encoding and distractor tasks on a computer using Microsoft PowerPoint. They were told that they were to read a report and that their memory for that report would be subsequently tested. Once participants were ready to begin, they pressed the space bar and the report was presented one message at a time. The presentation progressed to the subsequent message at a pre-determined pace of 0.25 seconds per word. Following the presentation of the report, participants were given an unrelated anagram task that lasted for four minutes. Participants were then asked to open and complete the questionnaire booklet placed face down in-front of them. Once participants had completed the response booklet, they were debriefed. The entire experiment took approximately thirty minutes.

Results (data available at https://osf.io/7ts9k/)

All statistical analyses were conducted using both frequentist and Bayesian techniques, using the IBM SPSS software and the BayesFactor package for R (Morey, 2015), respectively. All analyses are thus reported with an associated *p*-value and Bayes Factor (BF). The BF reflects the likelihood of the observed data under the alternative hypothesis (BF_{10}) relative to the likelihood of the data under the presumption of no effect (the null hypothesis). The value of the BF therefore represents the strength of the evidence for one model over another on a continuous scale. Although different categorisations exist (see Held & Ott, 2018), generally a BF of 1 represents equivocal evidence for either model, a BF between 1-3 is considered as weak evidence for the alternative, BFs between 3-10 as intermediate, BFs > 10 as strong evidence, and BFs > 100 as conclusive evidence for the alternative model over the null (Held & Ott, 2016; Kass & Raftery, 1995). Correspondingly, a $BF_{10} < 1$ is considered evidence for the null hypothesis with a strength equal to its reciprocal.

Coding procedure. All open-ended questions were scored by a trained scorer who was blind to condition. To assess inter-rater reliability, a second trained scorer scored a subset of the questionnaires (n = 20; 5 per condition). Inter-rater reliability was found to be high for both fact-recall questions (r = .978) and inference items (r = .975).

Fact-recall items were scored as incorrect (0) or correct (1). A score of 0.5 was possible for answers deemed partially correct. The maximum fact-recall score was 8. Inference items were scored 0 for no implication of an Aboriginal suspect, and 1 for an uncontroverted reference to an Aboriginal suspect. An example of a response scored 1 was "[…] community might start to feel threatened by Aboriginal males or females in the future" in response to the question "How is this event likely to impact social harmony in the area?"; by contrast, "[…] the arrest didn't seem logical, and this might be seen as racial profiling and

cause uproar" in response to the same question was scored 0. Ambiguous references were scored 0.5; for example, the response "Because the robbers spoke in a different dialect" in response to the question "Why might the injured patron have struggled to understand what the robbers were asking him to do?" may have referred to an Aboriginal dialect but did not state so explicitly, and was therefore given a partial score. The event-summary question was treated as an inference question and was scored in the same manner. The final inference question directly asked participants about the race of the suspects. Participants were required to respond to this item on a scale (scored from 0-7). In order to be consistent with the other inference questions, scores on this item were recoded to a continuous 0-1 scale and integrated with the scores from the ten inference questions and event-summary question to provide an overall average inference score ranging from 0 to 1. This score reflected participants' tendency to rely on the critical information in their reasoning, and was our main dependent variable.

Accuracy of recall. The mean level of fact-recall across conditions (expressed as proportion correct) was .69 (only-negated = .69; negated-then-reinstated = .67; only-affirmed = .72; affirmed-then-retracted = .67). A 4 × 2 analysis of variance (ANOVA) on participants' mean fact recall accuracy rates with the factors condition (negated-then-reinstated, only-negated, only-affirmed, affirmed-then-retracted) and prejudice group (low, high) found no significant effects of condition [$F(3,199) = 0.701$, $MSE = 2.33$, $p = .553$, $\eta^2_p = .010$], or prejudice group [$F(1,199) = 2.39$, $MSE = 2.331$, $p = .123$, $\eta^2_p = .012$], as well as no interaction between the two [$F(3,199) = 2.13$, $MSE = 2.331$, $p = .098$, $\eta^2_p = .031$]. These results were supported by varying levels of evidence in favour of the null hypothesis in the Bayesian analysis: $BF_{10} = 0.06$ (condition); $BF_{10} = 0.49$ (prejudice group); $BF_{10} = 0.55$ (interaction). Following precedent set by Ecker and colleagues (2014), each analysis detailed

in the following sections was repeated with the exclusion of participants with fact-recall scores < 2. The exclusion of low-recall participants did not substantially alter the pattern of results for any of the three analyses conducted and is therefore not presented.

Inferential reasoning. A two way between-subjects ANOVA (see Figure 1) on inference scores with the factors condition and ATIA group yielded a significant main effect of condition [$F(3,199) = 58.77$, $MSE = 0.008$, $p < .001$, $\eta^2_p = .470$, $BF_{10} = 3.72 \times 10^{23}$]. There was no significant main effect of ATIA group [$F(1,199) = 2.58$, $MSE = 0.008$, $p = .110$, $\eta^2_p = .013$, $BF_{10} = 0.25$], or interaction between the two factors [$F(3,199) = 2.04$, $MSE = 0.008$, $p = .110$, $\eta^2_p = .030$, $BF_{10} = 0.15$], suggesting that participants who scored high versus low on the ATIA measure made an equivalent number of references to an Aboriginal robber both across and within conditions.[1]

