

Keeping Track of 'Alternative Facts':

The Neural Correlates of Processing Misinformation Corrections

Andrew Gordon^{1,2}, Susanne Quadflieg¹, Jonathan C. W. Brooks^{1,3}, Ullrich K. H.
Ecker⁴ & Stephan Lewandowsky^{1,4}

Authors' Affiliations:

¹School of Psychological Science, University of Bristol, Bristol, UK.

²MIND Institute, University of California, Davis, Sacramento, USA.

³Clinical Research and Imaging Centre, University of Bristol, Bristol, UK

⁴School of Psychological Science, University of Western Australia, Perth, Australia.

Corresponding author:

Andrew Gordon (andrew.gordon@bristol.ac.uk)

School of Psychological Science, University of Bristol
12a Priory Road, Bristol BS8 1TU
Telephone: 01179546621

Word count abstract: 221

Word count manuscript (excluding references, tables, and figure captions): 8410

Abstract

Upon receiving a correction, initially presented misinformation often continues to influence people's judgment and reasoning. Whereas some researchers believe that this so-called continued influence effect of misinformation (CIEM) simply arises from the insufficient encoding and integration of corrective claims, others assume that it arises from a competition between the correct information and the initial misinformation in memory. To examine these possibilities, we conducted two functional magnetic resonance imaging (fMRI) studies. In each study, participants were asked to (a) read a series of brief news reports that contained confirmations or corrections of prior information and (b) evaluate whether subsequently presented memory probes matched the reports' correct facts rather than the initial misinformation. Both studies revealed that following correction-containing news reports, participants struggled to refute mismatching memory probes, especially when they referred to initial misinformation (as opposed to mismatching probes with novel information). We found little evidence, however, that the encoding of confirmations and corrections produced systematic neural processing differences indicative of distinct encoding strategies. Instead, we discovered that following corrections, participants exhibited increased activity in the left angular gyrus and the bilateral precuneus in response to mismatching memory probes that contained prior misinformation, compared to novel mismatch probes. These findings favour the notion that people's susceptibility to the CIEM arises from the concurrent retention of both correct and incorrect information in memory.

1. Introduction

With the advent of direct-to-consumer media and continuous news coverage, people routinely receive information that is subsequently declared invalid. For example, within hours of the Oklahoma City bombing in 1995, the media cast suspicion on a U.S. citizen of Jordanian descent as a possible culprit, before an American of Irish origin (Timothy McVeigh) was ultimately apprehended. It can prove surprisingly challenging for people to update their memories under such circumstances: Even in the face of explicit retractions or corrections, initially presented misinformation often continues to influence people's judgment and reasoning (Chan, Jones, Hall Jamieson, & Albarracín, 2017; Ecker, Lewandowsky, Swire, & Chang, 2011; Ecker, Lewandowsky, & Tang, 2010; 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 effect is known as the continued influence effect of misinformation (CIEM).

The CIEM is remarkably robust. Numerous lines of research have demonstrated that the CIEM can occur even when people receive prior warnings that they will be exposed to misinformation (e.g., Ecker et al., 2010; Marsh & Fazio, 2006) or demonstrably attend to and understand the correction (e.g. Johnson & Seifert, 1994; Marsh, Meade, & Roediger, 2003). This robustness seems to arise because the processing of corrective information requires increased cognitive effort and is particularly error-prone (Johnson-Laird, 2012; Verschueren, Schaeken, & d'Ydewalle, 2005). In other words, integrating corrective information into one's existing body of knowledge is widely considered a psychological task of particular difficulty (see Gentner & Stevens, 2014; Johnson-Laird, 1983). However, the source

of this difficulty remains a matter of debate (Chan et al., 2017; Ecker, Lewandowsky, & Apai, 2011; Henkel & Mattson, 2011; Moscovitch & Melo, 1997; van Oostendorp & Bonebakker, 1999).

According to the model-updating account, people struggle to retain a coherent understanding of events upon trying to replace original (mis-)information with corrective information and, thus, end up discarding corrections instead of fully integrating them into the model (Ecker et al., 2010; Kendeou, Walsh, Smith, & Brien, 2014; Verschueren et al., 2005). By contrast, according to the concurrent-storage hypothesis, people succeed at encoding corrections, but subsequently fail to remove or inhibit the initial (mis-)information and are, thus, left with competing memory traces for the same event (Ayers & Reder, 1998; Catarino, Küpper, Werner-Seidler, Dalgleish, & Anderson, 2015; Masson, Potvin, Riopel, & Foisy, 2014; Shtulman & Valcarcel, 2012; Vosniadou, 2012). Although either scenario could give rise to the CIEM, only the concurrent-storage hypothesis implies that the effect is not merely encoding-dependent, but also related to problems with selective memory retrieval (Catarino et al., 2015; Ecker et al., 2011; Jacoby & Whitehouse, 1989; Lewandowsky et al., 2012).

Unfortunately, traditional psychological measures struggle to disentangle these two stages of information processing as the CIEM's only behavioural indicator is people's level of reliance on misinformation – a measure that inherently conflates the impact of encoding- and retrieval-related processes. As a consequence, in their search for alternative means of inquiry, CIEM scientists have recently turned their attention towards neuroimaging techniques (e.g., Edelson, Dudai, Dolan, & Sharot, 2014; Gordon, Brooks, Quadflieg, Ecker, & Lewandowsky, 2017; Kaplan, Gimbel, &

Harris, 2016). Such techniques promise to be particularly relevant to the field given that they can monitor brain activity during different stages of information processing.

Difficulties with *encoding* corrective information, for instance, could be expected to increase activity in brain regions supporting the detection of unexpected information (such as the anterior cingulate cortex; Braver, Barch, Gray, Molfese, & Snyder, 2001; Bush, Luu, & Posner, 2000; Carter & van Veen, 2007). Furthermore, difficulties with *retrieving* corrective information should modulate activity in brain regions involved in selecting information from, or inhibiting information in, memory (such as the hippocampus and regions of the pre-frontal cortex; e.g., Anderson et al., 2004; Butler & James, 2010). Based on these predictions, we conducted two functional magnetic resonance imaging (fMRI) studies that extended the classic CIEM paradigm (see Johnson & Seifert, 1994; Wilkes & Leatherbarrow, 1988) to a version suitable for the fMRI environment.

2. Study 1

In Study 1, participants were asked to imagine that they worked as an editor at a newspaper. They were told that their work required them to complete three tasks: First, to read and sort incoming stories as potentially positive or negative news. Second, to learn what the fact-checked version of each of these stories looked like. Third, to decide whether an image that was meant to accompany each story was suitable for publication based on whether it matched or mismatched the verified facts. Using this novel paradigm, we examined the neural correlates of misinformation processing at two important stages of information processing: Upon presentation of the fact-checked stories, participants' neural response during the *encoding* of information that either corrected or confirmed prior information was

captured. In addition, upon presentation of the images, brain activity related to *retrieving* corrective or confirming information while evaluating image suitability was recorded. The method and analysis plan for this study were preregistered using the Open Science Framework (at <https://osf.io/ew8qb/>).

2.1 Methods

2.1.1 Participants

32 participants (20 female) aged 18-35 years ($M = 24.4$, $SD = 4.29$) took part in this study. Eight of them (4 females) failed to pass the head motion check for fMRI data (as outlined below) so that only their behavioural data (but not their fMRI data) were considered during data analysis. Five additional participants had been recruited but were excluded from all analyses due to withdrawing from the study ($n = 1$), failing to respond to more than 80% of trials ($n = 2$), or experiencing software failure during task completion ($n = 2$). Participant recruitment relied on adverts on the University of Bristol's psychology department website and flyers posted around the university. All participants were right-handed as assessed by the Edinburgh Handedness Inventory (Oldfield, 1971), had normal or corrected-to-normal vision, no history of neurological or neuropsychiatric disorders, were screened for any contraindications to MRI, and received £15 for their time. Written informed consent was obtained from all individuals and the study protocol was approved by the Human Research Ethics Committee of the University of Bristol's Faculty of Science.

2.1.2 Stimuli









Fifty-two fictional news reports were created for this study (all available at <https://osf.io/37rhs/>) based on previous research on the CIEM (see Johnson &

Seifert, 1994; Wilkes & Leatherbarrow, 1988). Half of the reports involved negative news (e.g., an aircraft evacuation), whereas the other half involved positive news (e.g., a new therapy for blindness). Positive and negative reports were equally distributed across all four experimental conditions as outlined below.

Each report consisted of two sentences (i.e., an initial pitch and a verification message) as well as an accompanying target image presented sequentially. The initial pitch described an event (e.g., the evacuation of an airplane) and a cause (e.g., due to a broken tailfin). The subsequent verification message described the same event, but either confirmed the original cause in different words (e.g., the rear rudder was damaged) or corrected it (e.g., the right engine broke off), creating so-called confirmation versus correction reports. The target image, finally, acted as a memory probe and portrayed information that either matched or mismatched the verification message.

This design had four experimental conditions (see Table 1): In condition A (ConMatch), the verification message confirmed the initial pitch, and the target image matched the verification message as well as the initial pitch. In condition B (ConMis), the verification message confirmed the initial pitch, but the target image mismatched both. In condition C (CorMatch), the verification message corrected the initial pitch, and the target image matched the correct verification message. Finally, in condition D (CorMis), the verification message corrected the initial pitch, but the target image mismatched the verification message, instead matching the initial incorrect pitch (i.e., only in this condition did the image directly refer to prior misinformation).

