Trust in information, political identity, and the brain

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Philosophical Transactions of the Royal Society B: Biological Sciences

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Trust in information, political identity, and the brain: An interdisciplinary fMRI study

Adam Moore, Sujin Hong and Laura Cram

Neuropolitics Research Lab, University of Edinburgh, 18 Buccleuch Place, Edinburgh, UK, EH8 9LN
Psychology, University of Edinburgh

Keywords: Trust, identity, functional Magnetic Resonance Imaging (fMRI), media and information source

Summary

Misinformation has triggered government inquiries and threatens the perceived legitimacy of campaign processes and electoral outcomes. A new identity polarisation has arisen between Remain and Leave sympathisers in the UK Brexit debate, with associated accusations of misinformation use. Competing psychological accounts of how people come to accept and defend misinformation pit self-reinforcing motivated cognition against lack of systematic reasoning as possible explanations. We harness insights from political science, cognitive neuroscience, and psychology to examine the impact of trust and identity on information processing regarding Brexit in a group of Remain identifiers. Behaviourally, participants’ affective responses to Brexit-related information are affected by whether the emotional valence of the message is compatible with their beliefs on Brexit (positive/negative) but not by their trust in the source of information. However, belief in the information is significantly affected by both (dis)trust in information source and by belief compatibility with the valence of the information. Neuroimaging results confirm this pattern, identifying areas involved in judgements of the self, others, and automatic processing of affectively threatening stimuli, ultimately supporting motivated cognition accounts of misinformation endorsement.

Introduction

There is increasing concern about social media’s magnification of misinformation, misleading information, and conspiracy theories (hereafter misinformation) and its impact on democracy (e.g. Allcott & Gentzkow, 2017; Gaughan, 2017; Hansen & Lim, 2019; for reviews see Colley, Granelli, & Althuis, 2020 and Tucker et al., 2018). Psychological accounts of misinformation acceptance differ as to whether effortful cognition is primarily involved in the acceptance or the rejection of misinformation. Accounts on the acceptance end of the spectrum (self-reinforcing (SR) accounts) propose that misinformation acceptance is largely the product of alignment with an individual’s preferred worldview/ideology/prior beliefs. When such alignment occurs, individuals may selectively attend to and process information compatible with what they already believe (Kahan et al., 2012; Knobloch-Westerwick, Mothes, & Polavin, 2020; Peterson, Skov, Serritzlew, & Ramsay, 2013). Alternatively, they may selectively search for perceived weaknesses in belief-incompatible information to dismiss it, even at the expense of truth (Ditto & Lopez, 1992; cf. Lewandoski & Oberauer, 2016). Effortful rejection (ER) accounts propose that misinformation is typically accepted on the basis of superficial characteristics and shallow processing (cf. Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2015), but when reasoning is engaged, people can reliably discriminate fake from real news regardless of ideological compatibility (Pennycook & Rand, 2019; though right-wing conservatives show greater susceptibility to misinformation than left-wing liberals - Baron & Jost, 2019;
Jost, 2017; but see Ditto et al., 2019). A recent identity-based hybrid (I-bH) account proposes that contextually salient goals control where effort is deployed, with self/identity/ideology-reinforcement weighed against desires for epistemic accuracy (Van Bavel & Pereira, 2018). This model posits key roles for orbitofrontal cortex (OFC) and ventromedial prefrontal cortex (vmPFC) in computing competing goals and the value of differing beliefs, respectively, which can then bias downstream processing.

Neuroimaging evidence is broadly compatible with these accounts. Areas associated with affect/emotion, value computation, and conflict processing, such as lateral and medial OFC and anterior cingulate cortex (ACC), show increased activity when political partisans are confronted with information potentially damaging to their preferred candidate (Westen, Blagov, Harenski, Kilts, & Hamann, 2006) and when trying to control affective reactions while explicitly evaluating socially relevant stimuli (e.g., abortion, murder; Cunningham, Raye, & Johnson, 2004). Similarly, vmPFC is more active when pre-existing beliefs bias reasoning performance (Goel & Dolan, 2003) and such activity predicts increasing belief in given information (Harris, Sameer, & Cohen, 2008). This is consistent with the I-bH model, and to some extent with SR frameworks. By contrast, and supporting both ER and I-bH accounts, disbelieving information activates distributed brain regions involved in effortful cognition, including the left inferior frontal gyrus (IFG), bilateral anterior insula (AI), dorsal ACC extending to superior frontal gyrus (SFG), and superior parietal lobe (Harris et al., 2008), while right lateral PFC activation corresponds to correct logical inference in the face of competing beliefs (Goel & Dolan, 2003).