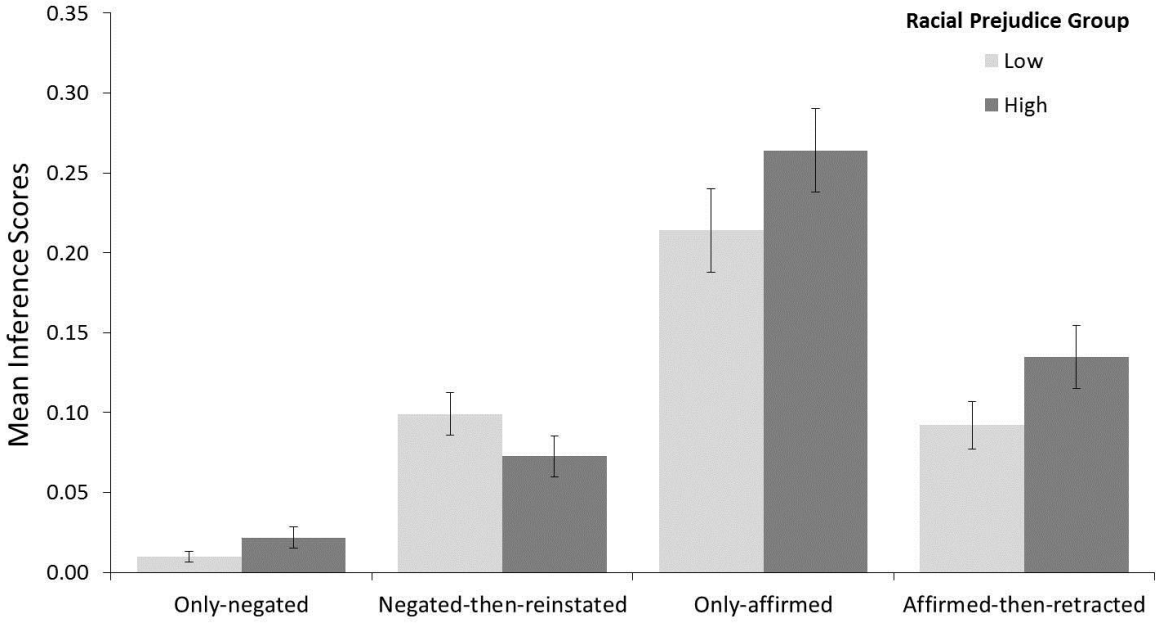

Figure 1. Mean inference scores across experimental conditions and racial prejudice groups (low, high) in Experiment 1. Error bars represent standard error of the mean.

[1] Participants' ATIA scores were also included as a continuous predictor, rather than relying on a median split, in a supplementary analysis of covariance. In this analysis, ATIA scores likewise did not significantly predict inference scores [$F(1,202) = 2.89$, $MSE = 0.008$, $p = .091$, $\eta^2_p = .014$, $BF_{10} = 0.57$].

Due to the non-significant main effect of prejudice group and absence of an interaction with condition, inference scores were collapsed across groups. A one-way between-subjects ANOVA on mean inference scores across conditions yielded a significant effect of condition [$F(3,203) = 56.98$, $MSE = 0.008$, $p < .001$, $\eta^2_p = .457$, BF_{10} 3.73 x 10^{23}]. Due to the need to conduct multiple follow-up pairwise contrasts, the target alpha was adjusted using the Holm-Bonferroni method (Holm, 1979). The results of the contrast analysis (see Table 2) showed that the tendency to refer to an Aboriginal robber was highest in the only-affirmed condition ($M = 0.24$, $SD = 0.13$) and lowest in the only-negated condition ($M = 0.02$, $SD = 0.03$), although scores in the only-negated condition were still significantly greater than zero, $t(51) = 4.194$, $p < .001$, $BF_{10} = 209.18$. Inference scores were significantly *higher* than the baseline only-negated condition in both the negated-then-reinstated ($M = 0.09$, $SD = 0.07$) and affirmed-then-retracted ($M = 0.11$, $SD = 0.09$) conditions. Additionally, scores in both of these conditions were significantly lower than in the only-affirmed condition, and did not differ from each other.

Table 2. Contrasts calculated on inference scores in Experiment 1.

Contrast		***t***	***df***	***p***	**Cohen's *d***	**BF_{10}**
1	Negated-then-reinstated vs. Only-negated	7.00	102	<.001*	>1	228.93
2	Negated-then-reinstated vs. Only-affirmed	7.31	102	<.001*	>1	6.49 x 10^{12}
3	Negated-then-reinstated vs. Affirmed-then-retracted	1.79	101	.077	0.35	0.50
4	Only-negated vs. Only-affirmed	11.69	102	<.001*	>1	1.14 x 10^{24}
5	Only-negated vs. Affirmed-then-retracted	7.50	101	<.001*	>1	2.09 x 10^{5}
6	Only-affirmed vs Affirmed-then-retracted	5.49	101	<.001*	>1	3.80 x 10^{8}

**Significant, Holm-Bonferroni corrected*

Discussion

Turning first to the impact of pre-existing racial attitudes on the processing of race-related (mis-)information, unexpectedly, and in contrast to Ecker and colleagues (2014), racial prejudice was not found to be a significant predictor of references to race-related (mis-)information generally. However, the lack of an interaction between prejudice group and condition, and specifically the finding that a retraction was equally effective for both high and low prejudice groups, replicates the results of Ecker and colleagues. Although the lack of a significant effect of ATIA group was surprising at first glance, we caution against an over-interpretation of this result given (a) the weak evidence for the null revealed by the BF analysis, and (b) the notable numerical difference in the same direction as in previous work.

Turning to our main research question, we found that a reinstatement of initially-negated information caused a significant increase in reliance relative to when that information was only negated, but it did not increase reliance on that information to the level observed when that information was presented as true throughout. On the contrary, the initial negation persisted to affect reasoning, consistent with a primacy account of the CIE. Our data therefore suggest that the CIE can occur regardless of whether the initially-provided information constitutes an affirmation or a negation. Consequently, these data support the notion of a 'reverse CIE' rather than that of an ironic rebound effect.

A remaining puzzle is why inference scores in the only-negated condition were significantly greater than zero. We cite two possible reasons for this effect: (a) participants relied on negated information despite knowing it was false, perhaps due to the difficulty of accurately interpreting and recalling negated information (Mckinstry, Dale, & Spivey, 2008; Reichle, Carpenter, & Just, 2000); or (b) there was enough innuendo in the story to drive reliance up regardless of the negation (Ecker et al., 2014). Unfortunately, Experiment 1 was

unable to differentiate between these two possibilities, as it lacked a no-misinformation reference condition in which Aboriginality was never mentioned. If scores in only-negated and no-misinformation conditions were of a comparable magnitude, then it could be concluded that the elevated level of misinformation reliance was simply due to narrative innuendo. By contrast, if the scores in a no-misinformation condition were lower than in the only-negated condition, then it could be inferred that the negation itself led to an elevated reliance on the misinformation.

Experiment 2 was designed to tease apart these two possibilities, while also providing a general replication of the first study. Experiment 2 was conducted with a larger U.S. sample of participants in order to generalize our findings. To this end, materials were converted to be applicable in a U.S. context.