Table 1. Two example reports as used in Study 1 shown in their four possible versions, which were counterbalanced across participants.

Condition	Phase	Participant 1	Participant 2
A: Confirmation Report/ Matching Image		<i>Report 1wy</i>	<i>Report 1wz</i>
	Pitch	A flight was evacuated before take-off due to a broken tail fin.	A flight was evacuated before take-off due to a detached motor.
	Verification Message	A plane was evacuated after the rear rudder was damaged.	A plane was evacuated after the right engine broke off.
Target Image			
B: Confirmation Report/ Mismatching Image		<i>Report 2wy</i>	<i>Report 2wz</i>
	Pitch	A bridge near Copenhagen was consumed by flames.	A bridge near Copenhagen was washed away after heavy rain
	Verification Message	A bridge near Copenhagen was destroyed by fire.	A bridge near Copenhagen was destroyed by a flood.
Target Image			
C: Correction Report/ Matching Image		<i>Report 1xz</i>	<i>Report 1xy</i>
	Pitch	A flight was evacuated before take-off due to a detached motor.	A flight was evacuated before take-off due to a broken tail fin.
	Verification Message	A plane was evacuated after the rear rudder was damaged.	A plane was evacuated after the right engine broke off.
Target Image			
D: Correction Report/ Mismatching Image		<i>Report 2xz</i>	<i>Report 2xy</i>
	Pitch	A bridge near Copenhagen was washed away after heavy rain	A bridge near Copenhagen was consumed by flames.
	Verification Message	A bridge near Copenhagen was destroyed by fire.	A bridge near Copenhagen was destroyed by a flood.
Target Image			

In order to prepare parallel confirmation and correction reports for Study 1, two versions of each report were created that differed only in the initial pitch (but neither in the verification message, nor the target image; see versions w and x in Table 1). To further ensure that each pitch appeared equally often in the context of a correction report as well as in the context of a confirmation report, two additional versions of each report were created that shared the same pitch but differed in their verification messages and target images (see versions y and z in Table 1). By counterbalancing these scripts carefully across participants we ensured that no participant saw the exact same pitch twice during the experiment (i.e., each pitch was only shown once per participant, either followed by a correction or a confirmation message). Specifically, the different versions of each report (resulting in 208 unique stimuli in total) were counterbalanced across participants so that each individual encountered 104 reports throughout the task with each report being shown exactly twice – once in the context of a confirmation report and once in the context of a correction report (with the constraint that paired reports would never be presented sequentially).

2.1.3 fMRI task and procedure

The task was explained to participants upon arrival as follows: *“You are a senior editor at a daily newspaper. You have two major tasks: First, many journalists keep pitching stories to you. You are the one who must sort them into potentially positive or negative news, while your fact-checking department verifies their content. Second, once you get the actual true story from your fact-checking department, it is your responsibility to check that a photo (that someone from your photography*

department has proposed to go with the story) actually matches with the true story (yes vs. no) as verified by your fact-checking department.”

To ensure that participants understood what was required of them they were allowed to ask questions and required to complete four practice trials outside the fMRI scanner. Participants who expressed confusion were allowed to retake the practice trials until they felt that they understood all task requirements. Upon being placed in the scanner, MRI-compatible button response boxes (LU400, Cedrus Lumina) were placed in each of the participant's hands (i.e., buttons 1 and 2 in the left hand, buttons 3 and 4 in the right hand) to record their responses. Presentation of stimuli was controlled and responses were recorded using Psychtoolbox (version 3.0.8; Brainard, 1997) running in Matlab (version 2012a). All stimuli were centred on a uniform black screen. Written stimuli were presented in white font type 'sans serif' size 40; images were displayed with 400 × 400 pixel resolution. Stimuli were presented using rear-projection onto a screen that was visible to participants through a mirror attached to the head-coil.

Participants read 104 reports distributed across two functional runs of approximately 25 minutes each, with a short break in between during which a fieldmap and structural sequence were run. The order of reports, including which run they were shown in, was randomised for each participant. Each trial began with the prompt 'NEW PITCH'. After 500 ms, this prompt was replaced by the actual pitch, which remained on screen for 4500 ms. Participants were required to decide whether the pitch presented primarily positive or negative news by pressing one of four buttons (1 = very negative, 2 = slightly negative, 3 = slightly positive, 4 = very positive). Responses were recorded from the moment the pitch appeared on the screen until 2000 ms after it disappeared. In accordance with our pre-registration,

any responses outside this time window were marked as a non-response and excluded from the analysis. The pitch was then replaced by a fixation cross that remained on screen for a pseudo-random period of time that lasted between 4000 to 8000 ms (as determined by drawing a random number from within this interval). This temporal jitter was introduced to ensure that the haemodynamic response to each sentence or image could be modelled separately from the adjacent stimuli (Ollinger, Shulman, & Corbetta, 2001).

The alert 'VERIFIED VERSION' (shown for 500 ms) ultimately replaced the fixation cross in order to prepare participants for the verification message. This message then replaced the alert and stayed on screen for 4500 ms. Participants were not required to respond to the verification message and simply waited for it to be replaced by a fixation cross that remained on screen for a pseudorandom duration (again ranging from 4 to 8 seconds). Subsequently, the fixation cross was replaced by the alert 'PICTURE CHECK' (presented for 500 ms) to prepare participants for the presentation of the target image. This alert was then replaced by a target image that remained on screen for 1500 ms. Participants were required to decide with a button press whether the target matched the verification message. Pressing buttons 1 or 2 indicated a mismatching target, whereas pressing buttons 3 or 4 indicated a matching target. Responses to the picture were recorded from the point at which the image appeared on the screen until 2000 ms after offset. Any responses outside this time window were treated as inaccurate responses. A fixation cross ultimately replaced the target image and stayed on screen for a pseudorandom duration (ranging from 6000 to 12 000 ms) before the next trial was launched.

2.1.4 fMRI protocol

Data were acquired on a 3 Tesla Siemens Skyra MRI scanner with a 32 channel receive-only head coil at the Clinical Research and Imaging Centre of the University of Bristol. Memory foam was used to minimize head movement. Functional images were acquired using a whole-brain T2*-weighted gradient echo sequence: echo planar imaging (EPI), TE/TR = 30/2500 ms, flip angle = 90°, 3 × 3 mm in-plane resolution; field of view (FOV) = 192 mm, phase encoding anterior to posterior, parallel acceleration factor two in the phase-encoding direction, and reconstructed using the generalized autocalibrating partially parallel acquisitions (GRAPPA, Griswold et al., 2002) method. Each volume consisted of 36 axial slices aligned parallel to AC-PC line (anterior commissure – posterior commissure) with 3 mm slice thickness and 0 mm gap. For each subject, a high resolution (0.9 × 0.9 × 0.9 mm) T1-weighted 3D volume scan was acquired with the MP-RAGE sequence: slice thickness = 0.9mm; TE/TR = 2.25/1800 ms; flip angle = 9°, FOV = 240 mm. In order to correct for spatial distortion in EPI data, dual-echo gradient echo field-maps were acquired for each subject: slice thickness = 3.0 mm; resolution = 3 × 3 × 3 mm; TE1/TE2/TR = 4.92/7.38/520 ms; flip angle = 60°; FOV = 192 mm.

2.1.5 Behavioural Analysis

Behavioural analyses examined participants' replies to the target images and analysed them in terms of their accuracy rates (in %) and median reaction times on correct trials (in ms). Additionally, a single dependent measure for the image categorization task combining participants' accuracy rates and reaction times was created to account for a potential speed-accuracy trade-off in participants' replies (Garrett, 1922; Schouten & Bekker, 1967; Wickelgren, 1977). Specifically, a diffusion

model was used to integrate participants' accuracy rates and reaction times. This model assumes that, when presented with a binary choice, participants will accumulate evidence in favour of one or the other response until a certain threshold is reached that allows them to settle for one of the available response options. In consequence, such analyses simultaneously consider participants' response speed and accuracy rates in order to estimate the mean rate of evidence accumulation in the decision-making process. This mean rate of evidence accumulation is also known as drift rate. The lower the drift rate, the lower is the signal-to-noise ratio in the evidence-accumulation process. In the current study, each participant's drift rate was estimated using the 'EZ' diffusion model (Wagenmakers, van der Maas, & Grasman, 2007). All three dependent measures (i.e., accuracy rates, reaction times, drift rates) were submitted to a 2 (verification message: confirmation vs. correction) × 2 (target image: match vs. mismatch) repeated measures analysis of variance (ANOVA).

2.1.6 fMRI Pre-Processing

Image processing and statistical inference was performed using the FSL software (version 5.0.9; Oxford Centre for Functional MRI of the BRAIN; FMRIB; Smith et al., 2004). In preparation for the motion correction procedure, the `fsl_motion_outliers` command (set to the option `-fdrms`) was used to examine each participants frame-to-frame displacement (calculated as the average rotation and translation parameter differences between successive acquisition frames using matrix RMS formulation – see Jenkinson, 2003). Participants with a relative displacement larger than 3.0mm on at least two occasions of the same run were excluded from all fMRI analyses ($n = 8$). For all remaining participants, prior to model estimation, functional images were

distortion and motion corrected by pre-processing with FEAT (FMRIB's Expert Analysis Tool), which included spatial smoothing, motion correction using MCFLIRT (Jenkinson, Bannister, Brady & Smith, 2002; Jenkinson & Smith, 2001), and high-pass temporal filtering (cut-off 90 s). Although we initially pre-registered a spatial smoothing kernel of FWHM = 5 mm, we ultimately used FWHM = 6 mm to ensure that the level of smoothing was at least twice the size of our voxels (Friston, Holmes, Poline, Price, & Frith, 1996; Mikl et al., 2008).

