The rational continued influence of misinformation

Saoirse A. Connor Desai[1], Jens K. Madsen[2], Toby D. Pilditch[3, 4]

University of New South Wales[1], London School of Economics[2], University of Oxford[3], & University College Londo4[3]

Author Note

*Correspondence concerning this paper should be sent to Saoirse Connor Desai, School of Psychology, University of New South Wales, 1006, Mathews Building, 8, Kensington NSW 2052; Email: saoirse.c.d@gmail.com

Supplementary materials and data are available at https://osf.io/6yq47

Abstract

Misinformation has become an increasingly topical field of study. Studies on the 'Continued Influence Effect' (CIE) show that misinformation continues to influence reasoning despite subsequent retraction. Current explanatory theories of the CIE tacitly assume continued reliance on misinformation is the consequence of a biased process. In the present work, we show why this perspective may be erroneous. Using a Bayesian formalism, we conceptualize the CIE as a scenario involving contradictory testimonies and incorporate the previously overlooked factors of the temporal dependence (misinformation precedes its retraction) between, and the perceived reliability of, misinforming and retracting sources. When considering such factors, we show the CIE to have normative backing. We demonstrate that, on aggregate, lay reasoners (N = 101) intuitively endorse the necessary assumptions that demarcate CIE as a rational process, still exhibit the standard effect, and appropriately penalize the reliability of contradicting sources. Individual-level analyses revealed that although many participants endorsed assumptions for a rational CIE very few were able execute the complex model update the Bayesian model entails. In sum, we provide a novel illustration of the pervasive influence of misinformation as the consequence of a rational process.

Keywords: Continued Influence Effect; Negation; Reliability; Dependency; Reasoning

1. Introduction

The harmful effects of misinformation have become a significant concern in contemporary society (Lewandowsky et al., 2017). These concerns are in part due to the ways that misinformation can spread rapidly online, as the news industry and general population alike, can share information outside of traditional information outlets. Media outlets can report false or inaccurate details while newsworthy events are still unfolding: as demonstrated when a prominent daily newspaper broke an online story incorrectly claiming Russian hackers had penetrated the US electricity grid. When, in fact, the electricity utility at the centre of the story found malware connected with Russian hackers on a single laptop, unconnected to the grid. Although the newspaper updated its article within a few hours of its original post, the incorrect information had already ricocheted through social media and the global news environment. Errors such as these are particularly worrying because several studies have shown that even clear and credible corrections often fail to eliminate the effects of misinformation (see Lewandowsky et al., 2012 for review). This phenomenon is known as the Continued Influence Effect (CIE) of misinformation (Ecker et al., 2010; Johnson & Seifert, 1994).

Continued influence studies examine corrections[1] to misinformation using variants of a laboratory paradigm first developed by Wilkes and Leatherbarrow (1988: but see also Johnson & Seifert, 1994). In a typical CIE experiment, participants read a fictitious news report presented a series of sequential statements. Misinformation, offering a causal explanation for the event's outcome, is presented and retracted later. A subsequent comprehension test typically shows that

[1] We use 'correction' and 'retraction' interchangeably throughout.

misinformation continues to influence memory and inferences even when participants understand and remember the retraction.

One scenario commonly used in the CIE paradigm concerns a warehouse fire in which initial reports suggest that flammable chemicals carelessly stored in a closet caused the fire (Connor Desai & Reimers, 2019; Ecker, Lewandowsky, Swire, et al., 2011; Guillory & Geraci, 2010; Johnson & Seifert, 1994; Wilkes & Leatherbarrow, 1988; Wilkes & Reynolds, 1999). One group of participants receive a retraction stating that “the closet was empty before the fire” thereby contradicting earlier misinformation. Retraction group responses are typically compared to a control group for whom there was no retraction, or a group who never saw the misinformation. The key CIE finding is that retractions are only partially effective: a retraction either results in no difference between a condition featuring a retraction and one in which there is no retraction (Johnson & Seifert, 1994) or reduces but fails to eliminate the misinformation's influence (Ecker, Lewandowsky, & Apai, 2011; Ecker, Lewandowsky, Swire, et al., 2011; Ecker et al., 2010; Guillory & Geraci, 2010, 2013; Rich & Zaragoza, 2016).

To date, there have been two leading cognitive explanations for the CIE (Gordon, Brooks, Quadflieg, Ecker, & Lewandowsky, 2017; Lewandowsky et al., 2012). The ***selective-retrieval*** account suggests that the CIE occurs when there is simultaneous storage of correct and incorrect information in memory; upon retrieval, misinformation is activated but inadequately suppressed (Ecker, Lewandowsky, Swire, et al., 2011). The ***model-updating*** account instead argues that corrections are poorly encoded because correcting misinformation leaves a gap in people’s mental model of the described event. Misinformation is therefore maintained because people prefer a coherent (incorrect) to an incomplete (correct) mental model. People are often unable to fill the gap in their mental-model, left by correcting misinformation unless a correction offers an alternative explanation for the outcome of the event (Connor Desai & Reimers, 2019; Ecker, Lewandowsky, & Apai, 2011; Ecker et al., 2010; Johnson & Seifert, 1994; Rich & Zaragoza, 2016). While selective-retrieval relates the CIE to and the inadequate suppression of misinformation at retrieval, the model-updating account posits a failure to update the mental-model stored in memory.

The selective-retrieval and model-updating accounts tacitly assume that the CIE is an error, or that normatively, a correction *should* reduce reliance on misinformation to the same level as would be observed if there was no misinformation at all. Both accounts presuppose that it is always appropriate to disregard earlier 'incorrect' information in favour of the later presented 'correct' information. For this assumption to hold, the 'correct' information must sufficiently "cancel out" the original 'incorrect' information. In this paper, as explained below, we explore the possibility for a rational foundation for the CIE, which considers the temporal dependence between misinformation and its subsequent correction, and the perceived reliability of the misinforming and retracting sources.

1.1. Temporal Dependence and Continued Influence of Misinformation

In this paper, we conceptualize the CIE as a scenario involving a contradiction between the testimonies of misinforming and retracting sources. As mentioned previously, there are two assumptions necessary for the retraction to "cancel out" the misinformation. First, the retracting source (either the same source at a second time-point or a different source at a later time-point) is perceived to be at least as reliable as the misinforming source. Second, the sources make their reports independently from one another. That is, the source of the retracting report bears no relation to the misinforming source, whether by sharing evidence (Schum, 1994), or background (Bovens & Hartmann, 2003; Madsen et al., 2020). When conceptualized as a matter of contradictory testimonies, the CIE could reflect a *dependency effect* whereby the observing reasoner perceives a source which contradicts themselves or two sources which contradict each other as less reliable, due to the inconsistency between their first and second reports. When a source contradicts their earlier statement, the reasoner does not know which testimony is truly correct but knows a source has been openly wrong on at least one occasion and therefore penalizes the reliability of all reporting sources. Crucially, in situations in which two sources do not know what the other has said (i.e. the sources are conditionally *independent*), yet provide contradictory reports, then the two reports should cancel out. In such a case, there should be no CIE. However, if the retracting source is aware of what the misinforming source has said (as is usually the case when a retraction is issued), the sources are no longer independent. When there is an asymmetry between the original and correcting reports: the correction is not just a statement of the hypothesis given the reliability of its source, but also a response to the original report, whilst the reverse is not true for the original (misinforming) reporter. Depending on the assumptions outlined below, this asymmetry can produce a difference in the reliability penalty applied to each source, given they are contradicting each other. Specifically, the corrector (second source) is penalized more than the misinformer (first source), and as such the correcting testimony is deemed weaker, and a continued belief in the (misinformation) hypothesis remains (i.e., a CIE).

Work in evidential reasoning on testimony has illustrated the necessity of these assumptions (Hahn et al., 2009; Hahn, Harris, et al., 2013; Hahn, Oaksford, et al., 2013; Schum, 1994; Schum & Martin, 1982) for disagreement (misinformation vs correction) to have a nullifying effect. For independence to hold in the CIE case, the “corrector” cannot be aware of the misinformation they are correcting – impossible in the same source case, and unlikely in the different source case. Instead, we must consider there to be a "temporal" dependency between the two pieces of information because the misinformation precedes its correction – something known to influence the reliability of sources providing evidence for a hypothesis (Madsen, Hahn, & Pilditch, 2018; 2019).

When considering the temporal dependence between misinformation and its retraction, and the reliability of misinforming and retracting sources, we seek to shed new light onto an effect that has, to date, escaped any normative account of how people *should* process corrections to misinformation. Sensitivity to temporal order has previously shown to affect the CIE wherein retractions followed by valid information/explanation are more effective than retractions preceded by valid information (Ecker et al., 2015). However, the temporal dependence between misinformation and its correction, and the impact of this temporal dependence on perceived source reliability, have yet to be considered within a formal framework.