Experiment 2

Similar to Experiment 1, participants were presented with a fictitious news report about a stabbing that took place outside of a convenience store in the U.S. state of Arkansas. Five versions of the report were created. Four versions of the report mirrored those used in Experiment 1, with the exception that the information concerning the racial characteristic of the suspect (i.e., Aboriginality) was changed to information concerning their religious affiliation (i.e., Muslim). To further disentangle the effect of a simple negation, and in contrast to Experiment 1, a further version of the report was created that did not reference the religious affiliation of the suspect at all (no-misinformation *control* condition). Participants also completed the islamophobia scale developed by Lee, Gibbons, Thompson, and Timani (2009). Participants' scores were used as a grouping factor similar to the ATIA scores in Experiment 1. In contrast to Experiment 1, which used a combination of pre-screening and a median split, participants in the current experiment were assigned to

groups by selecting those that scored in the upper and lower terciles on the distribution of islamophobia scores (see results section for more detail on the split). Experiment 2 therefore had a 5 (condition: only-negated, negated-then-reinstated, only-affirmed, affirmed-then-retracted, control) × 2 (islamophobia: high vs. low) between-subjects design.

Method (The method and analysis plan for this study were pre-registered and are available at https://osf.io/75bda/)

Participants. An a-priori power analysis using G*Power software (Faul, Erdfelder, Lang, & Buchner, 2007), based on the interaction effect size observed in Experiment 1, indicated that to achieve adequate power (1 – β = .8) required a sample of approximately 560 participants. In total, 756 U.S.-based participants were recruited on MTurk. Based on a-priori criteria, twenty-three participants were excluded from analysis due to: data recording errors (*n* = 3); task completion in under four minutes (*n* = 1; median completion time was 493 seconds); self-acknowledged lack of attention (i.e., a "no" response to a question whether their data should be used for analysis - see below for details; *n* = 1); poor recall of materials (i.e., providing incorrect answers to both fact-check questions - see results section for details; *n* = 4); and inconsistent responding on the islamophobia measure (see stimuli section below for details; *n* = 14). The final sample size was thus *N* = 733 (369 females, 360 males, 4 of undisclosed gender; age range 19-80, *M* = 39.79, *SD* = 12.52). Participants were paid US$0.45 for their participation.

Stimuli (all stimuli are available at https://osf.io/njh2g/). A fictional news report was used that comprised 13 messages and concerned a stabbing at a convenience store in the town of Fulton, Arkansas. The news reports differed across conditions at sentence 4 – where the critical information concerning the religious affiliation of the suspect was introduced (in all but the only-affirmed and control conditions) – and sentence 11 – where

the critical information was initially provided, retracted, or reinstated (in all but the only-negated and control conditions – for detail see Table 3). Participants' understanding of the report and their level of reliance on the critical religion-related information was assessed in a subsequent test phase. This consisted of two fact-recall questions, one open-ended event-summary item, and eight scale-based inference items (scored from 0-7) as well as the islamophobia scale. The order of questions was kept constant for all participants to avoid responses to the islamophobia scale biasing responses to inference items.

The fact-recall questions targeted arbitrary event details unrelated to the religious affiliation of the suspect (e.g., "What weapon was used in the attack?"). These questions were included to allow for the potential exclusion of participants who had not adequately encoded the news report. The event-summary item allowed participants to provide a response that did or did not mention the religious affiliation of the suspect (e.g., "A Muslim man stabbed someone at a convenience store" vs. "A man stabbed members of the public at a convenience store"). The eight scale-based inference items specifically probed participants' reliance on the critical religious-affiliation information. For example, the item "Police should conduct investigations in the local mosque" should have received a high level of endorsement from participants who believed that the suspect was a Muslim, and a low level of endorsement if they did not. The final inference item targeted the critical information directly ("Was the attacker a Muslim?").

The islamophobia scale (Lee et al., 2009) consists of 16 items scored on five-point Likert scales, and measures attitudes towards Muslims and Islam in general. One item ("I dread the thought of having a professor that is a Muslim") was removed from the measure due to it lacking relevance for the majority of the sample. As the remaining 15 items shared the same polarity, two reverse-coded items were added ("I feel comfortable in areas with

large Muslim populations" and "Islam is a peaceful religion"). As these two items were not a part of the validated scale created by Lee and colleagues, responses did not contribute towards participants' final islamophobia scores. However, scores on these two items were used to identify participants who responded inconsistently (i.e., uniformly to all items before reverse scoring) and thus were likely not paying attention to the content of the items. To identify inconsistent responders, we calculated the difference between the average score on the original items and the average score on the two reverse-coded items (after reverse-scoring). Participants were excluded from further analysis if their difference score was more than 2.2 interquartile ranges above the third quartile.

Table 3. Illustration of presentation order for critical race-related information across all conditions.

	Only-negated	**Negated-then-reinstated**	**Only-affirmed**	**Affirmed-then-retracted**
Message 4	"[…] the suspect is not a Muslim and the attack was not religiously motivated"	"[…] the suspect is not a Muslim and the attack was not religiously motivated"	*NO CRITICAL INFORMATION*	"A police bulletin issued at 8pm stated that they now have a Muslim man in custody"
Message 11	*NO CRITICAL INFORMATION*	"A police bulletin issued at 10:20pm stated that they now have a Muslim man in custody"	"A police bulletin issued at 10:20pm stated that they now have a Muslim man in custody"	"[…] suspect is not a Muslim and the attack was not religiously motivated"

***NB**: For pragmatic reasons sentences 4 and 11 from the control condition are not included as no reference was made to the religion of the attacker in either case.*

Procedure. Participants completed the task using the Qualtrics online survey platform (Qualtrics, Provo, UT). Participants were first presented with an ethics-approved information sheet, which specified that participation in the online experiment would be taken as implied consent. The news report was presented one message at a time; each

message remained on-screen for a pre-determined period (0.25 seconds per word), during which the participant was not able to continue. Participants then completed an unrelated anagram task for 1 min. The test items were then presented one at a time in the same order regardless of condition (i.e., fact-recall, event-summary, scale-based inference, and islamophobia items). At the end, participants were asked to indicate whether or not their data should be included in our analysis ("In your honest opinion, should we use your data in our analysis? This is not related to how well you think you performed, but whether you put in a reasonable effort."). This question could be answered with "Yes, I put in a reasonable effort"; "Maybe, I was a little distracted"; or "No, I really wasn't paying any attention". Participants then received a debriefing sheet and their validation code for MTurk. The experiment took approximately 8 minutes.

Results (data available at https://osf.io/teyu4/)

Results are presented in accordance with our pre-registered analysis plan, with minor exceptions that are marked in the text.