To facilitate group-level analysis, the spatial transformation between the pre-processed EPI data and each subject's T1-weighted structural scan was determined using the boundary-based registration algorithm (BBR; Greve & Fischl, 2009) and FLIRT (FMRIB's Linear Image Registration Tool; Jenkinson et al., 2002; Jenkinson & Smith, 2001). To improve registration, brain tissue was segmented from structural scans by using an in-house brain extraction tool ("VBM8BET"), based on the output from VBM8 ("VBM at Structural Brain Mapping Group", n.d., <http://www.neuro.uni-jena.de/vbm/>). The final registration step included spatial normalisation of each subject's brain extracted T1-weighted structural scan to a "standard space" template brain (Montreal Neurological Institute [MNI] averaged-152 subject 2 mm template), achieved using an initial 12-parameter affine registration with FLIRT, followed by non-linear registration using FNIRT (FMRIB's non-linear image registration tool) with 5 mm warp spacing.

Parameter estimates for each explanatory variable of interest (EV) were calculated for correct trials using a general linear model (GLM) and canonical hemodynamic response function (HRF) implemented in FEAT, which incorporated pre-whitening with FILM (Woolrich, Ripley, Brady, & Smith, 2001). One model including all EVs as listed in Table 2 was created to assess main effects and planned

contrasts as described in further detail below (note that by doing so we slightly altered our original pre-registration, which – unnecessarily – suggested building two separate models for main effects and planned contrasts). All EV's were modelled with their respective duration. Besides these six regressors of interest, the subject-level model also included two nuisance regressors (identifying the presentation of the pitch message and any incorrect target image decisions) and six participant-specific motion parameters (as commonly implemented for task-based fMRI; Caballero-Gaudes & Reynolds, 2017). Following subject-level modelling, parameter estimate maps and associated variance images were transformed to standard space and input to a group-level mixed effects model, estimated using FLAME (FMRIB's Local Analysis of Mixed Effects).

Table 2. *Explanatory variables of interest (EV) included in the fMRI data model in Study 1*

Variable	Relevant Stimuli
EV1 (vmCon)	Onsets of verification messages acting as confirmations
EV2 (vmCor)	Onsets of verification messages acting as corrections
EV3 (tiConMatch)	Onsets of matching target images following confirmations
EV4 (tiConMis)	Onsets of mismatching target images following confirmations
EV5 (tiCorMatch)	Onsets of matching target images following corrections
EV6 (tiCorMis)	Onsets of mismatching target images following corrections

2.1.7 fMRI Analyses

Two types of analyses were pre-registered for this study. Analysis 1 examined whole-brain contrasts and parametric analyses (i.e., brain-behavior correlations), whereas Analysis 2 repeated the same two types of analyses within preregistered regions-of-interest (ROIs). For the sake of clarity, this manuscript reports the former here and the latter in the Supplementary Material. Adopting a whole-brain analysis

approach, we first examined a series of principal contrasts (see Table 3).

Corresponding contrast maps were computed for each of the two functional runs at the subject level. Within each subject, data from the two separate functional runs were combined using a fixed-effects model. These maps were then entered into a group-level mixed-effects model estimated using FLAME. Statistical inference was performed using Gaussian random field theory (Worsley, Evans, Marrett, & Neelin, 1992).

Based on FSL's default analysis setting, we originally pre-registered a liberal cluster-forming threshold of $Z > 2.30$ with a cluster-significance threshold of $p < 0.05$ (FWE corrected). However, following recent recommendations (Eklund, Nichols, & Knutsson, 2016; Kessler, Angstadt, & Sripada, 2017; Nichols, Eklund, & Knutsson, 2017), we ultimately applied a more conservative threshold of $Z > 3.09$ with a cluster-significance threshold of $p < 0.05$ (FWE corrected). All activated clusters were interrogated using AUTOAQ for automated anatomical labelling (Winkler, 2012) and a threshold of $> 10\%$ probability was applied to determine relevant regions within each cluster. If all regions in a cluster failed to reach this threshold, the region with the highest probability was reported instead.

Table 3. *Pre-registered whole-brain contrasts as computed in the fMRI analysis*

Contrast	Effect of...	Weights of Explanatory Variables
<i>Main Effects</i>		
[A & B] >/< [C & D]	Verification message	[1 -1 0 0 0 0] and [-1 1 0 0 0 0]
[A & B] >/< [C & D]	Target image	[0 0 1 1 -1 -1] and [0 0 -1 -1 1 1]
<i>Planned Contrasts</i>		
A >/< C	Target image	[0 0 1 0 -1 0] and [0 0 -1 0 1 0]
B >/< D	Target image	[0 0 0 1 0 -1] and [0 0 0 -1 0 1]

Adopting a whole-brain analysis approach, we further explored whether brain activity captured during the target image categorization task was associated with participants' reaction times on this task. Thus, participants' reaction times were included as trial-by-trial parametric modulators of the four relevant EVs modelling neural activity during presentation of the target image. Corresponding statistical maps were computed for each participant and entered into a second-level one-sample *t*-test for each condition, treating participants as a random effect. Significance testing was performed in the same manner as for the whole-brain analyses. Finally, in order to be able to fully understand differences in parametric activity across experimental conditions, these maps were also entered into a group-level mixed-effects model.

2.2 Results

2.2.1 Behavioural Data (available at <https://osf.io/dpknj/>)

In preparation for examining participants' image categorization accuracy rates, reaction times, and drift rates on the main task (see Table 4), a brief preliminary analysis was run. This analysis ensured that participants had paid sufficient attention to the original news pitches as included in the task. It involved correlating each participant's pitch-related valence ratings with the mean valence ratings for all pitches as provided by the remaining participants. No participant's correlation coefficient fell more than 3 standard deviations below the sample's mean coefficient ($r = .89$, $SD = 0.03$, $\min r = .83$, $\max r = .94$), indicating that the pitch messages were adequately processed by all participants. Thus, we proceeded with our main analyses.

These analyses revealed that participants' accuracy rates showed no significant main effects [verification message: $F(1,31) = 3.46, p = .072$; target image: $F(1,31) = 1.89, p = .179$], but a significant verification message \times target image interaction [$F(1,31) = 6.02, p = .020$]. The interaction signalled that participants were equally accurate at judging the suitability of matching and mismatching target images whenever these followed a confirmation message [$t(31) = 1.29, p = .208$], but were less accurate at judging the suitability of mismatching compared to matching target images when these images followed a correction message [$t(31) = 2.23, p = .033$]. Moreover, while accuracy rates for matching target images were unaffected by the content of the verification message [$t(31) = 0.42, p = .675$], accuracy rates for mismatching target images were significantly lower following a correction than a confirmation message [$t(31) = 2.87, p = .007$].

Table 4. Means and standard deviations for the three behavioural variables captured during target image categorisation in each of the four experimental conditions in Study 1

Experimental Condition	Accuracy		Reaction time		Drift rate	
	Mean (%)	SD	Mean (ms)	SD	Mean	SD
A (ConMatch)	83.05	10.09	1566	332	0.096	0.036
B (ConMis)	84.86	7.93	1611	344	0.104	0.033
C (CorMatch)	83.89	8.49	1627	357	0.099	0.037
D (CorMis)	78.37	13.46	1709	464	0.077	0.042

In addition, participants' median RTs on correct trials yielded a significant main effect of verification message [$F(1,31) = 6.60, p = .015$], but no main effect of target image [$F(1,31) = 2.48, p = .126$] and no interaction effect [$F(1,31) = 0.88, p = .356$]. Image categorization times were overall faster following a confirmation message ($M = 1588, SD = 321$ ms) than a correction message ($M = 1668, SD = 388$

ms). Finally, participants' drift rates on correct trials revealed a significant main effect of verification message [$F(1,31) = 4.77, p = .037$] that was qualified by a significant verification message \times target image interaction [$F(1,31) = 12.73, p = .001$]. The analysis yielded no significant main effect of target image [$F(1,31) = 2.13, p = .154$]. Follow-up pairwise comparisons revealed that drift rates were equivalent for matching and mismatching images that followed a confirmation message [$t(31) = 1.45, p = .157$], but were significantly reduced whenever mismatching rather than matching images followed a correction message [$t(31) = 3.03, p = .005$]. Also, while drift rates for matching images were unaffected by the content of the verification message [$t(31) = 0.45, p = .653$], drift rates for mismatching images were significantly lower following a correction than a confirmation message [$t(31) = 3.90, p < .001$]. In summary, participants' image suitability judgments were most compromised (as reflected in systematically reduced accuracy and drift rates) for images that followed a correction report and referred directly to prior misinformation.