However, this research does not systematically disentangle the affective impact of the belief (in)compatibility of information itself versus the (dis)trust of the source of that information. In other words, how do people respond when (in)compatible information is delivered via a (dis)trusted source? This is critical because source identity can impact information acceptance regardless of actual content (e.g., Cohen, 2003), and such identity cues can trigger motivated cognition (Peterson et al., 2013; see also Cram, Moore, Olivieri, & Suessenbach, 2018). However, the affective impact of belief incompatible information (without source identity) can drive a disconnection with external information and increased processing in default-mode networks (Kaplan, Gimbel, & Harris, 2016). Here we investigate how information source and belief compatibility combine to affect judgments of belief and feelings about politically salient information. Neuroimaging is useful in disentangling predictions made by different frameworks: SR accounts predict that trusted sources and compatible information should result in greater belief and greater activation in areas associated with effortful cognition (e.g. lateral PFC, ACC, etc.), and a mismatch between (dis)trusted source and (in)compatible information may elicit effortful cognition related to decreased belief (i.e. motivated rejection). ER accounts predict less activity in these areas in response to trusted/compatible stimuli, but possibly more in response to mismatch stimuli (which potentially provide cues triggering reasoning processes), though it is unclear what relation effortful reasoning may have to belief in this case. Under the I-b-H account we can assume that our participants are motivated in large part, though not completely, by identity protection concerns related to their self-identification as Remain voters (on which they were pre-screened; see Methods). This suggests predictions largely in line with SR accounts, though vmPFC activity related to judging various beliefs (identity protection and truth/accuracy) against one another may emerge.

Methods

Participants

Thirty-eight, right-handed, UK national, native English-speaking participants aged 18-42 (Mean = 25.5, 20 females) underwent blood oxygenation level-dependent (BOLD) contrast fMRI scanning. Three were excluded due to failure to record button responses. All were Brexit Remain supporters familiar with social media platforms. Pre-screening was conducted separately to avoid affecting behaviour. All participants provided informed consent and were paid £20. A debrief followed the fMRI scans. The study was approved by the Research Ethics Committees of the School of Social and Political Sciences and the Research Ethics procedures of the Edinburgh Imaging facility at The Queen’s Medical Research Institute (EiF-QMRI) at the University of Edinburgh.
Stimuli

We extracted text from tweets related to Brexit (posted between 01/01/2019-14/08/2019) gathered from various media sources, and separately we pre-screened all of our participants to determine a bespoke list of their three most-used and least-used media sources (based on the British social attitudes 2018 questionnaire, Q216, [https://www.bsa.natcen.ac.uk/media/39286/questionnaire_2018_v2.pdf], and an open-choice option; see Supplementary Table 1). In the initial stage, two reviewers rated tweet texts as positive, negative, or ambiguous. From the negative and positive rated texts, we selected short one-line tweets and, where necessary, edited these to remove source information. The remaining tweets were validated in a pilot experiment by separate Remain-voting participants (i.e. not the fMRI sample; Supplementary Methods, Supplementary Figure 2; A full list of 120 tweet texts (60 negative, 60 positive) used in the present study are in Supplementary Table 4.).

Negative valence tweets (i.e. regarding substantial costs/consequences) on Brexit are belief-compatible for Remain voters who oppose Brexit, while positive valence tweets (i.e. regarding benefits/rewards) are belief-incompatible. These factors, information source (Trusted, Distrusted) and emotional valence/belief compatibility (Negative/Compatible, Positive/Incompatible), were fully crossed in a 2x2 ANOVA design (see Figure 1, panels b and c).

Task and fMRI design

Participants completed 3 practice trials before confirming that they understood the task. We used a rapid event related design. During the task a fixation cross was shown for the initial 20 seconds (s). The tweets and accompanying logos were presented for 2.5s; two rating tasks followed, measuring belief and feeling, counterbalanced across participants. For the belief rating task, the statement ‘I believe this is true’ was presented with a 7-point scale from ‘Not at all (1)’ to ‘Completely (7)’ (reversed for 50% of participants). For feeling ratings, the phrase ‘This makes me feel’ was presented with a 7-point scale from ‘Negative (1)’ to ‘Positive (7)’ (reversed for 50% of participants). Cursor position on the scales always initiated in the middle. Inter-trial interval (ITI) was jittered 2-4s. Participants used left and right thumb buttons to move the cursor to the left and to the right on the response scales, respectively, and pressed the right index button to finalise their response (Figure 1).
Figure 1. fMRI task and design. (a) Schematic of a single trial during fMRI scan. (b) Examples of tweet texts and logos. (c) A 2×2 repeated measures ANOVA design (emotional valence × information source).