In this paper, we formalize the CIE within a Bayesian Network (BN) model (Pearl, 1988) to test whether there may be a rational explanation for the CIE. Bayesian Networks use graph structures to represent probabilistic relations between hypotheses and evidence, showing which inferences a given model rationally permits. BNs can capture (in)dependencies between sources (e.g. Pilditch et al., 2020; Pilditch et al., 2018), and the influence of perceived reliability on belief revision (e.g. Madsen et al., 2018, 2020), both critical features of misinformation retraction scenarios. Following these studies, which explore the effects of contradictions when considering issues of dependence, we manipulate the source of the retraction. In our case, the source of the retraction is either the original misinformer, or a different source retracts the misinformation statement made by another source.

Bayesian normative frameworks facilitate the integration of people's subjective perception of the strength of evidence, their prior beliefs in hypotheses, and their perception of dependency and reliability. Bayesian approaches have been used to explain reasoning biases or errors from a rational perspective, including arguments from ignorance (Hahn, Oaksford, & Bayindir, 2005; Oaksford & Hahn, 2004), ad hominem (Harris et al., 2012; Oaksford & Hahn, 2012), slippery slope (Corner et al., 2011), and circular arguments (Hahn, Oaksford, & Corner, 2005). Bayesian networks have also been successfully applied to responses which appear to violate optimal responding, and involve contrary updating, such as belief polarization (Cook & Lewandowsky, 2016; Jern et al., 2014). By developing process-oriented models, such as Bayesian models, researchers can uncover causal mechanisms, and thereby better test interventions to prevent undesirable outcomes (e.g. the persistence of misinformation after a correction).

Exploring the CIE through a formal reasoning model, we find an alternative explanation that does not entail irrationality or bias. In line with model predictions, as explained below, we show that belief in the hypothesis (i.e. misinformation) *should* remain above prior levels. Instead, the reliabilities of sources that provide contradictory information are (appropriately) penalized whereby people perceive the second (retracting) reporter as less reliable than the first. Temporal dependence influences the effect when incorporated within the formal reasoning model. Correcting is often done by a source that is, in some way, linked with the source of misinformation (e.g. a second reporter working at the same network as the first). The correction must not only be considered a function of direct evaluation of the hypothesis in question but also of the (in)accuracy of the reports (i.e. the second report may be erroneous not only because of independent error but also the influence of the preceding report). The model, therefore, highlights a significant conceptual limitation to the traditional framing of CIE, which is silent on the dependency between misinformative and retracting reports. The crucial assumption here is that a correcting source is, *ceteris paribus*, less likely to be providing a truthful testimony when they are following (and contradicting) the report of another source, than when they are the first source providing an independent testimony. Examples of this inequality would be, for instance, concerns over cover-ups or attempts to control narratives, such as to divert blowback/blame arising from the original report. Lewandowsky et al. (2012) have acknowledged that the dismissal of retractions to misinformation could represent rational integration of prior biases with new information and that in principle, it is possible to instantiate the CIE within a BN. To date; however, the CIE has not been realized within a BN framework.

Contrary to the standard interpretation of CIE, we demonstrate that there are reasonable grounds under which people should maintain the misinformation, despite the provision of a correction; thus, the effect does not require deviation-based explanatory theories. Conceptualizing CIE, in this manner, also has implications for the kinds of interventions that are likely to be effective at reducing reliance on misinformation.

1.2. Source Reliability

Source reliability is essential for evaluating the evidentiary value of testimony. The quality of the source of information is critical to evaluating the suggested content – for example, if the source lies (is untrustworthy) or is mistaken (is inexpert), it may be entirely reasonable to disregard the suggested content. Although initially demonstrated normatively (e.g. Bovens & Hartmann, 2003; Hahn et al., 2009), empirical studies suggest that people incorporate subjectively perceived source credibility into evaluations of testimony (Harris & Hahn, 2009; Harris et al., 2016; Madsen, 2016). Indeed, adjusting a source’s reliability is prudent if new information, additional contradictory or corroborative reports, or insight into the relationship between sources becomes available (Madsen et al., 2020).

We consider the CIE to be a case of contradicting testimonies, such that, the “corrector’s” statement contradicts the “misinformer’s” statement (whether the same source or not), and argue that correcting source’s reliability suffers as a consequence of the contradiction. Several studies support this interpretation and show that the CIE may occur because some people do not *believe* the retraction (Guillory & Geraci, 2010, 2013; Ithisuphalap et al., 2020; O'Rear & Radvansky, 2020), demonstrating that source reliability is a critical component of processing retractions to misinformation. As the CIE involves a temporal dependence (the contradicting testimony *follows* the original, incorrect testimony), there is an additional reason for including source reliability within the scope of the study: the two sources of information differ in the information they have available (when reporting, the correcting source is often *aware* of the preceding, incorrect source’s statement, but not vice-versa), and this may influence judgments of reliability. In summary, given the CIE involves contradicting testimonies from sources with potentially different access and motivations for producing said testimony, there is a need for formalization, detailed in the section that follows.

1.3. A Bayesian Approach to Continued Influence of Misinformation

As mentioned, past CIE research has not provided a normative account of how people *should* process retractions of misinformation. Bayes’ theorem gives a normative belief revision model by integrating people’s subjective prior degrees of belief with the likelihood ratio to estimate the posterior degree of belief and expresses how a rational agent should revise their belief in a hypothesis H when faced with new evidence E. The probability P(H|E) represents the revised (posterior) degree of belief in the hypothesis H. The revised belief is a function of the prior belief P(H) and the conditional probability of observing the evidence E given H is true. Bayesian approaches to belief revision have been popular in research on argumentation (Hahn & Oaksford, 2006, 2007), and reasoning (Hayes et al., 2019; Oaksford & Chater, 2007), as well as other areas of cognition (Chater et al., 2010).

The BN framework (Pearl, 1988) is apt for capturing the difficulties of dependencies, and reasoning under uncertainty, that is integral to updating inferences in CIE. Bayesian networks are probabilistic graphical models which represent the relations between items of evidence and possible hypotheses allowing one to draw inferences about specific hypotheses based on observed evidence. The graph consists of a set of nodes representing variables of interest (i.e. hypotheses, evidence, reliability) and a set of directed links representing the probabilistic relations between variables, and in particular, the conditional dependencies. The quantitative component of BNs consists of conditional probability distributions for each variable in the graph. Bayesian networks, therefore, provide the means to test causal models of scenarios – including models of source reliability – and compare intuitive inferences of lay reasoners to a normative standard (Lagnado et al., 2013). The BN framework therefore offers a method for formalising the temporal dependency between misinformative and retracting reports and the impact that this contradiction has on misinforming and retracting sources (i.e., their perceived reliability, and the impact of their testimonies on the hypothesis).

1.4. The Present Study

In order to test the foundational assumptions of CIE formally, we constructed two BN scenarios; in the first scenario, the contradicting report (retraction) comes from the same source whereas, in the second scenario, the contradicting report comes from a second source. Figures 1 and 2 (below) show example BN models for the same and different source retraction conditions, respectively, and illustrate the assumptions of the temporal dependence between first and second reports and the impact on source reliability. Each model consists of a hypothesis node (i.e. the subject of the misinformation report), reporter nodes and reliability node(s). Figures 1 and 2 encapsulate the three stages involved in the CIE: the first (baseline) stage reflects the situation when there are no observations, at stage two a single piece of evidence (the misinformation) is available, and at stage three the contradictory (retraction) report is available. Referring to Figures 1 and 2, the values shown in stage one represent the prior probabilities elicited from participants before reading any reports (see Method section for further details). Each scenario includes unidirectional links between sources to represent the dependency between timepoints or reports. This link represents the assumption that an individual is aware of their previous statements or that, in general, people are aware of existing statements in the "world" (i.e. a retraction usually requires an awareness of the retracted statement). At stage two, the 'misinforming' reporter node is fixed to 'true' to reflect the positive report submitted by the reporter. The hypothesis node (denoted by H) value increases from stage 1 to stage 2, reflecting an increase in belief in the hypothesis (misinformation) after receiving a positive report. At stage three, the hypothesis and reliability nodes update when there is second a contradictory (retraction) report. In both the same and different source cases, belief in the hypothesis decreases relative to stage two but does not return the level observed in stage one, indicating a CIE. There are corresponding updates to the source reliabilities. In Figure 1, the same source reports the misinformation and retraction and source reliability increases from stage one to two but decreases in stage three after the retraction. Figure 2 shows that when a different source retracts the misinformation, the reliabilities of both the first and second reporters decrease from stage two to three.