Coding procedure. As in Experiment 1, fact-recall questions were scored 0 for an incorrect response, and 1 for a correct response. Partial scores were not possible due to the multiple-choice format of the questions. The maximum fact-recall score was therefore 2. Upon inspection of the data, it became apparent that of those participants who had gotten only one fact-recall question wrong (n = 130), the overwhelming majority (n = 121) provided an incorrect answer to the first question ("How did the attacker flee the scene?") but not the second. Inspection of these incorrect responses found that all but 9 participants had provided the same incorrect answer ("On foot") instead of the correct answer ("In a van"). Because it is logically conceivable – and indeed likely in reality – that the attacker ran to his van to escape, both "On foot" and "In a van" were scored as correct answers to the first

question, whereas the remaining two answers were scored as incorrect (as indicated earlier, participants who scored 0 on both questions were removed from subsequent analysis).

The open-ended inference question ("Please provide a one-sentence summary of the events") was scored either 0 or 1 by a trained scorer who was blind to condition. Any uncontroverted mention of a Muslim being the chief suspect (e.g., "A Muslim man stabbed members of the public at a convenience store") received a score of 1. In contrast, any controverted statement (e.g., "Police arrest a Muslim man however find he was not responsible"), or any statement that did not reference the religious affiliation of the suspect received a score of 0. As in Experiment 1, ambiguous responses were scored 0.5 (e.g., "At first police [said] that it was a Muslim person but the surveillance could not make the person out […]"). Scale-based inference items were scored on a scale from 0 (strongly disagree/very unlikely) to 7 (strongly agree/very likely), with higher numbers indicating greater belief in a Muslim suspect[2]. Scores on these items were recoded to lie between 0 and 1 to be comparable to the open-ended item. Scores on the scale-based and open-ended items were then averaged to provide a single composite inference score ranging from 0 to 1; this was the main dependent variable of interest.

Inferential reasoning. Due to the lack of a significant effect of prejudice group in Experiment 1, the data were first analysed without regard to islamophobia group (i.e., collapsed across groups). A one-way between-subjects ANOVA on mean inference scores across conditions yielded a significant effect of condition [$F(4, 728) = 44.99$, $MSE = 0.042$, $p < .001$, $\eta^2_p = .198$, $BF_{10} = 3.57 \times 10^{30}$]. To explore the effect of condition, several planned contrasts were conducted. The results of these contrasts are reported in Table 4. First, to

[2] One item was erroneously scored on a scale from 0-8. This was accounted for in the analysis.

examine whether a negation drives an increase in the use of negated information above the level when that information is not mentioned, we assessed the difference between the control (*M* = 0.23, *SD* = 0.18) and only-negated (*M* = 0.18, *SD* = 0.17) conditions. Participants mentioned the critical information significantly less following a negation than if that information was never mentioned, although the corresponding Bayesian analysis indicated only very weak evidence for a difference between the two, meaning that this interpretation must be made with some caution. However, scores in the only-negated condition were nonetheless significantly above zero, $t(148) = 12.91$, $p < .001$, $BF_{10} = 1.29 \times 10^{22}$.

Second, to examine the presence of the standard CIE, we compared scores in the affirmed-then-retracted (*M* = 0.28, *SD* = 0.23) and control conditions. The results showed that participants made significantly more references to the critical information following its retraction than if it was never mentioned, in line with the clear majority of CIE research, although this difference was not given much support by the Bayesian analysis ($BF_{10} < 2$).

Finally, to explore how a reinstatement of initially-negated information influenced the subsequent use of that information, three further contrasts were computed. First, we compared inference scores between the negated-then-reinstated (*M* = 0.38, *SD* = 0.23) and only-affirmed (*M* = 0.46, *SD* = 0.22) conditions. It was found that participants made significantly fewer references to the critical information if it had been previously negated than when it appeared as true throughout. This difference represents a CIE of reversed polarity and replicates Experiment 1. Second, we compared inference scores between the negated-then-reinstated and only-negated conditions, revealing that participants made significantly more references to the critical information if it had been reinstated after an initial negation than when it was only negated. This reflects the effectiveness of a reversed-polarity (i.e., affirmative) correction. Finally, we compared inference scores between the

negated-then-reinstated and affirmed-then-retracted conditions. In contrast to Experiment 1, reinstatement of initially-negated information resulted in a significantly higher number of references to the critical information than when this information was first presented as true but then subsequently retracted[3].

Table 4. Contrasts calculated on RTM scores in Experiment 2.

Contrast		*t*	*df*	*p*	Cohen's *d*	BF_{10}
1	Control vs. Only-negated	2.59	292	.001*	.30	1.08
2	Affirmed-then-retracted vs. Control	2.24	288	.026*	.26	1.67
3	Negated-then-reinstated vs. Only-affirmed	2.97	296	.003*	.34	13.45
4	Negated-then-reinstated vs. Only-negated	8.65	290	<.001*	>1	4.70×10^{13}
5	Negated-then-reinstated vs. Affirmed-then-retracted	3.61	286	<.001*	.43	367.20

**Significant, Holm-Bonferroni corrected*

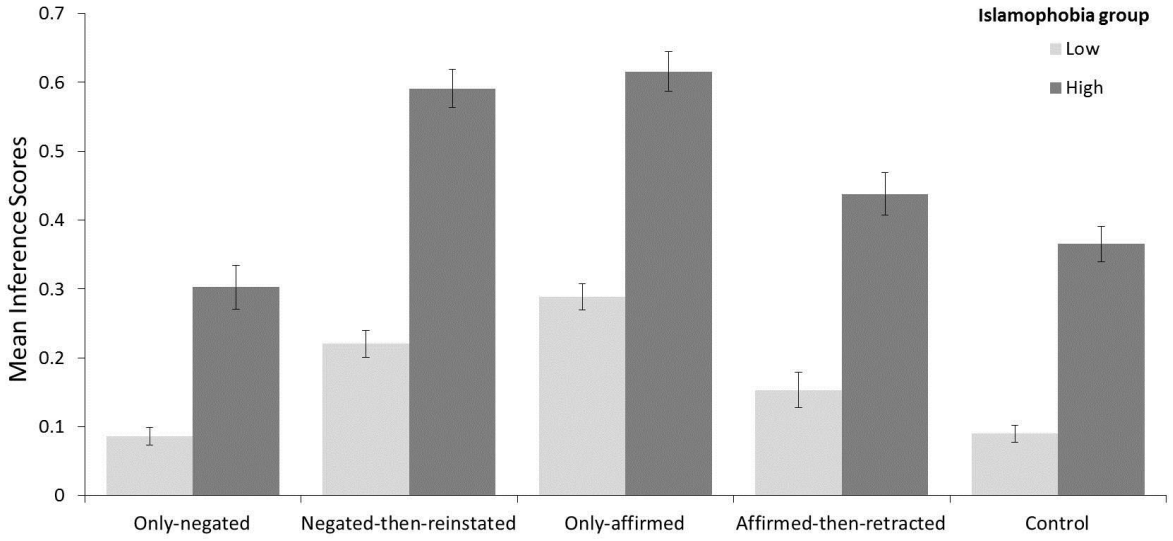

Figure 2. Mean inference scores across experimental conditions and islamophobia groups (low, high) in Experiment 2. Error bars represent standard error of the mean.