2.2.2 Whole-Brain Contrasts

Two univariate whole-brain contrasts examined the main effect of verification message at encoding (i.e., confirmation message > correction message, correction message > confirmation message), but failed to detect any suprathreshold activation. Two further contrasts examined the main effect of verification message at information retrieval (i.e., during the processing of target images; see Table 5). It was found that processing target images following confirmation messages > correction messages returned greater activity in the left lingual gyrus (LG), whereas the reverse contrast identified enhanced activity in the right angular gyrus (AG; see Figure 1A).

To further scrutinize the effects of verification message on target image processing (while ensuring equivalent participant responses), additional planned contrasts were computed separately for accurately *accepted* and *rejected* target images (see Table 6). We first compared the successful acceptance of matching target images following the presentation of confirmation versus correction messages (i.e., conditions A vs. C), but no suprathreshold activation emerged (regardless of whether we contrasted $A > C$ or $C > A$). We then compared the successful rejection of mismatching target images following the presentation of confirmation versus correction messages (i.e., conditions B vs. D). Although the processing of mismatching target images following confirmation $>$ correction messages returned no suprathreshold activation, the reverse contrast yielded enhanced activity bilaterally in the Precuneus (PrC) and the AG (see Figure 1B). Phrased differently, rejecting target images that directly referred to prior misinformation elicited enhanced activity in two distinct brain regions. Finally, in order to determine whether this activity difference was truly specific for rejection trials, we also compared the two planned contrasts directly [$D > B$] vs. [$C > A$], but this comparison returned no suprathreshold activation.

Table 5. *Peak voxel in MNI coordinates and number of voxels for brain regions that responded differently to the same target images depending on whether they followed a confirmation message or a correction message.*

Region	Hemisphere	Voxels	Max z-value	x	y	z
<i>Target Images following a Confirmation Message > Correction Message</i>						
Lingual Gyrus (extending into the posterior temporal cortex)	L	230	4.24	-28	-44	-10
<i>Target Images following a Correction Message > Confirmation Message</i>						
Angular gyrus (extending into the supramarginal gyrus)	R	338	4.14	52	-42	38

NB: Results identified by a series of whole brain contrasts at a cluster-forming threshold of $Z > 3.09$ and $p < 0.05$ (FWE-corrected).

Table 6. *Peak voxel in MNI coordinates and number of voxels for brain regions that responded differently to the same mismatching target images depending on whether they followed a confirmation or a correction message.*

Region	Hemisphere	Voxels	Max z-value	x	y	z
<i>Target Images following a Confirmation > Correction</i>						
No suprathreshold activation						
<i>Target Images following a Correction > Confirmation</i>						
Angular gyrus (extending into the supramarginal gyrus)	R	786	4.22	46	-58	44
	L	525	4.12	-44	-56	42
Precuneus	R/L	454	4.04	10	-70	38

NB: Results identified by a series of whole brain analyses at a cluster-forming threshold of $Z > 3.09$ and $p < 0.05$ (FWE-corrected).

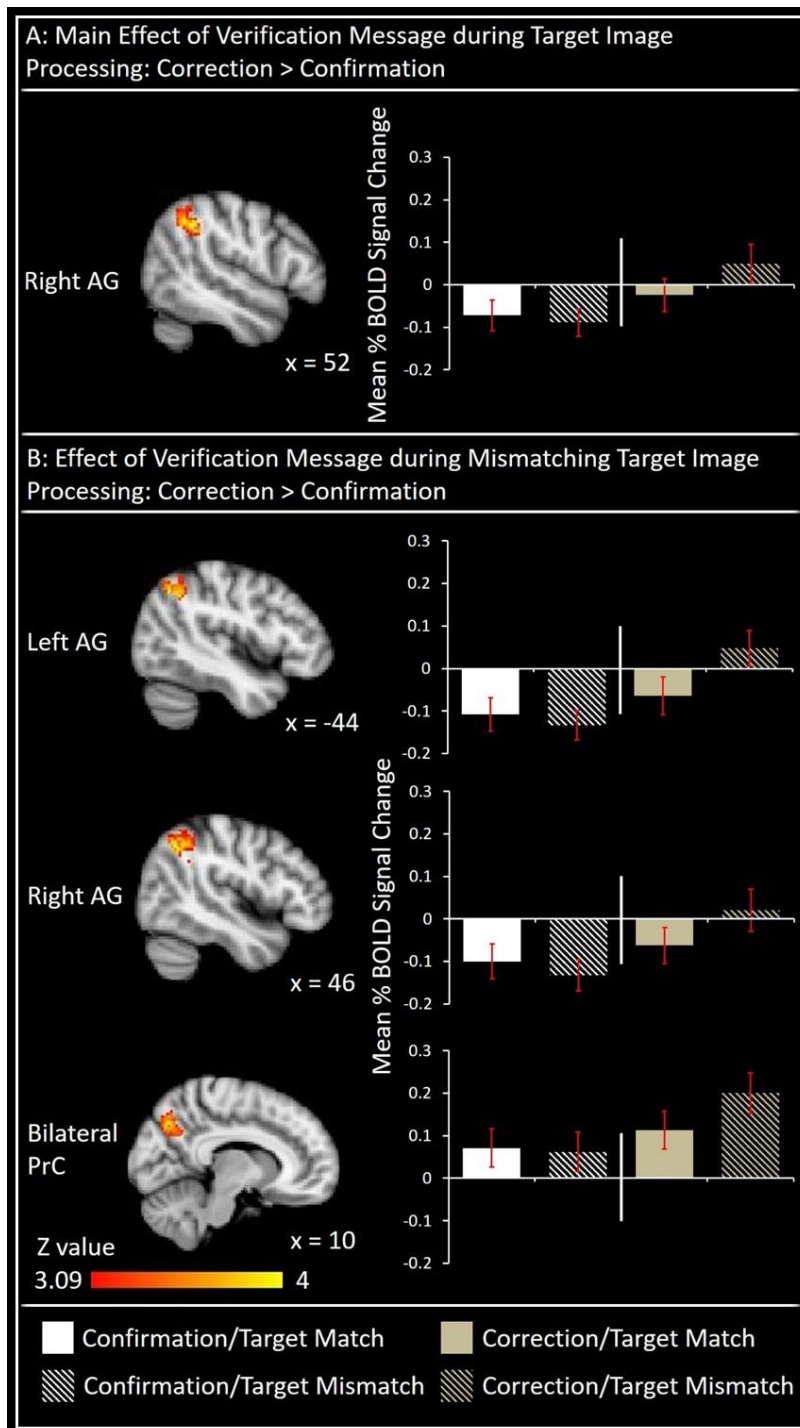


Figure 1: Whole-brain contrasts revealed enhanced activity in the right angular gyrus (AG) in response to target images that followed correction messages rather than confirmation messages (see panel A). Enhanced activity in partially overlapping locations was also observed when the contrast was limited to the processing of mismatching target images (see Panel B). By contrast, no suprathreshold activation emerged when the same contrast was computed for matching target images. To illustrate these effects we plotted the mean % signal change (mean \pm standard error bars) in the respective brain regions for all conditions. The figure shows group data from 24 participants displayed on sagittal slices of the MNI 2mm brain template. Brain regions were identified based on paired *t*-tests (mixed effects models) with cluster-forming thresholds of $Z > 3.09$ and $p < 0.05$ (FWE-corrected).

2.2.3 Whole-Brain Correlations

A series of parametric analyses linked participants' reaction times on the image categorization task to their task-related brain activity. When analysed separately, significant (positive and/or negative) correlations were observed for each of the four experimental conditions (see Table S1 in the Supplementary Material). But comparing these correlations for correction and confirmation messages separately for matching target images (A vs. C) and for mismatching target images (B vs. D) failed to return significant differences. Similarly, comparing correlations by verification message irrespective of image type (A & B vs. C & D) failed to return significant results, indicating that correlation patterns did not differ systematically across the four experimental conditions.

2.3 Interim Discussion

Study 1 captured a well-known behavioural signature of the CIEM: Upon processing news reports which featured a correction (rather than a confirmation), participants' ability to reject subsequently presented target images that referred to prior misinformation was compromised. Beyond demonstrating this behavioural effect, Study 1 also explored participants' neural activity during task completion. Doing so revealed that the exact same memory probes (i.e., target images) elicited enhanced activity in the right AG when they followed correction reports compared to confirmation reports. In addition, it revealed that the successful rejection of images with reference to prior misinformation elicited enhanced activity in the bilateral AG and PrC. Despite these interesting findings, Study 1 had several important limitations.

First and foremost, the study's most important finding (i.e., enhanced bilateral AG and PrC activity in response to misinformation-related target images following corrections) remained inconclusive as it was uncertain whether it applied only to memory probes that referred to prior misinformation or generalized even to memory probes without such reference. Though the outcomes of two planned contrasts favoured the first interpretation, the corresponding interaction effect (i.e., comparing both planned contrasts directly) failed to reach statistical significance. Furthermore, the study produced a series of unexpected null findings. Specifically, it failed to capture any neural activity differences for the encoding of correction messages. It also failed to return any condition-specific brain-behaviour correlations.

Unfortunately, the obtained null-findings may have simply reflected a suboptimal study design. For instance, participants were asked to encode each verification message twice throughout the task (i.e., once as a confirmation message and once as a correction message). Accordingly, these messages may have inadvertently been perceived as correcting (or confirming) information from a prior trial and, thus, as inherently ambiguous. Additionally, the absence of condition-specific brain-behaviour correlations may have simply reflected limited statistical power due to a relatively small number of trials per condition (cf., Liu, Frank, Wong, & Buxton, 2001) caused by pragmatic concerns about the study's overall length. Thus, with these limitations in mind we designed an improved follow up study.