Using a BN formalism, we examine the impact of source reliability on the estimated probability of the reported event being “misinformation” in each scenario - given the statements provided by those sources. For example, this conditional influence of reliability should mean that a source perceived as reliable will be more effective in persuading an individual of their (misleading, or correcting) statement (i.e., misinformation is *less likely* to be provided, conditional on that source being reliable). Crucially, along with the inclusion of reliability, we also capture a reasonable assumption of dependence between sources (within a single source, or between different sources) as correction temporally follows misinformation, which together with the consideration of reliability we argue yields a rational explanation for CIE. The model can consequently capture key differences between conditions. For example, it captures the clear difference in how reliability is updated (and belief in misinformation is also updated) when a single source contradicts their earlier statement, vs when a different source provides the contradiction. In the single-source scenario, the reliability of the source is penalized more heavily than in the two-source scenario because of the internal contradiction. Lastly, we elicit key parameters from participants themselves, such that we fit each model to participant’s assumptions. As a result of this, we can investigate the consistency of participant responses relative to the predictions of their own, fitted models. Following the CIE paradigm, participants in the present study read a set of brief news reports and complete a comprehension test. Crucially, we varied whether or not a sentence that appeared towards the end of the report retracted information provided earlier information, and whether the retracting source was the same or different to the misinforming source.

In this paper, we examine four hypotheses. We take pains to note that we do not base these predictions on the parameterizations of Figures 1 and 2, as these are solely illustrative examples of the more general structural (and ordinal parameter) relations that we note can lead to a rationally predicted CIE. The model is used to delineate several underpinning assumptions and effects that we outline below:

H1: We predicted higher endorsement of misinformation probes in the retraction condition than in the control condition in which there was no retraction (i.e. there is no retraction of the initial report).

H2: We predict that the conditional probability measures provided by participants, which assess the participants own interpretation of the relationship between a source's likelihood of a statement being in error, given their reliability and possible contradiction of previous statements, will yield a predicted CIE when using these measures to parameterize the Bayesian network model.

H3: In line with model predictions, participants will penalize source reliability when there is a contradiction. The perceived reliability of the retracting source (at stage three) will decrease relative to misinforming source (at stage two), as shown in Figures 1 and 2. The same vs different source manipulation is exploratory, and there is no directional prediction for the impact on reliability.

H4: We expect to elicit the CIE in terms of the posterior probability measure such that participants will retain belief in the hypothesis (i.e. the misinformation) despite a retraction.

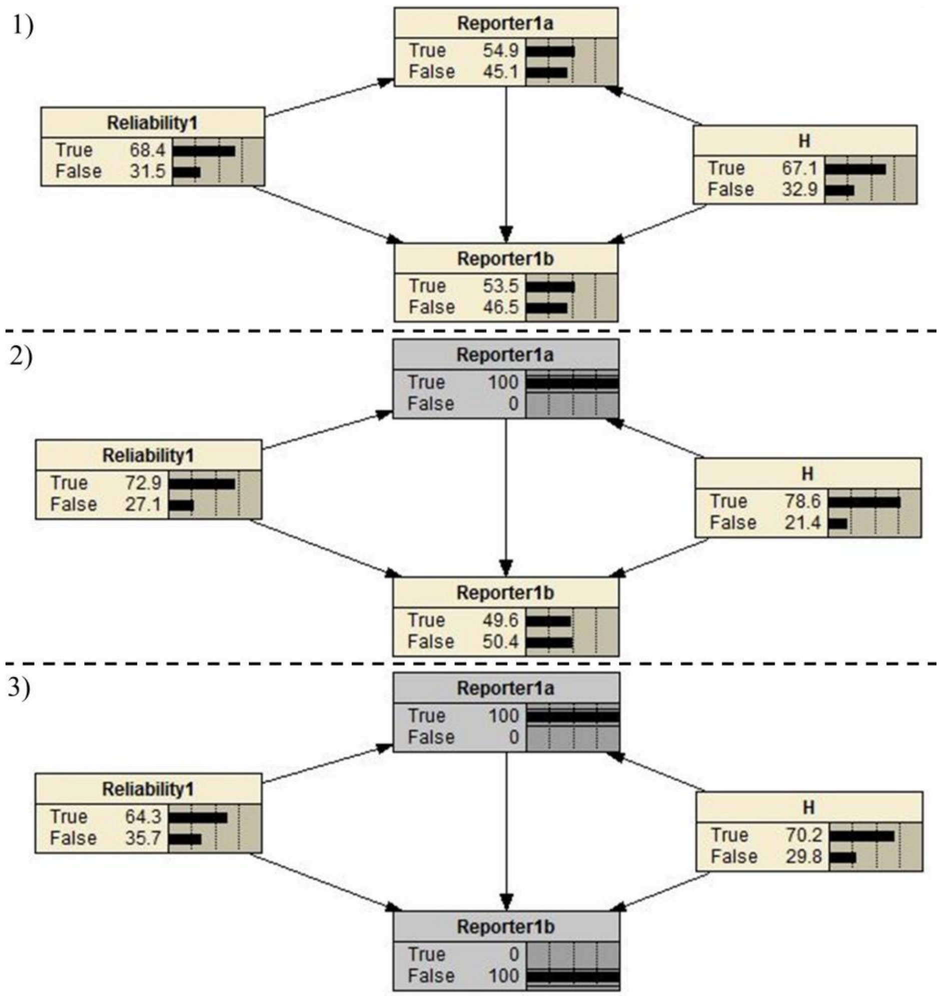

Fig 1 BN model for the retraction (same source) condition across three separate time points, where H represents the belief in the hypothesis in question (misinformation), which is informed by the misinformer (represented as Reporter1a) and later this same source as a correction (Reporter1b). Given this is the same source at two points in time, the reliability of the source (Reliability1) connects to both instances of the reporting source. 1) Baseline (no observation) stage, 2) Single positive (first) report stage (i.e. control

condition) –misinformation stated, and 3) Final (retraction) state given a second, correcting report from the same reporter.[2] Figure created using the AgenaRisk Bayesian Network software (*AgenaRisk*, 2019).

[2] BN model parameters taken from the mean estimates across the retraction same condition for the police officer scenario.

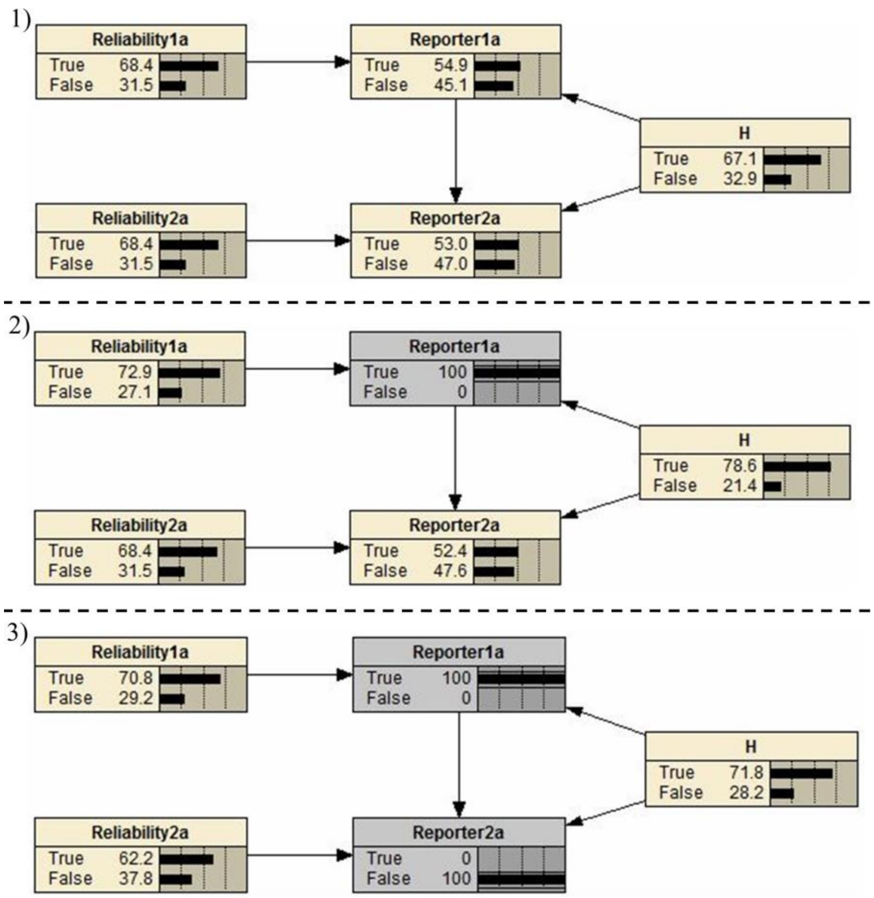

Fig 2 BN model for the retraction (different source) condition across three separate time points. H represents the belief in the hypothesis in question (misinformation), which is informed by the misinformer (represented as Reporter1a) and later a separate source as a corrector (Reporter 2a). Given this, the two sources have their reliability specified (Reliability 1a and 2a, respectively). 1) Baseline (no observation) stage, 2) Single positive (first) report stage (i.e. control condition) – misinformation stated, and 3) Final (retraction) state given a second, correcting report from a separate reporter[3].