[3] In light of the significant CIE and reverse-CIE observed in both experiments, a further analysis (not preregistered) was conducted in order to examine whether a reinstatement resulted in a greater degree of updating than a retraction. This analysis revealed that the correction message was equally effective regardless of whether it was correcting an initial negation (i.e., a reinstatement) or an initial affirmation (i.e. a retraction, for more detail see supplementary material).

Analyses split by islamophobia group. Additional analyses were conducted taking into account participants' islamophobia scores (see Figure 2). Scores on the islamophobia measure ranged from 15 to 75 (representing the minimum and maximum possible scores on the measures). Participants were split into high and low islamophobia groups by taking the top and bottom terciles of the sample4. The mean islamophobia score for the low group (n = 257) was M = 15.72 (SD = 1.25), and for the high group (n = 244) it was M = 56.68 (SD = 10.44). A two way between-subjects ANOVA on inference scores with the factors condition and islamophobia group yielded a significant main effect of condition [$F(4,491) = 43.15$, $MSE = 0.029$, $p < .001$, $\eta 2p = .260$, $BF10 = 3.52 \times 1015$] and islamophobia group [$F(1,491) = 375.36$, $MSE = 0.029$, $p < .001$, $\eta 2p = .433$, $BF10 = 5.43 \times 1047$], with those high in prejudice making significantly more references to the religious affiliation of the attacker across all conditions. There was also a significant interaction between the two factors [$F(4,491) = 2.85$, $MSE = 0.029$ $p = .023$, $\eta 2p = .023$, $BF10 = 3.83$].

We explored the interaction with a set of analyses designed to look at the effect of islamophobia group membership on memory-updating following a correction (see Table 5). Analyses were split between updating following a retraction of prior information (analysis 1: only-affirmed – affirmed-then-retracted) and updating following a reinstatement of previously negated information (analysis 2: only-negated – negated-then-reinstated). The results across these analyses confirmed that for both groups, memory was updated following corrective information regardless of whether that information functioned as a retraction or a reinstatement. Additionally, it was found that participants high in islamophobia made more references to a Muslim suspect than did those in the low group

[4] In the event that the score identifying a tercile boundary was shared by participants both above and below that boundary, all participants with that score were included in either the low or high islamophobia group.

regardless of condition. However, when the correction functioned as a reinstatement, we observed a significant interaction between condition and islamophobia group that was absent when the correction functioned as a retraction, indicating that participants high in islamophobia updated their memory to a greater degree than those low in islamophobia.

Although these data appear to suggest that correction-congruence effects may be unique to the presentation of a reinstatement, such a conclusion cannot be supported based solely upon the presence of a significant interaction in one analysis and its absence in the other (Gelman & Stern, 2006; Nieuwenhuis, Forstmann, & Wagenmakers, 2011). In order to clarify interpretation of this difference, a final three-way [2 (group: high vs. low) × 2 (correction stage: pre vs. post) × 2 (polarity: reinstatement vs. retraction)] ANOVA was conducted (analysis 3 in Table 5). This analysis yielded no significant interaction between the three factors, suggesting that the magnitude of memory updating between prejudice groups was equivalent following corrective information regardless of its polarity, and rendering us unable to draw strong conclusions about the attitude-congruence effect.

Table 5: Results from analyses focused on memory updating following a correction (reinstatement or a retraction) for participants high versus low in islamophobia.

	F	η^2_p	***p***	BF_{10}
(1) Updating following a retraction				
Condition (only-affirmed vs. affirmed-then-retracted)	35.51	.149	<.001	1246.36
Group (high vs. low)	135.15	.401	<.001	1.66×10^{18}
Condition x Group	0.64	.003	.425	0.23
(2) Updating following a reinstatement				
Condition (only-negated vs. negated-then-reinstated)	85.04	.305	<.001	5.66×10^{7}
Group (high vs. low)	164.14	.458	<.001	1.05×10^{19}
Condition x Group	11.32	.055	<.001	0.74
(3) *Updating - interaction with polarity*				
Group x Correction Stage x Polarity	2.73	.007	.099	0.35

Discussion

Experiment 2 was designed to provide a conceptual replication of Experiment 1, in a different population and using different materials, while adding an additional control condition. In agreement with Experiment 1, we showed evidence for a reverse-CIE in that a reinstatement was not able to bring reliance on initially-negated information back to the level observed when that information was presented as true throughout. This confirms the robustness and generality of the reverse CIE. However, in contrast to Experiment 1, worldview produced clear differences in reliance on worldview-relevant information. In line with previous work (e.g., Ecker et al., 2014), across all conditions, participants who scored high on islamophobia made more references to the religion of the suspect than those who scored low; we take up this difference between studies in the General Discussion. Also in accord with previous work (e.g., Ecker et al., 2014; Wood & Porter, 2019), we found that the effectiveness of a retraction was not contingent on its congruence with participants' attitudes. However, we found tentative evidence that attitude-congruence modulated the effectiveness of a reinstatement. Although these data suggest that the effect of pre-existing attitudes is unique to cases in which previously negated material is subsequently reinstated, we are unable to draw strong conclusions regarding this effect due to the weak evidence observed in the Bayes factor analysis, and the lack of a significant effect of polarity on the effectiveness of a correction.

Finally, Experiment 2 afforded us an opportunity to examine whether the above-zero inference score in the only-negated condition of Experiment 1 resulted from the negated critical information ironically driving inferences up, or whether it was based just on spontaneous, innuendo-based references to the critical information. The fact that inference scores in Experiment 2 were higher in the no-misinformation control condition than the

only-negated condition speaks against the former and suggests that the level of reliance in the control condition likely reflects a propensity to endorse *any* information suggested by the inference statement, either due to the level of innuendo present in the narrative or because people will spontaneously endorse anything. The provision of a negation therefore appeared to depress – but not eliminate – such "background" endorsement.