Our second study focused on advancing our understanding of encoding- and retrieval-related differences in brain activity during the processing of misinformation. Thus, it examined the processing of misinformation-containing memory probes in further detail in order to overcome the interpretational ambiguity of Study 1. In addition, it monitored the encoding of misinformation corrections in a manner that did

not involve repeating the exact same verification messages for the same participant. By contrast, it refrained from exploring brain-behaviour correlations, acknowledging that the current paradigm does not lend itself well to the necessary increase in trial numbers without jeopardizing participants' ability to complete the task.

3. Study 2

In Study 2, participants were again required to read a series of news reports that contained corrective or confirmatory information. This time, however, participants were asked to subsequently choose from a pair of images the one that best matched the report's verified facts. By using image pairs rather than single images as memory probes in Study 2, we were able to present on each trial one image that matched the verified facts and one distractor image. This distractor image either referred to prior misinformation or not. Based on our results from Study 1, we predicted that enhanced activity in the AG and PrC would be most pronounced in response to image pairs that followed correction reports and contained explicit references to prior misinformation in the distractor image. The method and an analysis plan for this study were pre-registered at <https://osf.io/pwr72/>.

3.1 Methods

3.1.1 Participants

30 participants (22 female) aged 18-32 years ($M = 21.3$, $SD = 2.77$), that had not taken part in experiment 1, took part in this study. Three of them (3 females) failed to pass the head motion check for fMRI data (as applied in Study 1) so that only their behavioural data (but not their fMRI data) were considered during data analysis.

Three additional participants were recruited but excluded from all analyses due to

software failure during task completion ($n = 1$) or failure to respond to more than 25% of target image decisions ($n = 2$). Advertisement, recruitment, screening criteria, and subject reimbursement were equivalent to Study 1. Written informed consent was obtained from all individuals. The study protocol was approved by the Human Research Ethics Committee of the University of Bristol's Faculty of Science.

3.1.2 Stimuli

The positive and negative fictional news reports from Study 1 were adapted for Study 2 (all stimuli are available at <https://osf.io/3kzsd/>). The adapted reports consisted of an initial news pitch followed by a verification message and an image pair as memory probe. Once again, positive and negative reports were equally distributed across all experimental conditions. For each report, three different pitch sentences were created. Each version gave a different cause for the same event (e.g., 'A bridge near Copenhagen was consumed by flames' versus 'A bridge near Copenhagen was washed away after heavy rain' versus 'A bridge near Copenhagen collapsed in high winds'). The subsequent verification message was identical for all three versions of each report and described one of the original three causes in different words (e.g., 'A bridge near Copenhagen was destroyed by a fire'). Thus, the verification message either confirmed or corrected the original pitch.

In addition, the image pair that acted as the memory probe was identical across all three versions of each report. Each image pair contained one image that matched the actual verification message (i.e., an image of a bridge on fire) and one distractor image that showed one of the two alternative causes for the event (i.e., a bridge washed away by a flood or a bridge blown over by wind). Thus, depending on the initial pitch that a participant had seen, the exact same distractor image either

referred to prior misinformation (i.e., to a cause that was subsequently corrected) or to information that had not previously been discussed.

This experimental approach resulted in three experimental conditions: confirmation reports followed by neutral image pairs without misinformation (condition CONF_NEU), correction reports followed by neutral image pairs without misinformation (condition CORR_NEU), and correction reports followed by image pairs with misinformation (condition CORR_MIS). Importantly, the last condition closely resembled condition D in Study 1, but this time the participant response was kept constant across all experimental conditions (i.e., participants selected the correct image rather than indicating whether a single image was a match or a mismatch). Furthermore, by counterbalancing the presentation of all reports across participants, all reports (including their verification messages) were only seen once per participant throughout the task. Thus, three counterbalanced versions of the task were prepared and administered to one third of our participants.

3.1.3 fMRI Procedure

Participants were again instructed to imagine working as a senior newspaper editor, who was expected to judge each incoming story's valence and to choose a suitable image that would be published with the verified story. To ensure that participants understood what was required of them, they completed at least four practice trials outside the fMRI scanner. Upon being placed in the scanner, MRI-compatible button response boxes (LU400, Cedrus Lumina) were placed in each of the participant's hands (i.e., buttons 1 and 2 in the left hand, buttons 3 and 4 in the right hand) to record their responses. Throughout the actual fMRI task, participants were presented with 78 reports distributed across two runs of approximately 22 minutes each, with a

short break in between during which the fieldmap and T1-weighted structural scans were acquired. The order of all reports, including which run they appeared in, was randomized for each participant.

The order and timings of the different trial components remained the same as in Study 1. In addition, the presentation of all target image pairs was carefully controlled to ensure that the matching image appeared equally often on the left or the right side of the screen. Visual stimulus presentation methods and parameters were identical to Study 1, but this time target images were displayed with 400 × 400 pixel resolution, with a 100 pixel gap between the two images. During valence judgments and image selection, participants logged their responses by pressing one of four buttons on the button boxes using the index and middle fingers of their right and left hand (valence: 1 = very negative, 2 = negative, 3 = positive, 4 = very positive; image selection: 1/2 = left image, 3/4 = right image).

3.1.4 fMRI protocol, pre-processing, and analysis

The fMRI data acquisition protocol, pre-processing, and analysis was equivalent to Study 1. Parameter estimates for each relevant EV on correct trials were again calculated using a GLM as implemented in FEAT (see Table 7). As in Study 1, the subject-level model included all EVs (modelled with their respective durations), two nuisance regressors (identifying the presentation of the pitch message and any incorrect image pair decisions) and six participant-specific motion parameters. Following subject-level modelling, parameter estimate maps and associated variance images were transformed to standard space and input to a group-level mixed effects model, estimated using FLAME.

Table 7. *Explanatory variables of interest (EV) included in the fMRI data model in Study 2*

Variable	Relevant Stimuli
EV1 (vmConf)	Onsets of verification messages acting as confirmations
EV2 (vmCorr)	Onsets of verification messages acting as corrections
EV3 (tipConf_Neu)	Onsets of neutral target image pairs following confirmations
EV4 (tipCorr_Neu)	Onsets of neutral target image pairs following corrections
EV5 (tipCorr_Mis)	Onsets of target image pairs with misinformation following corrections

First, we contrasted participants' neural activity during encoding, that is, during the processing of verification messages (i.e., confirmations vs. corrections). To facilitate this analysis, verification messages that acted as corrections were pooled together in one EV irrespective of the image pair that followed them. Second, we contrasted participants' neural activity during information retrieval, that is, during image pair processing depending on which verification message preceded the image pair (confirmations vs. corrections). Third, we examined potential interaction effects between type of verification image and distractor image content (with or without misinformation) during image pair processing by contrasting all three experimental conditions with each other (i.e., CONF_NEU vs. CORR_NEU; CONF_NEU vs. CORR_MIS; and CORR_NEU vs. CORR_MIS).

Given the absence of significant brain-behaviour correlations in Study 1, we refrained from preregistering (and running) corresponding correlational analyses for Study 2. Instead, our pre-registered analyses of Study 2 were limited to whole-brain contrasts (analysis 1) as well as ROI-based contrasts (analysis 2). For the sake of clarity, this manuscript reports the findings of the whole-brain contrasts below and the outcomes of the ROI-based analyses in the Supplementary Material. As in Study

1, whole-brain contrasts were thresholded at $Z > 3.09$ with a cluster-significance threshold of $p < 0.05$ (FWE corrected).

3.2 Results

3.2.1 Behavioural Data (available at <https://osf.io/wby7n/>)

Each participant's pitch-related valence ratings were again correlated with the mean valence ratings for all pitches as provided by the remaining participants. As in Study 1, no participant's correlation coefficient fell more than 3 standard deviations below the sample's mean coefficient ($r = .87$, $SD = .05$, min $r = .72$, max $r = .93$), signalling that pitch messages were adequately processed by all participants. Subsequent behavioural analyses examined participants' replies on the image selection task in terms of their accuracy rates, median reaction times on correct trials, and drift rates.

In line with our pre-registration, all three measures were initially inspected by submitting each of them to a one-way repeated measures ANOVAs with three levels (CONF_NEU, CORR_NEU, CORR_MIS; see Table 8). Doing so returned no significant effect of experimental condition for accuracy rates [$F(2,58) = 1.92$, $p = .156$] or reaction times [$F(2,58) = 1.55$, $p = .221$], but a marginally significant effect on drift rates [$F(2,58) = 2.67$, $p = .077$]. To explore the latter in further detail, we conducted an additional (i.e., non-preregistered) series of pairwise comparisons. It was found that participants' drift rates were significantly lower in the CORR_MIS condition than in the CONF_NEU condition [$t(29) = 2.08$, $p = .047$]. The remaining pairwise comparisons returned no significant results [i.e., CONF_NEU vs. CORR_NEU: $t(29) = 1.69$, $p = .102$; CORR_NEU vs. CORR_MIS: $t(29) = 0.53$, $p = .603$].