[3] BN model parameters taken from the mean estimates across the retraction different condition for the police officer scenario.

2. Method

2.1. Participants

We aimed to recruit 105 participants (N = 35 per condition). We based our estimate on the principle that formal statistics sources suggest that Central Limit Theorem tells us that the sampling distributions of the means will be approximately normal, even if the underlying data distributions are non-normal when the sample size is larger than 30 (Field, 2013). There was no prior work to inform an effect-size based power calculation as this was a novel design. In total, 101 participants from Prolific Academic https://www.prolific.co/ completed the experiment. There was a mean age of 31.57 (*SD* = 9.6), and there were 71 females and 30 males. Participants were paid £1.50 (~$1.97) for their time (*Median* = 12.87 minutes, *SD* = 5.78).

2.2. Materials, Design and Procedure

To replicate CIE, we used materials adapted from past research (Gordon et al., 2017; Johnson & Seifert, 1994). We opted for shorter scenarios than those typically used in CIE studies to keep the study duration to a minimum since participants answered an extensive set of conditional probability questions (see Supplementary Materials https://osf.io/6yq47). It was also necessary to ensure that the non-critical details in the scenario were independent of the hypothesis (misinformation) and the evidence (retraction), to model the participants' responses. We selected four scenarios for the main study, that produced the largest baseline CIE (i.e. the difference between retraction and control conditions) from a set of eight pilot scenarios (*N* = 70). In the main study, participants read four scenarios (motorcycle accident/police officer, medical controversy/independent reviewer, music festival/local journalist, and explosion/police spokesperson) consisting of six sequentially presented sentences (see Supplementary Materials). The materials provided minimal information about the source of the misinformation and retraction, to better control for differences in perceived source reliability. In this sense, the inferences participants make about reliability should be based on the contradiction, and the source's profession (i.e. police officer, local journalist, independent reviewer, police spokesperson).

We assessed the effect of retracting information between groups (Control, Retraction – Same Source, Retraction – Different Source). We randomly assigned participants to a condition and randomized the presentation order of the scenarios across participants. Table 1 shows that in each scenario, sentence 2 differed between control and retraction conditions for each event. In retraction conditions, sentence 2 contained (mis)information. Whereas in the control condition, sentence 2 contained incidental information to provide a baseline for the misinformation endorsement test. The key sentence (sentence 5) was identical in all conditions. Given exposure to sentence 2, sentence 5 did or did not correct previous information. The source of the (mis)information (sentence 2) and retraction (sentence 5) were either from the same source or a different source, in the retraction conditions.

Before reading the scenarios, participants provided prior estimates for the reliability of the sources of misinformation that would appear in the subsequent reports on a scale of 0 (Extremely unlikely) to 100 (Extremely likely). They then provided six conditional probability estimates per report for each of the two sources (i.e. the misinformer and the retractor). Eliciting conditional probability estimates in this way is necessary because there is no general normative function that captures the dependency relationship between the misinformer and the retractor. Participants provided their probability estimates (on the same scale as above) for the same and different source conditions and thus capturing the specific assumptions regarding the nature of the dependency relationship on the individual level.

1. *If a police officer is reliable, how likely are they to make an erroneous statement in reporting about a road accident, **if that same police officer is contradicting their own earlier statement?***
2. *If a police officer is reliable, how likely are they to make an erroneous statement in reporting about a road accident, if that same police officer is **contradicting an earlier statement from another police officer**?*

Responses to the above questions illustrate one aspect of the dependency relationship: the perceived likelihood of a source providing an erroneous report (i.e. the misinformation) given that they are reliable, *before* learning about the specifics of the event. Participants provided six conditional probability estimates per event scenario (24 in total), all on 0-100 sliders (0 and 100 denote the same as in the above). Due to the elicitation of conditional probabilities, no free parameters were requiring posthoc fitting.

The modelling process generated a Condition (3) x Scenario (4) matrix, creating 12 "group" models. Participants provided three types of estimates; participants supplied the first two types of estimates before reading the reports and supplied the third type of estimate after reading the report. First, participants provided estimates for the reliability of the sources of misinformation that would appear in the subsequent reports (e.g. police officer), which we call *reliability priors* (e.g. *How likely are police officers to be reliable in their reporting?*). Second, participants estimated the likelihood of the sources making an erroneous statement about the reported event (e.g. road accident) conditional on the source being reliable or not (e.g. *If a police officer is reliable, how likely are they to make an erroneous statement in reporting about a road accident?*), and conditional on whether the source was contradicting/corroborating their own or another source's statement. We call these estimates conditional probability estimates. These allow for a full range of possible assumptions regarding the relationship between the misinforming and retracting reports (from no influence, to complete dependence). Finally, participants estimated the probability of the focal hypothesis and the reliability of the source in each scenario.

We created each model using the elicited responses for each estimate, using as many of the responses of possible. All three conditions could use the reliability priors from all participants, along with the conditional probabilities for Reporter 1, as these were similar across conditions. However, the conditional probabilities for Reporter 2 were condition-specific (i.e. the Retraction Same Source condition could only use conditionals elicited from that condition – see question 1 above).

Lastly, as eliciting participants' "prior" probability estimates for the focal hypothesis in each scenario (i.e. an estimate of how likely the reported event is to be accurate, in each scenario) beforehand were likely to interfere with CIE (via the premature introduction of the misinformation), we reverse-engineered the priors from the control condition posterior estimates. As the control condition had a single positive Reporter 1 observation (rather than multiple contradicting observations), a prior probability could be calculated via Bayes Theorem using the known likelihood (i.e. Reporter 1 conditionalized parameters) and provided posterior estimates. For example, given the posterior (P (Hypothesis | Report) for the journalist scenario was 77.58%, and the elicited conditional probabilities for the journalist reporter were a probability of the journalist being correct if reliable (P (Report | Hypothesis , Reliable)) of 68.94% and correct if unreliable (P (Report | Hypothesis, ¬ Reliable)) of 31.06% (with these probabilities reflected for the chance of error), conditional on a probability of being reliable (P (Reliable)) of 54.1%, then dividing the above posterior by the conditionalized reporter likelihood results in a prior

(P(Hypothesis)) of 75.6%. All parameters, except the generated prior, were directly elicited from participants and fed into the model at the group level. This “prior” could then be implanted in the retraction condition models, further ensuring that the models did not have any free parameters.

Responses to a set of misinformation probes that followed each scenario measured the continued influence of misinformation (see Table 1). Participants rated each probe on a 7-point scale from ‘strongly disagree’ to ‘strongly agree’. In line with previous CIE methods, probes referred to the critical information (sentence 5). Higher levels of misinformation probe endorsement captured the extent to which participants integrated the misinformation (sentence 2 in the retraction conditions) into their understanding of the news report.

After rating the probes, participants provided posterior probabilities on a similar scale used for prior beliefs. For example, in the scenario in Table 1, participants were asked: 1) Given everything you know so far about the incident in question, how likely do you think it is that the accident occurred because the driver was intoxicated/travelling over the speed limit? 2) Given everything you know so far about the incident in question, how likely to do you think it is that the police officer is reliable in their reporting? Participants who received a retraction from a different source as the misinformation provided an additional estimate for the reliability of the second reporter.

Table 1

Example news report scenario and misinformation probes

Sentence	Control	Retraction (Same Source)	Retraction (Different Source)
Example News Report			
Sentence 1	A motorcyclist died yesterday after being knocked off his bike by a car.		
Sentence 2	***Officer Jones reported that the driver of the car had been travelling over the speed limit.***	***Officer Jones reported that the driver of the car was intoxicated.***	***Officer Jones reported that the driver of the car was intoxicated.***
Sentence 3	The accident happened on the A7 north of Carlisle.		
Sentence 4	The motorcyclist was 30 years old and had two children.		
Sentence 5	***Officer Jones revealed that the car driver was not intoxicated.***	***Officer Jones revealed that the car driver was not intoxicated.***	***Officer Smith revealed that the car driver was not intoxicated***
Sentence 6	The driver of the car was also injured in the incident.		
Example Misinformation Probes			
Question 1	Drink-driving charges should be brought against the driver of the car		
Question 2	The driver should be forced to complete a drink-driving awareness course		
Question 3	A breathalyzer would have returned a positive result		

3. Results

Bayesian analyses were performed with JASP statistical software (JASP Team, 2018) and assumed an uninformed prior. The use of Bayes factors (BFs) additionally allowed us to infer evidence for the null hypothesis, wherein a BF_{10} of less than one third is considered substantial support for the null (Dienes, 2014).