General Discussion

Previous research into the CIE has exclusively focused on examining the impact of corrective information in a single polarity – the retraction of initially-*affirmed* information. In two experiments we examined how the CIE would be affected by a reversal of this polarity – providing a reinstatement of initially-*negated* information. We hypothesised that such a reversal may result in one of two outcomes: either a reinstatement of initially-negated material would result in a 'reverse-CIE' analogous to the conventional CIE, or it would result in the effective updating of memory to a level equivalent to when information is true throughout (or even beyond, i.e., an 'ironic rebound effect'). Both our studies clearly ruled out the possibility of an ironic rebound effect. As noted at the outset, this possible outcome was tied to a 'schema-plus-tag' model in which the initial negation is not encoded as a single meaningful unit but rather as two dissociated representations – a core supposition ('X') and a negation-tag ('not', e.g., Mayo et al., 2004) – raising the possibility that a subsequent reinstatement ('X') may uniquely interact with the core supposition to form a representation that is as strong (or conceivably stronger) than just being exposed to 'X' in isolation. As this effect was not observed, from here on we refrain from consideration of explanations that rely on this model of negation comprehension.

Instead the data suggested that the continued influence of initially presented information is not modulated by its polarity. Corrective information appears to engender

only partial belief updating, regardless of whether it corrects initially-affirmed or initially-negated material. Such a 'reverse-CIE' effect – the finding that a reinstatement cannot entirely offset an initial negation – is concordant with the previous literature demonstrating continued influence in the typical affirmation-negation (i.e., misinformation-retraction) paradigm. However, although the results are quite clear, the particular cognitive mechanisms underlying this reverse-CIE, much like those that underlie traditional CIE effects, remain uncertain. Given the symmetry of these effects, our results lend support to the notion outlined in the Introduction that an initial negation is recoded into a single meaningful informational unit. Although our data do not speak directly to the processes involved, our remaining discussion focuses on the possible mechanisms that may underlie this recoding.

As noted at the outset, a negation in the absence of prior information may be viewed pragmatically as a single assertive unit, and may therefore function no differently from an affirmation. This may occur because a negation is simply viewed as an assertion about a certain state of the world, meaning that the information contained therein is encoded at face-value. For example, the sentence "the suspects are local, male residents and none of them are of Aboriginal descent" may conceivably be encoded as an objective fact (i.e., there are no Aboriginal suspects; we revisit this potentially over-simplistic view later). Similarly, negations may be encoded as affirmations of their alternative attribute with the 'core' feature ('X') and its negation marker ('not') being incorporated into a single representation affirming the opposite 'Y' of the negated concept (Mayo et al., 2004). For example, consider the sentence "the suspect is not guilty", which is amenable to an alternative encoding, namely, the suspect is *innocent*. Critically, in both cases the negation can be encoded as a single meaningful unit ('not-X' or 'Y', respectively), therefore an initial

negation may be cognitively indistinct from an initial assertion. If this is the case, then a reinstatement, like a retraction, should perform the same cognitive role of correcting an earlier assertion of fact, and thus the underlying cognitive mechanisms should be indistinguishable. Although the exact nature of how such mechanisms produce the CIE is debated (see Chan et al., 2017; Lewandowsky et al., 2012), theoretical models such as the KReC framework discussed earlier (Kendeou & O'Brien, 2014) favour the notion of concurrent storage where both the original misinformation and its later correction persist in memory. Both may therefore be candidates for subsequent activation (Gordon, Quadflieg, Brooks, Ecker, & Lewandowsky, 2019; for reviews see Gerrig & O'Brien, 2005; Kendeou et al., 2019; Kendeou & O'Brien, 2014; Kendeou, Walsh, Smith & O'Brien, 2014). If negations are indeed encoded as single meaningful units, then concurrent-storage accounts and processes specified by the KReC framework offer a reasonable explanation for the CIE regardless of its polarity.

However, an additional account – one that could explain a reverse-CIE via a distinctly different cognitive mechanism – arises from further consideration of how a negation may be recoded as an affirmation of an alternative. Being able to transform a negation into an affirmation of the opposite largely depends upon its substantive ambiguity. While negations of information with specifically bipolar traits may be recoded (as above, the sentence "the suspect is not guilty" could be recoded as "the suspect is innocent"), more ambiguous negations (for instance, those concerning fuzzier 'multi-level' factors such as race or religion) may not be readily transformed in this fashion. For example, in Experiment 1, the 'not-Aboriginal' negation specifies a unipolar trait (i.e., one that does not have a distinct alternative), as suspects could belong to a number of other races (e.g., Caucasian, Asian, Indian, etc.). Therefore, in the context of the current work, the encoding of a negation as an

affirmation of its alternative attribute may be too simple an explanation for the effects observed. Instead, exposure to negating information not previously encountered may trigger the spontaneous generation of *several* simulated alternatives to the negated concept, consistent with the dominant 'rejection-based' account of negation comprehension (Dale & Duran, 2011; Hasson & Glucksberg, 2006; Kaup, Yaxley, Madden, Zwaan, & Lüdtke, 2007; Lüdtke, Friedrich, De Filippis, & Kaup, 2008). This model of negation comprehension suggests that a reinstatement may activate *multiple* conflicting representations that draw activation away from the correct representation (i.e., the reinstatement), which may reduce the likelihood of its subsequent retrieval (Gerrig & O'Brien, 2005; Kendeou & O'Brien, 2014), leading to a reverse-CIE via a markedly different process than the one outlined previously. Further empirical adjudication examining the methods by which an initial negation is processed will be required to clarify the relative contributions of the above possibilities.

As foreshadowed in the Discussion of Experiment 2, our paradigm also provided us with a unique opportunity to test the impact of only-negated narratives on subsequent critical information reliance, an area that CIE research has thus far neglected. Across both experiments, we found that only-negated narratives produced non-zero levels of inference during the subsequent retrieval phase. While Experiment 1 was unable to determine the cause for this non-zero effect, Experiment 2 clarified that only-negated narratives produced significantly fewer misinformation inferences than a control condition in which the misinformation was never mentioned. This result suggests that, rather than the negation driving a paradoxical increase in inference, the non-zero scores in the only-negated condition likely reflected low baseline levels of inference that were driven by the presence of innuendo in the narrative. Moreover, although the only-negated condition was originally

included in order to provide a comparison with the negated-then-reinstated condition, it may in fact serve as a more appropriate control condition against which to assess the typical CIE (i.e., via comparison to the affirmed-then-retracted condition). The significantly higher level of inference in the no-misinformation (i.e., typical) control condition relative to the only-negated condition suggests that scores in the typical control condition may actually represent participants' willingness to endorse any information suggested by the inference probes at retrieval. In contrast, only-negated narratives appear to decrease – but not eliminate – such "background" endorsement. An only-negated condition may thus provide a more appropriate control condition for typical CIE research than a no-misinformation condition, as the latter may lead to an underestimation of CIE magnitude due to background propensity to endorse information in the absence of relevant-knowledge.