Table 8. Means and standard deviations for the three behavioural variables captured during target image selection in each of the three experimental conditions in Study 2

Experimental Condition	Accuracy Rates		Reaction Times		Drift Rates	
	Mean in %	SD	Mean in ms	SD	Mean	SD
CONF_NEU	89.49	5.53	1635	266	0.133	0.029
CORR_NEU	88.21	5.62	1576	266	0.121	0.026
CORR_MIS	86.54	8.19	1607	261	0.118	0.036

NB: CONF_NEU = confirmation reports followed by neutral image pairs without misinformation, CORR_NEU = correction reports followed by neutral image pairs without misinformation, CORR_MIS = correction reports followed by image pairs with misinformation

3.2.2 fMRI Data

Comparing participants' neural activity during verification message processing (i.e., during encoding) returned no suprathreshold activation, regardless of whether we contrasted confirmation > correction messages or correction > confirmation messages. Similarly, comparing participants' neural activity during image pair processing (i.e., during information retrieval) returned no suprathreshold activation, regardless whether image pairs followed confirmation > correction messages or correction > confirmation messages.

Suprathreshold activation was observed, however, when neural activity during image pair processing was compared keeping in mind both the image pair's preceding verification message and the image pair's type of distractor image (see Table 9). Specifically, contrasting CORR_MIS > CONF_NEU found enhanced activity in the left AG, the bilateral PrC, and the bilateral posterior cingulate cortex (PCC). The reverse contrast revealed no suprathreshold activation. Similarly, the contrast CORR_MIS > CORR_NEU also returned enhanced activity in the left AG and the bilateral PrC (see Figure 2), but the reverse contrast returned no suprathreshold activation. Finally, contrasting CONF_NEU > CORR_NEU revealed

enhanced activation in the right occipital cortex, whereas the reverse contrast returned no suprathreshold activation.

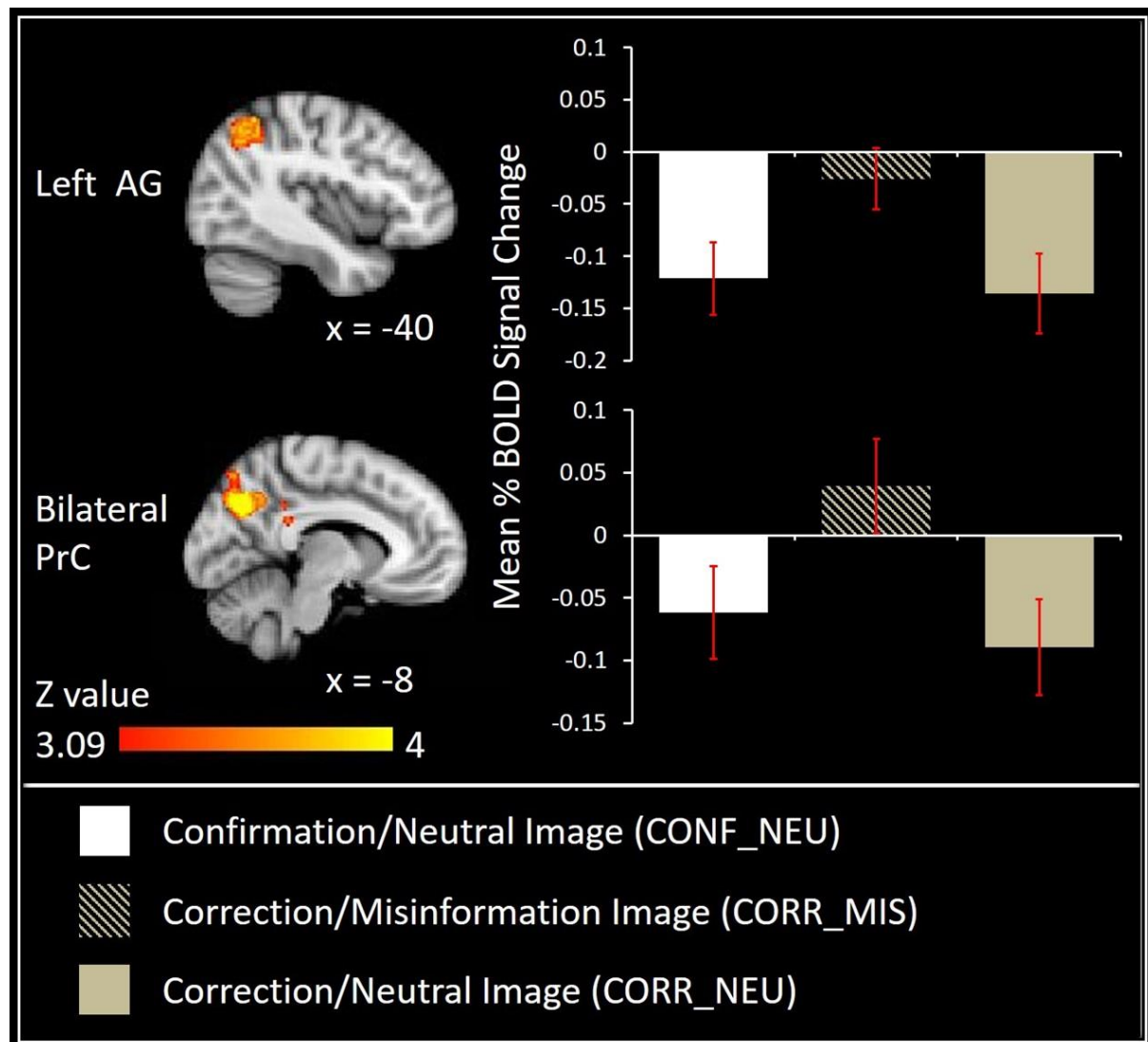


Figure 2: Whole-brain contrasts revealed enhanced activity in the left angular gyrus (AG) and the bilateral precuneus (PrC) in response to target image pairs that referred to prior misinformation. To illustrate this effect we plotted the mean % signal change (mean \pm standard error bars) in the respective brain regions for all conditions. The figure shows group data from 27 participants displayed on sagittal slices of the MNI 2mm brain template. Brain regions were identified based on paired t -tests (mixed effects models) with cluster-forming thresholds of $Z > 3.09$ and $p < 0.05$ (FWE-corrected).

Table 9. *Peak voxel in MNI coordinates and number of voxels for brain regions that showed differential activity across the three experimental conditions during target image selection in Study 2.*

Region	Hemi- sphere	Voxels	Max z- value	x	y	z
<i>CONF_NEU > CORR_NEU</i>						
Occipital cortex (extending into middle temporal gyrus)	R	199	4.04	56	-60	4
<i>CORR_NEU > CONF_NEU</i>						
No suprathreshold activation						
<i>CONF_NEU > CORR_MIS</i>						
No suprathreshold activation						
<i>CORR_MIS > CONF_NEU</i>						
Angular gyrus (extending into the Superior Occipital Gyrus)	L	285	4.16	-42	-60	44
Posterior Cingulate Cortex	R/L	188	4.25	0	-22	40
Precuneus	R/L	857	4.91	-4	-64	32
<i>CORR_NEU > CORR_MIS</i>						
No suprathreshold activation						
<i>CORR_MIS > CORR_NEU</i>						
Angular gyrus (extending into the superior occipital gyrus)	L	485	3.99	-40	-54	52
Precuneus (extending into the Posterior Cingulate Cortex)	R/L	2309	4.55	-6	-68	34

NB: *Results identified by a series of whole brain analyses at a cluster-forming threshold of $Z > 3.09$ and $p < 0.05$ (FWE-corrected).*

3.3 Interim Discussion

Study 2 re-established that participants' ability to process memory probes was uniquely compromised (as reflected in their drift rates) when these probes followed a correction report and contained prior misinformation. This (marginally significant)

behavioural effect was accompanied by a corresponding alteration in brain activity: Brain activity in the left AG and the bilateral PrC was systematically enhanced for memory probes that followed a correction and contained prior misinformation, compared to memory probes that followed a correction but lacked misinformation as well as compared to memory probes that followed a confirmation. But just as Study 1, Study 2 failed to observe any differences in brain activity during the encoding of corrections and confirmations.

4. General Discussion

The CIEM captures the observation that even upon receiving a correction, initially presented misinformation often succeeds at influencing people's subsequent reasoning (for reviews see Lewandowsky et al., 2012; Schwarz et al., 2016). To better understand the cognitive origins of the effect, we probed the CIEM across two studies using both behavioural and neuroimaging measures. In both studies, participants were required to identify target images that matched verified facts which either confirmed or corrected an initial news report. Across both studies, participants' drift rates signalled that processing news reports featuring a correction rather than a confirmation compromised participants' ability to evaluate subsequent memory probes that directly referred to prior misinformation. Importantly, these findings lend further support for the CIEM's robustness and replicability as highlighted by a recent meta-analysis (cf., Chan, Jones, Jamieson, & Albarracin, 2017), considering that both our studies involved slight adjustments to the original CIEM paradigm in order to ensure fMRI compatibility (e.g., increase of usual trial numbers).

Both studies also recorded brain activity during task completion. This additional measure was considered particularly important in order to differentiate

between two competing accounts of the CIEM. According to the so-called model-updating account, difficulties during the encoding of correcting information primarily give rise to the CIEM. By contrast, according to the concurrent-storage account, the effect largely reflects a failure in selective memory retrieval. Thus, in order to study both accounts simultaneously, participants' brain activity was continuously monitored during both the encoding of correcting information as well as the retrieval of correct information.