fMRI apparatus and MRI data acquisition

Participants viewed the visual stimuli via a mirror attached on a head coil; we used PsychoPy (v.3) to present stimuli (Peirce et al., 2019). Two response grips were used (NordicNeuroLab). We acquired imaging data of high resolution T1-weighted 3D anatomical images (TR 2500ms, TE 4.37ms, 256×256mm 100% FOV, 1mm slice thickness), gradient-echo fieldmaps (same slice as the EPI images, resolution 3×3mm, TR=599ms, TE=5.19ms, TE=7.65ms, flip angle 60°, bandwidth=260, TA=1m15s), and gradient echo-planar imaging (EPI) functional images (TR 2260ms, TE 27ms, 192mm 100% FOV, in-plane resolution 3×3mm, 3mm slice thickness, no gap, flip angle 80°, slice order=Siemens ascending interleaved), using a 3T Siemens skyra MRI scanner (Siemens Healthineers, Erlangen, Germany).

fMRI data analysis

fMRI data were analysed using Statistical Parametric Mapping (SPM12 (v 7771), Wellcome Trust Centre for Neuroimaging, London, UK) and MATLAB 2019b (MathWorks). In pre-processing, the first two functional images were discarded to ensure T1 saturation. Remaining functional images were corrected for slice-timing, realigned, distortion-corrected via gradient fieldmap (Realign & Unwarp of SPM12 using a voxel displacement map, Jezzard and Balaban, 1995; Hutton et al. 2002) and co-registered to the corresponding individual T1 weighted anatomical images. The anatomical images were segmented, with parameters used to normalise functional images with the Montreal Neurological Institute (MNI) brain template. Normalised functional data were smoothed with a 3-dimensional isotropic Gaussian kernel 6mm full width at half maximum. For statistical processing, the general linear model (GLM) was used at 1st level analysis. Presentation of tweet, rating of belief score, and rating of feeling score were modelled as events for each condition (2 levels of 2 factors) using the canonical hemodynamic response function (HRF). Additional regressors of no interest included the temporal derivative of each GLM predictor, high-pass filter (128s) regressors, and 6 rigid body transformation parameters from spatial realignment. Contrasts of parameter estimates (>fixation (not explicitly modelled)) of each event were used to compute contrast images at the 1st level, which were then entered into the 2nd level GLM model of a two-way repeated
measures (2 x 2) flexible factorial ANOVA for the whole brain. We performed separate ANOVAs during the first presentation of a tweet, belief rating task, and feeling rating task. Statistical threshold was set at $p < .001$ (uncorrected) at the voxel level, and with a cluster-level extent threshold of $p < .05$ (FDR) applied for multiple corrections. Cluster parameter estimates included all voxels, extracted using the Marsbar toolbox (Brett et al., 2002).

Results

Behavioural results

Rating scores and reaction times (RTs) collected during fMRI scans were standardized and analysed using R (v.4.0.2), brms (v2.13.5), and STAN. All regression models sampled from 4 chains with 1000 trial burn-in and 6000 iterations each, using regularizing, weakly informative Gaussian priors ($\mu = 0$, $\sigma = 1$). All chains converged as indicated by both caterpillar plots and $\hat{R}$ estimates (all = 1.00). Intervals are 95% highest posterior density intervals (HDI) unless otherwise stated.

As belief and feeling ratings were negatively correlated, $r = -.19$, HDI = [-.22, -.16], we performed a 2x2 repeated measures Bayesian multilevel multivariate regression, with fixed effects of information source ( Trusted, Distrusted) and valence (Negative, Positive) and random intercepts for participants and stimuli. Trusted sources increased belief ratings, $b = .22$, HDI = [.15, .29], but did not appreciably affect feeling ratings, $b = .01$, HDI = [-.05, .07]. Positive valence/belief-incompatible information decreased belief ratings, $b = -.79$, HDI = [-.97, -.61], and increased positive feeling ratings, $b = 1.17$, HDI = [1.05, 1.29]. There was no meaningful interaction for belief ratings, $b = -.01$, HDI = [-.10, .09], nor for feeling ratings, $b = .04$, HDI = [-.05, .12].