A secondary aim of this study was to assess the effects of pre-existing attitudes on information-processing across experimental conditions, as the congruence of a reinstatement or a retraction with a participant's pre-existing beliefs could conceivably impact its effectiveness. However, the evidence was not entirely consistent across the two experiments and we therefore keep our discussion brief. In Experiment 2, we found that prejudicial attitudes determined to an extent what information participants relied upon in their reasoning: We found that participants scoring higher on a prejudice measure were more likely to refer to a suspect from that prejudiced group generally, that is, regardless of narrative condition (In Experiment 1, the main effect of prejudice tended in the same direction but failed to reach significance.) There are at least two possible reasons for this divergence between experiments. First, the size and make-up of our samples differed considerably: Experiment 1 investigated attitude effects in a relatively small sample of university students, whereas Experiment 2 canvassed members of the U.S. public. This

difference in the sample demographics may have influenced the results. For instance, although prejudice against Aboriginals is well-documented in Australia, university students in Australia display less prejudice towards Aboriginal people than the general public (Augoustinos, Ahrens, & Innes, 1994; Locke, MacLeod, & Walker, 1994; Pedersen & Walker, 1997). Second, and perhaps more important, our two experiments addressed different types of prejudice. While Experiment 1 examined racial prejudice towards Aboriginals, Experiment 2 examined prejudice against Muslims. One consequence of this change is that Experiment 2 examined a significantly more salient prejudice in light of the current global socio-political landscape. Islamophobia and the stigmatisation of Muslims has been amplified globally in recent years (Braunstein, 2017), particularly in the U.S. (Ogan, Willnat, Pennington, & Bashir, 2013). It is therefore reasonable to assume that Islamophobic beliefs are more prevalent in the general public in the U.S. than anti-Aboriginal sentiments are in Australia. We therefore suggest that the differences in the effect of prejudice across our two experiments may be, at least partly, driven by the increased salience of Muslim stereotypes in the current global climate.

Nevertheless, our finding that participants higher in prejudice measure were more likely to refer to a suspect from that prejudiced group replicates prior research also demonstrating a significant effect of pre-existing attitudes on mean level of reliance on attitude-relevant information (Berinsky, 2012; Ecker et al., 2011, 2014; Kull et al., 2003; Travis, 2010). Additionally (and consistent across both experiments), we found a *retraction* to be effective in reducing reliance on critical information for both high and low prejudice groups, consistent with the findings of Ecker and colleagues (2014). These data therefore support the notion that *post*-retraction differences in misinformation reliance may actually reflect *pre*-retraction differences, rather than attitude-congruence effects impacting the

effectiveness of the retraction (Hart & Nisbet, 2012; Nyhan & Reifler, 2010; Nyhan, et al., 2013; also see Ecker & Ang, 2019; Wood & Porter, 2019). Conversely, in the other polarity, we found tentative evidence that a reinstatement (after an initial negation) was more effective when it was attitude-congruent. Given the assumed complexities of negation comprehension detailed above, it is entirely possible that pre-existing beliefs uniquely interact with how an initial-negation (as opposed to an initial-affirmation) is processed by the reader and therefore affect its amenability to later correction. For instance, consistent with the scepticism notion outlined in the introduction, a participant high in prejudice may view an initial-negation of a Muslim attacker with a higher degree of suspicion (reasoning that negating the religious affiliation of the attacker implies that it may be a plausible but unarticulated assumption), thereby enhancing the effectiveness of a later reinstatement. However, we must refrain from an over-interpretation of this finding given that (a) both high- and low-prejudice participants updated their memory following a reinstatement (the difference was simply numerically larger for high-prejudice participants), and (b) the evidence in favour of an interaction between prejudice and polarity was weak. Nonetheless, this tentative finding is deserving of further investigation.

Acknowledgements

This research was facilitated by a World University Network (WUN) grant and University of Bristol internal funding awarded to Andrew Gordon, and a grant from the Australian Research Council to Ullrich Ecker and Stephan Lewandowsky (DP160103596). This research was further supported by RCUK funding from the EPSRC (EP/M506473/1). The authors have no competing interests to declare.

References

Augoustinos, M., Ahrens, C., & Innes, J. M. (1994). Stereotypes and prejudice: The Australian experience. *British Journal of Social Psychology*, 33(1), 125-141.

Braunstein, R. (2017). Muslims as outsiders, enemies, and others: The 2016 presidential election and the politics of religious exclusion. *American Journal of Cultural Sociology*, 5(3), 355-372.

Chan, M. P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, 28(11), 1531-1546.

Dale, R., & Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cognitive Science*, 35(5), 983-996.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, 14(2), 238-257.

DiFonzo, N., Beckstead, J. W., Stupak, N., & Walders, K. (2016). Validity judgments of rumors heard multiple times: the shape of the truth effect. *Social Influence*, 11(1), 22-39.

Ecker, U. K ., & Ang, L. C. (in press). Political Attitudes and the Processing of Misinformation Corrections. *Political Psychology*.

Ecker, U. K., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, 18(3), 570-578.

Ecker, U. K., Lewandowsky, S., Fenton, O., & Martin, K. (2014). Do people keep believing because they want to? Preexisting attitudes and the continued influence of misinformation. *Memory & Cognition*, 42(2), 292-304.

Ecker, U. K., Hogan, J. L., & Lewandowsky, S. (2017). Reminders and repetition of misinformation: Helping or hindering its retraction? *Journal of Applied Research in Memory and Cognition*, 6(2), 185-192.

Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, 38(8), 1087-1100.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175-191.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4), 328-331.

Gerrig, R. J., & O'Brien, E. J. (2005). The scope of memory-based processing. *Discourse Processes*, 39(2-3), 225-242.

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, 59(4), 601.

Gordon, A., Quadflieg, S., Brooks, J. C., Ecker, U. K., & Lewandowsky, S. (2019). Keeping track of 'alternative facts': The neural correlates of processing misinformation corrections. *NeuroImage,* 193, 46-56.