Unexpectedly, contrary to prior reports in the literature (e.g., Gordon et al., 2017), neither of our two studies found systematic neural processing differences during the encoding of correcting information in support of the model-updating account of the CIEM. Yet these null findings should be tentatively received. On the one hand, they may simply reflect issues of statistical power, given that both studies relied on a relatively small sample of 20 to 30 participants (see also Poldrack et al., 2017). On the other hand, these null findings may be due to the fact that receiving a relatively large number of correct as well as correcting pieces of information in a very short period of time could alter how this information is received more generally (e.g., with a heightened sense of caution irrespective of trial type). Further research is therefore needed to explore if and how the encoding of corrective information may differ from the encoding of non-corrective (e.g., original or confirmative) information in the human brain and whether neuroimaging data could lend support for (or against) the model-updating account of the CIEM (Ecker et al., 2010; Kendeou et al., 2014; Verschueren et al., 2005). In this context, it may be particularly worthwhile to examine how manipulating encoding strength (e.g., as shallow or deep) affects the neural correlates of the CIEM (cf., Ecker et al., 2011).

At present, the current data provide strong support for the concurrent-storage hypothesis of the CIEM (e.g., Ayers & Reder, 1998; Catarino et al., 2015; Ecker et al., 2011). Across both studies, it was found that memory probes that explicitly referred to prior misinformation elicited enhanced activity in the left AG and the bilateral PrC. In fact, the peak coordinates of both regions were highly similar across studies, regardless of whether participants' memory was probed using single targets (Study 1) or two-alternative targets (Study 2). Although the specificity of the effect remained uncertain in Study 1, it was unambiguously established in Study 2: Both regions did not simply respond to any type of memory probe that followed a correction report, but specifically responded to memory probes that referred to previously presented misinformation. That is, activity in the AG and PrC was enhanced when people had to reject information that was initially thought to be true but that was subsequently corrected. This result provides support for a concurrent-storage account of the CIEM, considering that only the retention of misinformation in memory can explain the differences between memory probes with and without direct reference to prior misinformation following correction reports.

Based on existing fMRI findings on selective memory retrieval (e.g., Benoit & Anderson, 2012; Depue, 2012; Levy & Anderson, 2012) and memory intrusion resistance (Nee et al., 2013), we did not predict either region to respond in a differential manner during our task (as reflected in the absence of both regions from our pre-registered ROI-based analyses). Nevertheless, both regions have been associated with memory-related processes in the past (for relevant reviews see Cavanna & Trimble, 2006; Seghier, 2013). Therefore, we would like to briefly speculate on each region's potential role in processing misinformation-containing memory probes based on the available literature. However, these speculations

certainly require further systematic investigation in order to overcome their current status of mere reverse inferences (cf. Poldrack, 2011).

Prior research on the AG, for instance, has linked increased activity in this region to episodic memory retrieval, especially when events are recognised with high confidence (Cabeza, 2008; Cabeza, Ciaramelli, Olson, & Moscovitch, 2008; Vilberg & Rugg, 2008). Considering that our analyses focused exclusively on trials in which participants managed to resist misleading memory probes (i.e., by accurately rejecting them in Study 1 or refraining from selecting them in Study 2), the AG's response in our studies may potentially signal that participants felt confident in distinguishing correct from false information on these trials. Alternatively, the region's involvement in various aspects of conceptual integration (Binder, Desai, Graves, & Conant, 2009; Seghier, 2013) and/or resolving conceptual ambiguity (Nee, Wager, & Jonides, 2007; Nieuwland, Petersson, & van Berkum, 2007; Ye & Zhou, 2009) could mean that participants re-assessed conceptual relations between different pieces of memorized information upon re-encountering one seemingly irrelevant piece in a memory probe.

Similarly, previous research has revealed that increased activity in the PrC tends to facilitate accurate memory retrieval (Bonni et al., 2015; Cabeza & Nyberg, 2000; Henson, Hornberger, & Rugg, 2005). Enhanced PrC activity in response to misleading memory probes in the current study may thus signal increased efforts to monitor memory contents that have been recognized as inaccurate. Furthermore, this region is also well-known to play a fundamental role in the retrieval of contextual associations in episodic memories (Fletcher et al., 1995; Grol, Vingerhoets, & De Raedt, 2017; Lundstrom et al., 2003, 2005; Shallice et al., 1994). Therefore, the observed PrC activity may also signal that images with direct references to prior

misinformation necessitated an increased requirement for discrimination between competing representations stored in memory.

However, given that both the AG and the PrC are known to contribute to a wide variety of cognitive tasks (for relevant reviews see Cavanna & Trimble, 2006; Seghier, 2013), a full understanding of their response to misinformation-containing memory probes requires future investigations that overcome the following limitations of the current work: Though both our paradigms succeeded at capturing some instances of participants “falling for” misinformation, they did not record enough instances to determine whether the left AG and PrC respond equally strongly towards misinformation-containing memory probes that result, or fail to result, in inaccurate decision making. Hence, it remains to be established whether the regions’ responses primarily reflect cognitive conflict caused by encountering misinformation-containing memory probes or cognitive processes involved in overcoming such conflict.

Furthermore, by using entirely fictional news reports, the current work did not yet address how participants’ idiosyncratic worldviews can affect the neural correlates of the CIEM. Considering that prior beliefs can strengthen the CIEM (e.g., Ecker, Lewandowsky, Fenton, & Martin, 2014; Wood & Porter, 2017), however, the neural correlates of encoding and retrieving worldview-congruent or –incongruent corrections deserve further investigation. Equally deserving of future inquiry is the time scale at which the CIEM takes place. In the current study, corrections were provided and memories probed right after participants had received new (mis-)information. Whilst the fact that the CIEM can occur even after such short intervals is noteworthy, this approach does not advance our understanding of how the CIEM unfolds over longer timeframes (see Schwarz et al., 2007). If (mis-)information is

received over several consecutive days, for instance, brain regions involved in sleep-dependent memory consolidation and retrieval may contribute particularly strongly to the effect (e.g., the hippocampus; Born & Wilhelm, 2012; Marshall & Born, 2007).

In conclusion, the prevalence of online misinformation has been declared a growing and significant challenge in the modern world (WEF, 2013). Acknowledging this challenge, the current paper aimed to explore in further detail how the human mind handles corrections of misinformation. Across two studies, we observed that receiving corrections of prior misinformation resulted in neural activity indicative of the concurrent storage of correct and corrected information in people's memory. These data support the view that integrating corrective information into one's existing body of knowledge is inherently difficult because previously stored false information is not easily removed. Instead, receiving corrections often leaves people with competing memory traces for the same event, a circumstance that makes them particularly susceptible to inaccurate reasoning and decision making.

Acknowledgements

This research was made possible through University of Bristol internal funds, a research grant from the Australian Research Council (DP160103596) awarded to Ullrich Ecker and Stephan Lewandowsky, and funding from the Royal Society and Psychonomic Society awarded to Stephan Lewandowsky. This research was supported by RCUK funding from the EPSRC.

Competing interests

The authors have no competing interests to declare.

References

- Anderson, M. C., Ochsner, K. N., Kuhl, B., Cooper, J., Robertson, E., Gabrieli, S. W., Glover, G. H., & Gabrieli, J. D. (2004). Neural systems underlying the suppression of unwanted memories. *Science*, 303(5655), 232-235.
- Ayers, M. S., & Reder, L. M. (1998). A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin & Review*, 5(1), 1-21.
- Benoit, R. G., & Anderson, M. C. (2012). Opposing mechanisms support the voluntary forgetting of unwanted memories. *Neuron*, 76(2), 450-460.
- Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12), 2767-2796.
- Bonni, S., Veniero, D., Mastropasqua, C., Ponzio, V., Caltagirone, C., Bozzali, M., & Koch, G. (2015). TMS evidence for a selective role of the precuneus in source memory retrieval. *Behavioural Brain Research*, 282, 70-75.
- Born, J., & Wilhelm, I. (2012). System consolidation of memory during sleep. *Psychological research*, 76(2), 192-203.
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10, 433-436.
- Braver, T. S., Barch, D. M., Gray, J. R., Molfese, D. L., & Snyder, A. (2001). Anterior cingulate cortex and response conflict: effects of frequency, inhibition and errors. *Cerebral Cortex*, 11(9), 825-836.
- Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, 4(6), 215-222.
- Butler, A. J., & James, K. H. (2010). The neural correlates of attempting to suppress negative versus neutral memories. *Cognitive, Affective, & Behavioral Neuroscience*, 10(2), 182-194.
- Caballero-Gaudes, C., & Reynolds, R. C. (2017). Methods for cleaning the Bold fMRI signal. *NeuroImage*, 154, 128-149.
- Cabeza, R. (2008). Role of parietal regions in episodic memory retrieval: the dual attentional processes hypothesis. *Neuropsychologia*, 46(7), 1813-1827.
- Cabeza, R., Ciaramelli, E., Olson, I. R., & Moscovitch, M. (2008). The parietal cortex and episodic memory: an attentional account. *Nature Reviews Neuroscience*, 9(8), 613-625.
- Cabeza, R., & Nyberg, L. (2000). Imaging cognition II: An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience*, 12(1), 1-47.