Since RTs were also correlated ($r = .04$, HDI = [.01, .07]), we repeated the above analysis. Predicting belief RTs revealed negligible effects of trusted sources, $b = -.04$, HDI = [-.12, .04], positive valence/belief-incompatible information, $b = -.04$, HDI = [-.15, .05], and no interaction, $b = .05$, HDI = [.07, .16]. For feeling RTs the pattern was similar to source, $b = -.06$, HDI = [-.14, .02], positive valence/belief-incompatibility, $b = .10$, HDI = [-.01, .19], and the interaction, $b = .11$, HDI = [-.01, .20], though the latter two effects provide some suggestion that participants took slightly longer to give feeling ratings to belief-incompatible information, particularly when it came from a trusted source.

fMRI results

A two-way repeated measures ANOVA of whole brain during the belief rating task showed a significant main effect of the information source in the middle occipital gyrus (MOG) of the right hemisphere (Table 1, MNI coordinates $x, y, z = [48, 76, 4]$, Z score = 4.41), and a significant interaction in the cuneus and precuneus in the right hemisphere (Figure 2a, MNI coordinates $x, y, z = [21, -82, 38]$, Z score = 4.11), where the parameter estimates of Distrusted-Positive and Trusted-Negative were significantly greater than Distrusted-Negative and Trusted-Positive, respectively (Figure 2b). There was a negative correlation, $r = -.33$, HDI = [-.66, -.01], between parameter estimates in the cuneus/precuneus and belief ratings for belief-incompatible information from a distrusted source (i.e. distrusted source-positive valence) (Figure 2c), but no equivalent result for compatible information from a trusted source, $r = .12$, HDI = [-.22, .48].
Figure 2. Significant interaction between emotional valence and information source during the belief rating task. a. Cuneus and precuneus activity in the right hemisphere. b. Parameter estimates of the Cuneus/Precuneus activity (error bars denote ± SEM.). c. Correlation (and 95% HDI) between parameter estimates and belief ratings (Distrusted-Positive).

Table 1. fMRI ANOVA results during the belief rating task

<table>
<thead>
<tr>
<th>Brain regions</th>
<th>MNI coordinates</th>
<th>Z score</th>
<th>Cluster size (k)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main effect of information source</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R Middle occipital gyrus</td>
<td>48  -76  -4</td>
<td>4.41</td>
<td>21</td>
</tr>
<tr>
<td><strong>Interaction: information source * emotional valence</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R Cuneus and Precuneus</td>
<td>21  -82  38</td>
<td>4.11</td>
<td>34</td>
</tr>
</tbody>
</table>

R, right hemisphere

An identical analysis on the feeling rating task showed significant activations in the main effect of valence/belief compatibility across fronto-parietal-occipital lobes (Figure 3., Supplementary Table 3, Supplementary Figure 1.), which are largely clustered and include bilateral superior/middle/inferior occipital gyrus, bilateral postcentral gyrus, superior frontal gyrus, superior parietal gyrus, supplementary motor areas extending to middle cingulate gyrus, right putamen/insula, right rolandic operulum extending to heschl’s gyrus, right precentral gyrus extending to superior frontal gyrus, angular gyrus and supramarginal gyrus. There was no significant activation for the main effect of information source nor interaction.
Figure 3. The main effect of emotional valence on neural activation during the feeling rating task (see also Supplementary Table 3 and Supplementary Figure 1). PreCG, Precentral gyrus; IFGoperc, Inferior frontal gyrus, opercular part; PostCG, Postcentral gyrus; SMG, Supramarginal gyrus; SMA, Supplementary motor area; MCC, Middle cingulate gyrus; ROL, Rolandic operculum; HES, Heschl’s gyrus; SPG, Superior parietal gyrus; SOG, Superior occipital gyrus; IOG, Inferior occipital gyrus; ITG, Inferior temporal gyrus; MTG, Middle temporal gyrus; MOG, Middle occipital gyrus; ANG, Angular gyrus; INS, Insula; CUN, Cuneus; PCUN, Precuneus.

Discussion

Our participants rated their belief in, and feelings about, politically salient tweets. Behaviourally both (dis)trust in information source and emotional valence/belief compatibility affected belief ratings, but only valence/belief compatibility affected feeling ratings. RTs were slightly longer when rating feelings about positive valence/belief-incompatible tweets. The fMRI analyses of the feeling rating task revealed only a main effect of valence/belief compatibility in a wide network with clustered regions related to multiple cognitive processes, including language processing in the frontal, temporal, and parietal regions, motor movements in the precentral and postcentral gyrus, visual processing in the occipital areas in both hemispheres. By contrast, when rating belief in our political tweets, there was a significant main effect of information source in right MOG and a significant interaction effect in right cuneus and precuneus. That interaction effect, driven by the negative correlation with belief ratings for distrusted positive-valenced stimuli, suggests additional processing occurred for information matching source expectations (i.e. distrusted sources providing belief-incompatible information).