Hart, P. S., & Nisbet, E. C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research*, 39(6), 701-723.

Hasson, U., & Glucksberg, S. (2006). Does understanding negation entail affirmation? An examination of negated metaphors. *Journal of Pragmatics*, 38(7), 1015-1032.

Held, L., & Ott, M. (2016). How the maximal evidence of p-values against point null hypotheses depends on sample size. *The American Statistician*, 70(4), 335-341.

Held, L., & Ott, M. (2018). On p-values and Bayes factors. *Annual Review of Statistics and Its Application*, 5, 393-419.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65-70.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6), 1420.

Kahne, J., & Bowyer, B. (2017). Educating for democracy in a partisan age: Confronting the challenges of motivated reasoning and misinformation. *American Educational Research Journal*, 54(1), 3-34.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773-795.

Kaup, B., Yaxley, R. H., Madden, C. J., Zwaan, R. A., & Lüdtke, J. (2007). Experiential simulations of negated text information. *The Quarterly Journal of Experimental Psychology*, 60(7), 976-990.

Kendeou, P., Butterfuss, R., Kim, J., & Van Boekel, M. (2019). Knowledge revision through the lenses of the three-pronged approach. *Memory & Cognition*, 47(1), 33-46.

Kendeou, P., & O'Brien, E. J. (2014). 16 The Knowledge Revision Components (KReC) Framework: Processes and Mechanisms. In D. N. Rapp & J. L. G. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 353-377), Cambridge, MA: MIT Press.

Kendeou, P., Walsh, E. K., Smith, E. R., & O'Brien, E. J. (2014). Knowledge revision processes in refutation texts. *Discourse Processes*, 51(5-6), 374-397.

Kull, S., Ramsay, C., & Lewis, E. (2003). Misperceptions, the media, and the Iraq war. *Political Science Quarterly*, 118(4), 569-598.

Lee, S. A., Gibbons, J. A., Thompson, J. M., & Timani, H. S. (2009). The Islamophobia scale: Instrument development and initial validation. *The International Journal for the Psychology of Religion*, 19(2), 92-105.

Lewandowsky, S., Cook, J., Oberauer, K., Brophy, S., Lloyd, E. A., & Marriott, M. (2015). Recurrent fury: Conspiratorial discourse in the blogosphere triggered by research on

the role of conspiracist ideation in climate denial. *Journal of Social and Political Psychology*, 3(1), 142-178.

Lewandowsky, S., Stritzke, W. G., Oberauer, K., & Morales, M. (2005). Memory for fact, fiction, and misinformation: The Iraq War 2003. *Psychological Science*, 16(3), 190-195.

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4), 353-369.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), 106-131.

Locke, V., MacLeod, C., & Walker, I. (1994). Automatic and controlled activation of stereotypes: Individual differences associated with prejudice. *British Journal of Social Psychology*, 33(1), 29-46.

Lüdtke, J., Friedrich, C. K., De Filippis, M., & Kaup, B. (2008). Event-related potential correlates of negation in a sentence–picture verification paradigm. *Journal of Cognitive Neuroscience*, 20(8), 1355-1370.

Marsh, E. J., Meade, M. L., & Roediger III, H. L. (2003). Learning facts from fiction. *Journal of Memory and Language*, 49(4), 519-536.

Mayo, R., Schul, Y., & Burnstein, E. (2004). "I am not guilty" vs "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology*, 40(4), 433-449.

McKinstry, C., Dale, R., & Spivey, M. J. (2008). Action dynamics reveal parallel competition in decision making. *Psychological Science*, 19(1), 22-24.

Morey, R. D. (2015). Using the BayesFactor package version 0.9.2+. Retrieved from http://bayesfactorpcl.r-forge.r-project.org/

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, 14(9), 1105.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2), 303-330.

Nyhan, B., Reifler, J., & Ubel, P. A. (2013). The hazards of correcting myths about health care reform. *Medical Care*, 51(2), 127-132.

O'Brien, E. J., Rizzella, M. L., Albrecht, J. E., & Halleran, J. G. (1998). Updating a situation model: A memory-based text processing view. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(5), 1200.

Ogan, C., Willnat, L., Pennington, R., & Bashir, M. (2014). The rise of anti-Muslim prejudice: Media and Islamophobia in Europe and the United States. *International Communication Gazette*, 76(1), 27-46.

Pedersen, A., & Walker, I. (1997). Prejudice against Australian Aborigines: Old-fashioned and modern forms. *European Journal of Social Psychology*, 27(5), 561-587.

Pedersen, A., Beven, J., Walker, I., & Griffiths, B. (2004). Attitudes toward indigenous Australians: The role of empathy and guilt. *Journal of Community & Applied Social Psychology*, 14(4), 233-249.

Putnam, A. L., Wahlheim, C. N., & Jacoby, L. L. (2014). Memory for flip-flopping: Detection and recollection of political contradictions. *Memory & Cognition*, 42(7), 1198-1210.

Rapp, D. N., & Kendeou, P. (2009). Noticing and revising discrepancies as texts unfold. *Discourse Processes*, 46(1), 1-24.

Redlawsk, D. P. (2002). Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making. *The Journal of Politics*, 64(4), 1021-1044.

Reichle, E. D., Carpenter, P. A., & Just, M. A. (2000). The neural bases of strategy and skill in sentence–picture verification. *Cognitive Psychology*, 40(4), 261-295.

Rich, P. R., & Zaragoza, M. S. (2016). The continued influence of implied and explicitly stated misinformation in news reports. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(1), 62.

Schaffner, B. F., & Roche, C. (2016). Misinformation and motivated reasoning: Responses to economic news in a politicized environment. *Public Opinion Quarterly*, 81(1), 86-110.

Schwarz, N., Newman, E., & Leach, W. (2016). Making the truth stick & the myths fade: Lessons from cognitive psychology. *Behavioral Science & Policy*, 2(1), 85-95.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, 14(11), 1475.

Stadtler, M., Scharrer, L., Brummernhenrich, B., & Bromme, R. (2013). Dealing with uncertainty: Readers' memory for and use of conflicting information from science texts as function of presentation format and source expertise. *Cognition and Instruction*, 31(2), 130-150.

Swire, B., Ecker, U. K., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(12), 1948.

Van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, 1(2), 1600008.

Wilkes, A. L., & Leatherbarrow, M. (1988). Editing episodic memory following the identification of error. *The Quarterly Journal of Experimental Psychology*, 40(2), 361-387.

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 1-29.