- Carter, C. S., & Van Veen, V. (2007). Anterior cingulate cortex and conflict detection: an update of theory and data. *Cognitive, Affective, & Behavioral Neuroscience*, 7(4), 367-379.
- Catarino, A., Küpper, C. S., Werner-Seidler, A., Dalgleish, T., & Anderson, M. C. (2015). Failing to forget: Inhibitory-control deficits compromise memory suppression in posttraumatic stress disorder. *Psychological Science*, 26(5), 604-616.
- Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129(3), 564-583.
- 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.
- Depue, B. E. (2012). A neuroanatomical model of prefrontal inhibitory modulation of memory retrieval. *Neuroscience & Biobehavioral Reviews*, 36(5), 1382-1399.
- 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. *Quarterly Journal of Experimental Psychology*, 64(2), 283-310.
- 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.
- 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., 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.
- Edelson, M. G., Dudai, Y., Dolan, R. J., & Sharot, T. (2014). Brain substrates of recovery from misleading influence. *Journal of Neuroscience*, 34(23), 7744-7753.
- Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28), 7900-7905.
- Fletcher, P. C., Frith, C. D., Baker, S. C., Shallice, T., Frackowiak, R. S., & Dolan, R. J. (1995). The mind's eye - precuneus activation in memory-related imagery. *Neuroimage*, 2(3), 195-200.

- Friston, K. J., Holmes, A., Poline, J. B., Price, C. J., & Frith, C. D. (1996). Detecting activations in PET and fMRI: Levels of inference and power. *Neuroimage*, 4(3), 223-235.
- Garrett, H. E. (1922). A study of the relation of accuracy to speed. *Archives of Psychology*, 56, 1-105.
- Gentner, D., & Stevens, A. L. (Eds.). (2014). *Mental models*. Psychology Press.
- Gordon, A., Brooks, J. C., Quadflieg, S., Ecker, U. K., & Lewandowsky, S. (2017). Exploring the neural substrates of misinformation processing. *Neuropsychologia*, 106, 216-224.
- Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1), 63–72.
- Griswold, M. A., Jakob, P. M., Heidemann, R. M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., & Haase, A. (2002). Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 47(6), 1202-1210.
- Grol, M., Vingerhoets, G., & De Raedt, R. (2017). Mental imagery of positive and neutral memories: A fMRI study comparing field perspective imagery to observer perspective imagery. *Brain and Cognition*, 111, 13-24.
- Henkel, L. A., & Mattson, M. E. (2011). Reading is believing: The truth effect and source credibility. *Consciousness and Cognition*, 20(4), 1705-1721.
- Henson, R. N., Hornberger, M., & Rugg, M. D. (2005). Further dissociating the processes involved in recognition memory: An fMRI study. *Journal of Cognitive Neuroscience*, 17(7), 1058-1073.
- Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology*, 118(2), 126–135.
- Jenkinson, M. (2013). Measuring Transformation Error by RMS Deviation (Technical Report TR99MJ1). Retrieved from FMRIB website: <http://www.fmrib.ox.ac.uk/datasets/techrep/tr99mj1/tr99mj1/index.html>
- Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2), 143–156.
- Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2), 825–841.
- 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), 1420-1436.

- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, UK: Cambridge University Press
- Johnson-Laird, P. N. (2012). *Inference with mental models*. The Oxford handbook of thinking and reasoning, 134-145.
- Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports*, 6, 39589.
- 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.
- Kessler, D., Angstadt, M., & Sripada, C. S. (2017). Reevaluating "cluster failure" in fMRI using nonparametric control of the false discovery rate. *Proceedings of the National Academy of Sciences*, 114(17), E3372-E3373
- Levy, B. J., & Anderson, M. C. (2012). Purging of memories from conscious awareness tracked in the human brain. *Journal of Neuroscience*, 32(47), 16785-16794.
- 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.
- Liu, T. T., Frank, L. R., Wong, E. C., & Buxton, R. B. (2001). Detection power, estimation efficiency, and predictability in event-related fMRI. *Neuroimage*, 13(4), 759-773.
- Lundstrom, B. N., Petersson, K. M., Andersson, J., Johansson, M., Fransson, P., & Ingvar, M. (2003). Isolating the retrieval of imagined pictures during episodic memory: activation of the left precuneus and left prefrontal cortex. *Neuroimage*, 20(4), 1934-1943.
- Lundstrom, B. N., Ingvar, M., & Petersson, K. M. (2005). The role of precuneus and left inferior frontal cortex during source memory episodic retrieval. *Neuroimage*, 27(4), 824-834.
- Marsh, E. J., & Fazio, L. K. (2006). Learning errors from fiction: Difficulties in reducing reliance on fictional stories. *Memory & Cognition*, 34(5), 1140-1149.
- Marsh, E. J., Meade, M. L., & Roediger III, H. L. (2003). Learning facts from fiction. *Journal of Memory and Language*, 49(4), 519-536.
- Marshall, L., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in cognitive sciences*, 11(10), 442-450.

- Masson, S., Potvin, P., Riopel, M., & Foisy, L. M. B. (2014). Differences in brain activation between novices and experts in science during a task involving a common misconception in electricity. *Mind, Brain, and Education*, 8(1), 44-55.
- Mikl, M., Mareček, R., Hlušík, P., Pavlicová, M., Drastich, A., Chlebus, P., et al. (2008). Effects of spatial smoothing on fMRI group inferences. *Magnetic Resonance Imaging*, 26(4), 490-503.
- Moscovitch, M., & Melo, B. (1997). Strategic retrieval and the frontal lobes: Evidence from confabulation and amnesia. *Neuropsychologia*, 35(7), 1017-1034.
- Nee, D. E., Wager, T. D., & Jonides, J. (2007). Interference resolution: insights from a meta-analysis of neuroimaging tasks. *Cognitive, Affective, & Behavioral Neuroscience*, 7(1), 1-17.
- Nee, D. E., Brown, J. W., Askren, M. K., Berman, M. G., Demiralp, E., Krawitz, A., & Jonides, J. (2013). A meta-analysis of executive components of working memory. *Cerebral Cortex*, 23(2), 264-282.
- Nichols, T. E., Eklund, A., & Knutsson, H. (2017). A defense of using resting-state fMRI as null data for estimating false positive rates. *Cognitive Neuroscience*, 8(3), 144-149.
- Nieuwland, M. S., Petersson, K. M., & Van Berkum, J. J. (2007). On sense and reference: Examining the functional neuroanatomy of referential processing. *NeuroImage*, 37(3), 993-1004.
- Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9(1), 97-113.
- Ollinger, J. M., Shulman, G. L., & Corbetta, M. (2001). Separating processes within a trial in event-related functional MRI: I. The method. *Neuroimage*, 13(1), 210-217.
- Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, 72(5), 692-697.
- Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2), 115.
- 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), 62-74.
- Schouten, J. F., & Bekker, J. A. M. (1967). Reaction time and accuracy. *Acta Psychologica*, 27, 143-153.
- Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and

- public information campaigns. *Advances in experimental social psychology*, 39, 127-161.
- 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.
- Seghier, M. L. (2013). The angular gyrus: multiple functions and multiple subdivisions. *The Neuroscientist*, 19(1), 43-61.
- Shallice, T., Fletcher, P., Frith, C. D., Grasby, P., Frackowiak, R. S. J., & Dolan, R. J. (1994). Brain regions associated with acquisition and retrieval of verbal episodic memory. *Nature*, 368(6472), 633.
- Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, 124(2), 209-215.
- Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., ... Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(SUPPL. 1), 208–219.
- van Oostendorp, H., & Bonebakker, C. (1999). *Difficulties in updating mental representations during reading news reports*. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 319–339). Mahwah, NJ: Erlbaum.
- Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking & Reasoning*, 11(3), 239-278.
- Vilberg, K. L., & Rugg, M. D. (2008). Memory retrieval and the parietal cortex: a review of evidence from a dual-process perspective. *Neuropsychologia*, 46(7), 1787-1799.
- Vosniadou, S. (2012). *Reframing the classical approach to conceptual change: Preconceptions, misconceptions and synthetic models*. In *Second international handbook of science education* (pp. 119-130). Springer, Dordrecht.
- Wagenmakers, E. J., Van Der Maas, H. L., & Grasman, R. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, 14(1), 3-22.
- Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, 41(1), 67-85.
- Wilkes, A. L., & Leatherbarrow, M. (1988). Editing episodic memory following the identification of error. *The Quarterly Journal of Experimental Psychology*, 40(2), 361-387.

- Winkler, A.M. (2012). AutoAQ: Automatic atlas queries in FSL [Automated labelling of clusters of activations]. Retrieved from <http://brainder.org/tag/autoaq/> (accessed 20.08.16)
- Wood, T., & Porter, E. (2017). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 1-29.
- Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modelling of fMRI data. *NeuroImage*, 14(6), 1370–1386.
- World Economic Forum (2013). Digital wildfires in a hyperconnected world. Retrieved from <http://reports.weforum.org/global-risks-2013/risk-case-1/digital-wildfires-in-a-hyperconnected-world/>
- Worsley, K. J., Evans, A. C., Marrett, S., & Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism: Official Journal of the International Society of Cerebral Blood Flow and Metabolism*, 12(6), 900–918.
- Ye, Z., & Zhou, X. (2009). Conflict control during sentence comprehension: fMRI evidence. *Neuroimage*, 48(1), 280-290.