3.1. Misinformation Endorsement Ratings

A Bayesian repeated-measures ANOVA was used to determine the effect of retraction condition and scenario type on mean misinformation endorsement ratings. Strong evidence was found for the main effect of condition, $BF_{Inclusion}$[4] = $1.917 * 10^{12}$, and scenario, $BF_{Inclusion} = 5.44 * 10^9$, but no interaction, $BF_{Inclusion} = 0.122$. The model including just main effects was the strongest fit, BF_M[5] = 131.26, and was decisive; overall, $BF_{10} = 2.105 * 10^{22}$. As illustrated in Fig. 3 scenarios differed in misinformation endorsement ratings from one another, and there was a differential influence of condition.

Critically, the effect of condition indicated significantly higher endorsement ratings following the presentation and retraction of misinformation compared to when there was no misinformation presented. This result indicates that a CIE was observed across all scenarios, such that a retraction was insufficient to bring endorsement ratings back to baseline.

A Bayesian repeated-measures ANOVA was also used to establish whether there was an effect of scenario order (i.e. whether a participant read the scenario first, second, third or fourth) on misinformation endorsement ratings. This found no effect of scenario order, $BF_{Inclusion}$ =

[4] $BF_{Inclusion}$ reflects the change in odds from the sum of the prior probabilities of models that include the effect, to the sum of posterior probabilities of models including the effect.
[5] BF_M reflects the change from prior to posterior odds for the given model.

0.435, and strong evidence for a null effect of its interaction with scenario (type), $BF_{Inclusion}$ = 0.034. There was, however, strong evidence for the main effect of scenario, $BF_{Inclusion}$ = 5.46 * 10^9, with the model including only this main effect yielding the strongest fit, BF_M = 6.132, and decisive overall, BF_{10} = 7.54 * 10^9.

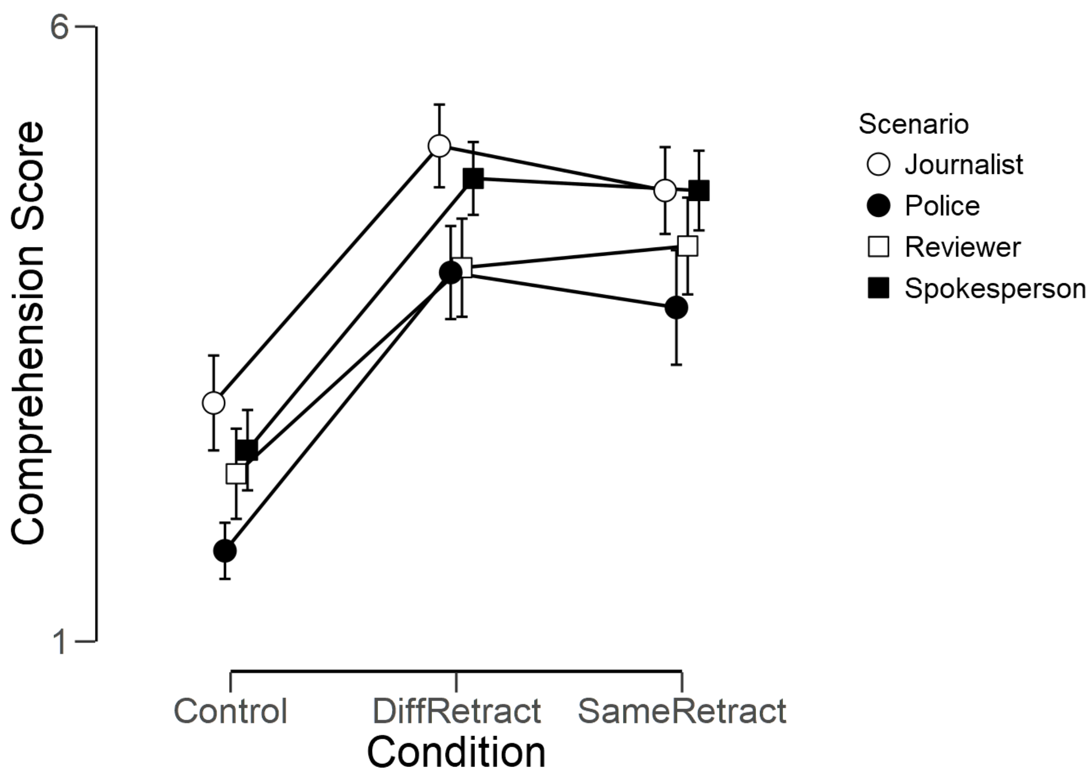

Fig 1 Mean misinformation endorsement ratings, split by scenario (line) and condition (horizontal axis). Error bars reflect 95% CI. The scale ranged from 1 = strongly disagree to 7 = strongly agree. Misinformation probes were more strongly endorsed in the retraction conditions than the control condition. A retraction was insufficient to bring endorsement ratings to baseline levels. Note that points have been offset on the x-axis to improve legibility.

3.2. Bayesian Model Fits

Using the conditional probabilities and priors elicited from participants, group means of these estimates were used to parameterize two group-condition models for each scenario. The conditional probabilities and priors for each first reporter and reliability node were fitted based on all participants, with two notable exceptions. First, conditional probabilities for the second reporter were based solely on estimates from the condition of relevance (i.e. we only used estimates from the retraction (different source) condition to parameterize the entailed different second reporter in that condition). Secondly, we reverse-engineered prior probabilities for each hypothesis (via Bayes Theorem) using the posteriors provided by the control condition. More precisely, taking the control condition BN model, the posterior for the hypothesis was fitted, given the single positive report. Retracting the observation could reveal the approximate prior (absent observations) for that hypothesis. This “prior” was fitted into the models for the two retraction conditions.

Figures 1 and 2 illustrate models for each experimental condition of the police officer scenario, fitted from participant data according to the protocol outlined above. Several significant trends are noticeable: Firstly, as expected, given a single positive reporter (stage 2), belief in the hypothesis (H) increases, and the predicted likelihood of corroboration from the second report increases. However, when the second, contradicting report is observed (stage 3), the belief in the hypothesis (H) does *not* return to prior (stage 1) levels. Instead, the reliability of sources decreases given the contradiction, this decrease is most influential in the second reporter (different condition) but is also substantial when the same reporter contradicts themselves (Fig. 2, stage 2 to stage 3).

Critically, the reason for this effect (retention of belief in H, but the reduction in perceived reliability) is due to the capturing of the temporal dependence from first to the second report. Put another way; the models capture the intuition that the second report is made with an awareness of the first report (whether internally in the case of the same reporter condition or via general narrative in the different reporter condition). The elicited conditional probabilities from participants then capture the manner and strength of this influence.

3.3. Participant Estimates

Returning to participant data, we again use Bayesian repeated-measures ANOVA to examine whether probability estimates correspond to the BN model predictions (and thus map onto a CIE) or corroborate the misinformation endorsement ratings (and indicate an absence of CIE – against fitted normative prescription).

3.3.1. Hypothesis

Turning first to posterior estimates of belief in the hypothesis, we find main effects of condition, $BF_{Inclusion} = 3.328 * 10^9$, and scenario, $BF_{Inclusion} = 41812.52$, but no interaction, $BF_{Inclusion} = 0.467$. The model consisting of the main effects along was the strongest fit, $BF_M = 34.27$, and enjoyed decisive support overall, $BF_{10} = 2.247 * 10^{14}$. As Fig. 4 illustrates, these effects corroborate misinformation endorsement ratings; wherein there is the retention of belief in misinformation despite its retraction. Crucially, this shows that participants generally deviate from the prescribed CIE entailed by the BN models, decreasing belief in the hypothesis below the control condition (and prior), given the retraction.

We again checked for order effects for posterior estimates of belief in the hypothesis across scenarios, findings strong evidence for a null effect of presentation order, $BF_{Inclusion =}$ 0.028, and its interaction with scenario type, $BF_{Inclusion =}$ 0.02. There was, however, the main effect of scenario type, $BF_{Inclusion} = 5354.61$, with the model including only scenario type yielding the strongest fit, $BF_M = 94.26$ and being decisive overall, $BF_{10} = 7963.11$.