While these results are limited, and somewhat unexpected given our initial hypotheses centred on PFC activity, they seem most compatible with the Ib-H (Van Bavel & Pereira, 2018) and SR accounts (Kahan, 2013, 2017; Lewandoski & Oberauer, 2016) of misinformation acceptance. Emotion and attention are known to heighten sensitivity to visual cues (Lane, Chua & Dolan, 1999), and recent neurocomputational work suggests that emotion-related activity in visual cortex, such as we found in the belief rating task, may represent an interface of sensory representations of the environment and previous knowledge (Kragel, Reddan & LaBar et. al., 2019). Encoding of high arousal negative information is associated with activity in the MOG, an area which may cooperate with closely related areas (e.g. posterior fusiform and inferior occipital gyrus) in a network specialising in identifying emotionally important visual clues (Mickley Steinmetz, Addis, & Kensinger, 2009); for example, both MOG and cuneus activity are involved in inferring threat from non-facial cues/body language (Kret et al., 2011). Given the identity-relevant nature of the information presented in our study, the activation we observe in MOG and cuneus is also consistent with middle occipital and cuneus involvement in both judgements and attitudes about others and the self (Bruneau & Saxe, 2010; Ochsner et al., 2004). Similarly, in close parallel to our findings, the cuneus is involved in information processing when invalid cues appear (i.e. when information is
detected to be misleading/wrong) and even more so when they are the result of human intention (vs preprogrammed stimuli; Ozdem et al., 2017), as well as when people disbelieve unpalatable political information relative to non-political information (Kaplan et al., 2016). People often deliberately scrutinise or attack incompatible and/or distrusted information more than they carefully consider appealing belief-compatible information (Ditto & Lopez, 1992), which is the pattern of activation we find in (pre)cuneus (see Figure 2b). The lack of significant PFC activation for belief-compatible information (from any source) seems consistent with this, and inconsistent with a role for careful reasoning in our task, though we cannot draw strong inferences from the lack of an effect in this context. It may well be that our participants, selected for strong Brexit views, were immediately ‘on guard’ when our stimuli were Brexit related. Such identity-based motivated scrutiny is likely somewhat automatic, with respect to core identity beliefs.

Indeed, this type of motivated rejection effect is also observed when participants refuse to update beliefs about the self or future finances when information is negative/undesirable (Coutts, 2019). Many also reject overwhelming scientific consensus on a variety of topics by positing complex, though absurd, conspiracy theories – essentially attacking the source of unpalatable information (e.g. Diethelm & McKee, 2009; Lewandowsky et al., 2015), but typically only when that information presents a challenge to their pre-existing worldview (Lewandowsky & Oberauer, 2016). Such reactions, though relying on some amount of deliberate cognition, are typically not products of systematic reasoning, but rather affective reactivity (van Prooijen & Douglas, 2018) that is largely effortless (Hughes & Zaki, 2015), though it can sometimes be the result of rational calculation to discredit information or sources that oppose or hinder one’s goals (Cook & Lewandowsky, 2016; Dunlap & Jacques, 2013). Our results suggest that participants integrated information regarding source trustworthiness into their belief ratings, and when confronted with belief-incompatible information from a distrusted source, believed it less, possibly as a function of differentially critical processing.

Conclusion

Under motivated cognition accounts of misinformation acceptance (SR and Ib-H), people can actively incorporate belief-compatible information and critically scrutinize incompatible information (Ditto & Lopez, 1992). In the latter case, people do not necessarily systematically reason about strengths and weakness of offending information, but rather critique to reject it and protect the self (cf. Amit, Menster, Arieli, & Porzycki, 2020). Remain and Leave identities, prevalent in the aftermath of the UK’s EU referendum, are personally important and cut across traditional party lines, generating affective polarization as intense as that of partisanship across a range of measures (Hobolt, Leeper & Tilley 2020). Our results seem to corroborate the view of motivated rejection of belief-compatible information. Belief-incompatibility itself and beliefs about trustworthiness of sources may be combined to motivate rejection of politically offensive information when that information is a threat to a deeply held/self-defining ideological belief.

Acknowledgments

This research was funded by the Economic and Social Research Council ES/R001901/1.

References