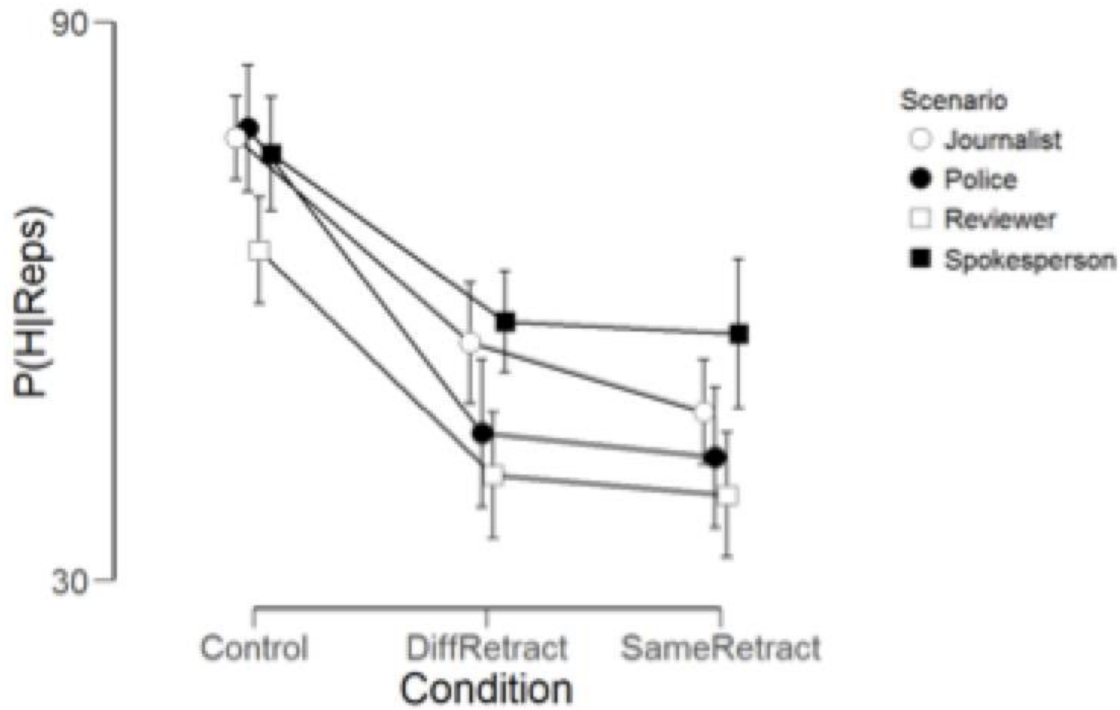

Fig 2 Posterior estimates of belief in the hypothesis (H), given all reports, split by scenario (line) and condition (horizontal axis). Error bars reflect 95% CI. Note that points have been offset on the x-axis to improve legibility.

47

3.3.2. Individual Differences

We performed an exploratory analysis to examine individual differences in participants' model predictions, posterior estimates of belief in the hypothesis, and their misinformation endorsement ratings. We first generated individual model fits for each participant and scenario. These individual fits were based on each participants own prior estimate of reliability for the source of the reports, conditional probabilities for sources, and the prior probability of the (misinformed) report being true, which was reversed engineered from the control group mean for that scenario (as such a prior could not be elicited from the same participant without undermining the CIE framework premise – see section 2.2). We then computed the proportion of participants whose fitted BN model predicted a CIE and the proportion of participants who exhibited a CIE (i.e. retained belief in the misinformation despite a retraction), separately for each of the four scenarios tested (see Table 2). A prediction of CIE was defined as a posterior probability for the hypothesis after both reports that remained above the level of the prior.

The first finding of note from is that although around half of participants provided parameter estimates that *should* lead to the CIE, very few actually do. We confirmed this by performing Bayesian tests of association using a joint multinomial sampling plan and default priors, separately for each scenario[6], to test the null that there was no association between observed and predicted CIE. The journalist scenario produced a $BF_{01} = 3.020$, the reviewer and police officer scenarios produced $BF_{01} = 3.618$, and the spokesperson scenario produced a $BF_{01} = 2.233$, indicating moderate evidence for the null.

[6] It was necessary to perform separate analyses for each scenario as the levels of scenario were not independent.

Table 2 Percentage of participants with predicted given their BN model and observed CIE by scenario

Predicted CIE	Observed CIE	Journalist	Police	Reviewer	Spokesperson
Yes	No	39.71	38.24	47.06	32.84
Yes	Yes	5.88	32.35	5.88	14.93
No	No	45.59	23.53	42.65	38.81
No	Yes	8.82	5.88	4.41	13.43

To corroborate this finding, we also performed a Bayesian regression to examine whether participant's parameter estimates predicted their misinformation endorsement ratings and found that the model including BN parameter estimates as a predictor was 2.11 more likely than an intercept only model[7].

Taken together, we find that although on an individual basis, many participants detailed probabilistic relationships between model components that *should* produce a CIE, very few participants in fact went on to exhibit one in their own probability responses. Furthermore, inclusion in the former category did not predict inclusion in the latter. Finally, there was anecdotal evidence that participant's parameter estimates predicted their misinformation endorsement ratings.

[7] The Bayesian regression was performed using the BayesFactor package in R using default Cauchy priors and participant as a random effect.

3.3.3. Reliability

Turning next to estimates of reliability, we add to the repeated measures ANOVA analysis a within-subject factor of the change in reliability estimates from, the prior, to posterior. Here we find significant main effects of condition (control > retraction different and same), $BF_{Inclusion} > 10^{20}$, scenario, $BF_{Inclusion} = 124.44$, and prior-posterior (posterior < prior), $BF_{Inclusion} > 10^{20}$. Figs 5, 6, and 7 illustrate the significant interaction of condition and prior-posterior, $BF_{Inclusion} > 10^{20}$, wherein reliability estimates increased in the control condition (Fig. 5; where no contradiction occurs, and in line with the increase observed in Figs 1 and 2, stage 2), but decreased in both retraction conditions (Figs 6 and 7; also, in line with model predictions illustrated in Figs 1 and 2, stage 3). Lastly, we also observed a strong interaction of scenario and prior-posterior, $BF_{Inclusion} = 75.92$, wherein the spokesperson scenario entailed smaller changes from the prior to the posterior than the 3 remaining scenarios. The model, including the above-supported terms, yielded the strongest fit, $BF_M = 484.97$, and was decisive; overall, $BF_{10} = 1.559 * 10^{28}$.

Finally, we note that the retraction condition showed no significant difference in posterior reliability estimates between the two different (first and second) reporters, $BF_{10} = 0.135$, contrary to model predictions (wherein there should be a more substantial reliability penalty for the second reporter because of the contradiction).

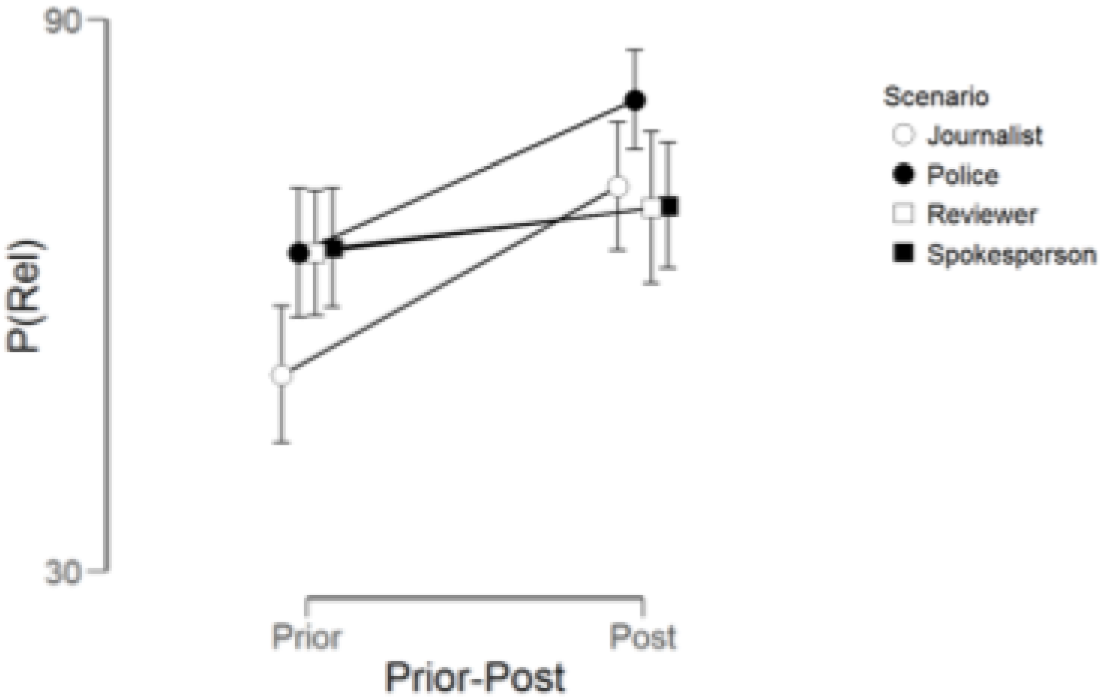

Fig 3 Control condition reliability estimates for reporters from prior to posterior (reports observed), split by scenario (lines). Error bars reflect 95% CI. Note that points have been offset on the x-axis to improve legibility.

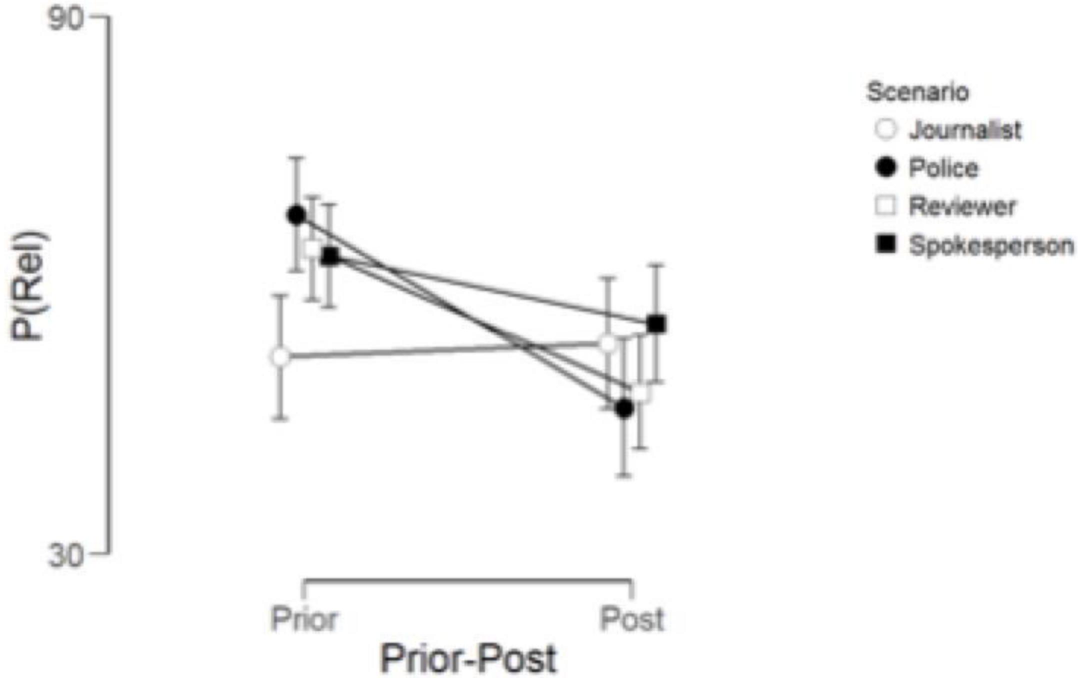

Fig 4 Retraction different condition reliability estimates for reporters from prior to posterior (reports observed), split by scenario (lines). Error bars reflect 95% CI.

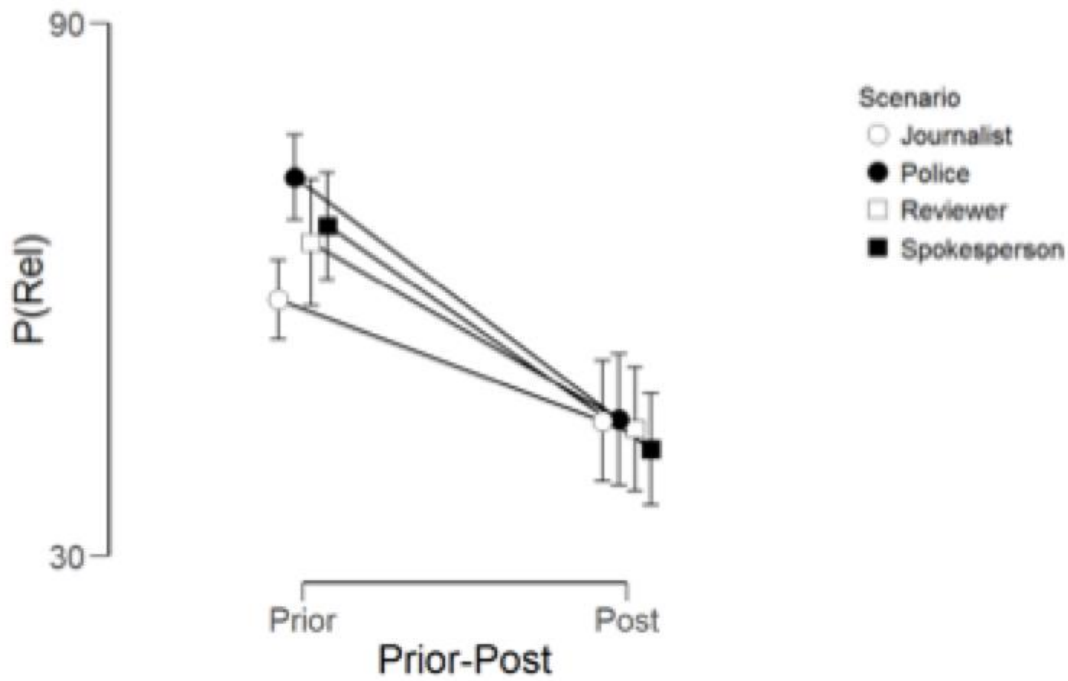

Fig 5 Retraction same condition reliability estimates for reporters from prior to posterior (reports observed), split by scenario (lines). Error bars reflect 95% CI.

4. Discussion

This paper formalizes the continued influence of misinformation (Johnson & Seifert, 1994; Lewandowsky et al., 2012) in a Bayesian network model which accounts for the temporal dependency between the misinformation and retraction reports, and its impact on source reliability. When accounting for the temporal dependency between misinformation and its retraction, we find a rational account for the continued influence effect. We find that participant's responses broadly fit with the predictions of this account, and show that belief in the hypothesis (i.e. the misinformation) remains above prior level, and instead, participants penalize the reliability of the second reporter (i.e. retraction's source). Participants perceived the second (retracting) reporter as less reliable than the first (misinforming) reporter, irrespective of whether the second reporter was the same or different from the first. However, participant's posterior estimates also decreased below their priors, and against their model predictions. This finding is contrary to standard CIE accounts (that people continue to rely on retracted misinformation when they should not); instead, we show that people do not always continue to rely on misinformation even though they should!

An individual-level analysis of the data revealed that people can, and do, endorse the necessary assumptions for a rational account of the CIE. However, most participants were unable to incorporate these assumptions into their posterior probability judgments or their misinformation endorsement ratings. Put another way, participants did not achieve the complex Bayesian update that the model entails; namely, integrating the conjunction of temporal dependency, and its impact on source reliability, to estimate the strength of the evidence for the retracted misinformative report. Crucially, these findings show that a "rational" CIE is possible when conceptualized in Bayesian terms, even with people's own assumptions about the relationships between the factors in play.This finding shows the integration of contradictory sources is difficult even when one considers the temporal dependency between the misinformation and retraction.

Reliability estimates revealed that participants decreased their estimate for a reporter who contradicts themselves, in line with model predictions. In the different source condition, participants decreased their reliability estimates for the first reporter and increased reliability estimates of the second reporter (both correct according to the model). Interestingly, the second reporter was considered more reliable than the first in the police officer and independent reviewer scenarios, but less reliable than the first in the journalist and police spokesperson scenarios. Descriptively, this discrepancy in reliability estimates demonstrates participant’s sensitivity to the different types of sources and suggests individual variability in source reliability priors. The fact that we elicited prior estimates of different source type’s reliability before presenting the scenarios, and still find differences in the reliability estimates between the control and retraction conditions, also demonstrates that, overall, participants are not solely remaining consistent with their prior estimates of reliability. Finally, we observed a classic CIE whereby misinformation endorsement ratings showed that a retraction, whether from the same or a different source, did not bring endorsement ratings back to the baseline level (as shown in the control condition). Participants continued to rely on retracted misinformation. Misinformation probes were more strongly endorsed when misinformation was presented and retracted than when the scenario did not involve a retraction of misinformation. This result is consistent with previous CIE studies that have included a "no misinformation" control condition who find baseline levels are higher than zero (e.g. Gordon et al., 2017; Johnson & Seifert, 1994; Rich & Zaragoza, 2016).

At the aggregate level, both posterior estimates for the hypothesis and misinformation endorsement ratings showed retention of the misinformation despite being retracted. The posterior estimates for the hypothesis, while lower than the control condition, still showed substantial retention of belief in the retracted hypothesis. It is worth noting that the posterior probability estimates used in the present study measure belief updating and are therefore qualitatively different from the traditional continued influence measures which measure comprehension. Our novel probability estimate measures are, arguably, a more sensitive measure of the CIE than traditional measures, as they demonstrate the uncertainty that often follows a correction of the misinformation. People might reduce their belief in misinformation after a retraction but not completely rule out the possibility that the misinformation is still valid because they do not believe the retraction (Guillory & Geraci, 2010, 2013; O'Rear & Radvansky, 2020).

The present study did not include a condition in which misinformation is presented but never retracted, as is common in most CIE research. The control and retraction conditions sufficiently demonstrated higher endorsement in the retraction condition. Excluding the no retraction condition meant that it was not possible to assess the effect of the retraction. Including such a condition would make it possible to directly compare the novel approach used in the present study with previous CIE research and presents an opportunity for follow-up research. The findings here also involve scenarios, and retractions, that are shorter and more straightforward than the ones people may encounter in everyday life. Replicating the findings with richer, more causally complex scenarios is necessary to establish whether the modelling process still predicts the CIE.

Taken together, we show that participants *should* exhibit the CIE (according to fitted BN models), maintaining belief in the retracted misinformation. We find this effect with standard behavioural measures used in the CIE literature (Brydges et al., 2018; Gordon et al., 2017), and observe retention of the hypothesis with novel probability estimate P(H) measures. We also find appropriate penalization in reliability estimates given a contradiction among first and second reporters – something hitherto unnoticed in CIE studies but predicted by our formalism. An individual-level analysis of the data revealed that although many participants endorsed assumptions for a rational CIE very few of these participants went on to provide posterior probability estimates in agreement with their model predictions. Furthermore, there was little evidence that participant's parameter estimates predicted their misinformation endorsement. To put our findings in context with previous explanatory theories (Lewandowsky et al., 2012), which tacitly assume that CIE is an error, we provide a process-oriented theory that can give a rational (and testable) framework for CIE. We do not argue that the sole explanation for the CIE relates to the inferences made about the reliability of sources providing contradictory of information; instead, we argue that source reliability plays a crucial role in the inferences that people generate after a correction to initially presented information, and that there is a richer context to consider when contemplating the CIE. We illustrate the (rationality-reversing) impact of one such reasonable context expansion, but this is not to outright refute previous descriptive theories *per se*.

To conclude, we provide a formal account of CIE using a BN framework and show that CIE is in some circumstances, rational. This approach captures the qualitative inferences participants make about the reliability of sources providing contradictory information and suggests that perceived reliability moderates the degree to which people are willing to integrate contradictory reports. The models described here are normative in the sense that they provide an argument for why CIE can be the product of a rational process. We do not make the argument that the models *describe* participant reasoning itself. This research demonstrates that it is possible to model CIE using a BN framework. Building upon current explanatory theories of CIE, and the insight may represent the reliabilities of sources providing contradictory information, is a promising direction for future research.

5. References

AgenaRisk. In. (2019). Agena Ltd. http://www.agenarisk.com/

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford University Press on Demand.

Brydges, C. R., Gignac, G. E., & Ecker, U. K. (2018). Working memory capacity, short-term memory capacity, and the continued influence effect: A latent-variable analysis. *Intelligence, 69*, 117-122.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811-823.

Connor Desai, S., & Reimers, S. (2019). Comparing the use of open and closed questions for Web-based measures of the continued-influence effect. *Behavior research methods, 51*(3), 1426-1440.

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in cognitive science, 8*(1), 160-179.

Corner, A., Hahn, U., & Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *Journal of Memory and Language, 64*(2), 133-152.

Ecker, U. K., Lewandowsky, S., & Apai, J. (2011). Terrorists brought down the plane!—No, actually it was a technical fault: Processing corrections of emotive information. *The Quarterly Journal of Experimental Psychology, 64*(2), 283-310.

Ecker, U. K., Lewandowsky, S., Cheung, C. S., & Maybery, M. T. (2015). He did it! She did it! No, she did not! Multiple causal explanations and the continued influence of misinformation. *Journal of Memory and Language, 85*, 101-115.

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

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. sage.

Gordon, A., Brooks, J. C., Quadflieg, S., Ecker, U. K., & Lewandowsky, S. (2017). Exploring the neural substrates of misinformation processing. *Neuropsychologia, 106*, 216-224.

Guillory, J. J., & Geraci, L. (2010). The persistence of inferences in memory for younger and older adults: Remembering facts and believing inferences. *Psychonomic Bulletin & Review*, *17*(1), 73-81.

Guillory, J. J., & Geraci, L. (2013). Correcting erroneous inferences in memory: The role of source credibility. *Journal of applied research in memory and cognition*, *2*(4), 201-209.

Hahn, U., Harris, A. J., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic*, *29*(4), 337-367.

Hahn, U., Harris, A. J., & Oaksford, M. (2013). Rational argument, rational inference. *Argument & Computation*, *4*(1), 21-35.

Hahn, U., & Oaksford, M. (2006). A normative theory of argument strength. *Informal Logic*, *26*(1), 1-24.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: a Bayesian approach to reasoning fallacies. *Psychological review*, *114*(3), 704.

Hahn, U., Oaksford, M., & Bayindir, H. (2005). How convinced should we be by negative evidence. Proceedings of the 27th Annual Conference of the Cognitive Science Society,

Hahn, U., Oaksford, M., & Corner, A. (2005). Circular arguments, begging the question and the formalization of argument strength. Proceedings of AMKLC'05, International symposium on adaptive models of knowledge, language and cognition,

Hahn, U., Oaksford, M., & Harris, A. J. (2013). Testimony and argument: A Bayesian perspective. In *Bayesian Argumentation* (pp. 15-38). Springer.

Harris, A. J., & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1366.

Harris, A. J., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The appeal to expert opinion: Quantitative support for a Bayesian network approach. *Cognitive science*, *40*(6), 1496-1533.

Harris, A. J., Hsu, A. S., & Madsen, J. K. (2012). Because Hitler did it! Quantitative tests of Bayesian argumentation using ad hominem. *Thinking & Reasoning*, *18*(3), 311-343.

Hayes, B. K., Banner, S., Forrester, S., & Navarro, D. J. (2019). Selective sampling and inductive inference: Drawing inferences based on observed and missing evidence. *Cognitive psychology*, *113*, 101221.

Ithisuphalap, J., Rich, P. R., & Zaragoza, M. S. (2020). Does evaluating belief prior to its retraction influence the efficacy of later corrections? *Memory*, *28*(5), 617-631.

Jern, A., Chang, K.-M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological review, 121*(2), 206.

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.

Lagnado, D. A., Fenton, N., & Neil, M. (2013). Legal idioms: a framework for evidential reasoning. *Argument & Computation, 4*(1), 46-63.

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.

Madsen, J. K. (2016). Trump supported it?! A Bayesian source credibility model applied to appeals to specific American presidential candidates' opinions. CogSci,

Madsen, J. K., Hahn, U., & Pilditch, T. D. (2018). Partial source dependence and reliability revision: the impact of shared backgrounds. CogSci,

Madsen, J. K., Hahn, U., & Pilditch, T. D. (2020). The impact of partial source dependence on belief and reliability revision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

O'Rear, A. E., & Radvansky, G. A. (2020, Jan). Failure to accept retractions: A contribution to the continued influence effect. *Mem Cognit, 48*(1), 127-144. https://doi.org/10.3758/s13421-019-00967-9

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 58*(2), 75.

Oaksford, M., & Hahn, U. (2012). Why are we convinced by the ad hominem argument? Source reliability or pragma-dialectictics. *Bayesian Argumentation*, 39-58.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann Publishers.

Pilditch, T. D., Hahn, U., Fenton, N., & Lagnado, D. (2020). Dependencies in evidential reports: The case for informational advantages. *Cognition, 204*, 104343.

Pilditch, T. D., Hahn, U., & Lagnado, D. A. (2018). Integrating dependent evidence: naïve reasoning in the face of complexity. CogSci,

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.

Schum, D. A. (1994). *The Evidential Foundations of Probabilistic Reasoning*. Northwestern University Press.

Schum, D. A., & Martin, A. W. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review*, 105-151.

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

Wilkes, A., & Reynolds, D. (1999). On certain limitations accompanying readers' interpretations of corrections in episodic text. *The Quarterly Journal of Experimental Psychology Section A, 52*(1), 165-183.