Please cite this article in press as: van der Linden and Cohen, The neuroscience of misinformation: A research agenda, Neuron (2025), https://doi.org/10.1016/j.neuron.2025.05.010

Neuron

CellPress

NeuroView

The neuroscience of misinformation: A research agenda

Sander van der Linden[1,*] and Michael S. Cohen[2]
[1]Department of Psychology, University of Cambridge, Downing Site, Cambridge CB2 3EB, UK
[2]Department of Psychology, University of Chicago, Chicago, IL, USA
*Correspondence: sander.vanderlinden@psychol.cam.ac.uk
https://doi.org/10.1016/j.neuron.2025.05.010

The global spread of misinformation is undermining democracies worldwide. In this NeuroView, we explain how neuroscience can inform our basic understanding of what makes the brain susceptible to false information, how it spreads in society, and how neuroscience can help shape and optimize interventions to effectively counter it.

Introduction

The global spread of misinformation is a threat to public health, science, and democracies worldwide.[1,2] Examples abound. Following a tragic stabbing in the UK, fabricated anti-immigration stories about the assailant fueled a series of violent national riots. In 2020, an angry mob stormed the US Capitol after Donald Trump spread the false conspiracy theory that the presidential election had been stolen. Recent California wildfires have similarly attracted a surge of misinformation—including conspiracy theories about "Jewish space lasers"—and during the pandemic, dangerous viral rumors about homegrown COVID-19 remedies on social media led to spikes in mass poisonings.[1] The spread of misinformation can also impact society more indirectly by exacerbating affective polarization, undermining trust in institutions, and lowering the quality of societal discourse.

Misinformation is certainly not new, but the rate at which it can spread and reach new audiences on social media is unprecedented, with AI-assisted microtargeting and persuasive deepfakes of celebrities and politicians being just two contemporary examples. A recent consensus report from the American Psychological Association (APA) defined misinformation as "any information that is demonstrably false or otherwise misleading, regardless of intention or source."[1] The report also notes how the science of misinformation has advanced quickly over the last decade, with plentiful scholarship on the harmful consequences of misinformation on human judgment and decision-making, computational studies on how falsehoods spread on social media, as well as evidence-based interventions that aim to protect and empower the public to identify misinformation.

What is largely missing from this body of work, however, is a critical contribution from cognitive, social, affective, behavioral, and clinical neuroscience. Accordingly, in this NeuroView, we introduce the psychology of misinformation and outline how neuroscience can contribute to the study of what makes individuals susceptible to misinformation, how false information spreads from one individual to another, and how we can design evidence-based interventions to help people believe and share less misinformation (Figure 1).

Susceptibility to misinformation

What causes individuals to believe in false information? Although discerning truth from fiction is a complex cognitive task that likely involves many factors, one important cognitive phenomenon is known as the "illusory truth" effect.[1–3] Illusory truth occurs when people are more likely to believe claims that have been repeated than claims that have not been repeated, irrespective of their veracity. For example, individuals are more likely to believe the false claim that "the Great Wall of China is visible from space" or "the Earth is a perfect square" after repetition. Illusory truth already occurs after a single repetition and impacts both children and adults, typically up to 85% of tested samples. Importantly, illusory truth persists regardless of the plausibility of a claim, credibility of the source, and when people have demonstrated correct prior knowledge at the start of the experiment. The real-world implications of illusory truth are evident insofar as repeating lies, such as vaccines causing autism or the 2020 US presidential election being stolen, leads people to be more likely to believe such false claims over time. The most likely mechanism behind illusory truth is "'fluency," where repeated information is processed faster and easier than novel information. Perceptual fluency or the ease with which claims can be processed is subsequently used as an indicator of truth in judgment formation.

Interestingly, there is little research on the neural correlates of illusory truth. One likely candidate is the perirhinal cortex (PRC) in the medial temporal lobe, given its key involvement in recognition memory, conceptual implicit memory, and memory judgments driven by familiarity rather than recollection. Consistent with the PRC-mediated hypothesis for illusory truth, one fMRI study scanned 24 participants while they rated 360 trivia statements as true or false, of which half had been shown previously (repeated), while the other half was novel.[3] The main gist was to look for activity changes that differ between repeated versus novel claims. Results revealed a clear interaction between repetition and truth: PRC activity increased linearly as a function of perceived truth for repeated claims but decreased for novel claims. Future research could experimentally manipulate fluency to further investigate the role of the PRC in illusory truth.

Please cite this article in press as: van der Linden and Cohen, The neuroscience of misinformation: A research agenda, Neuron (2025), https://doi.org/10.1016/j.neuron.2025.05.010

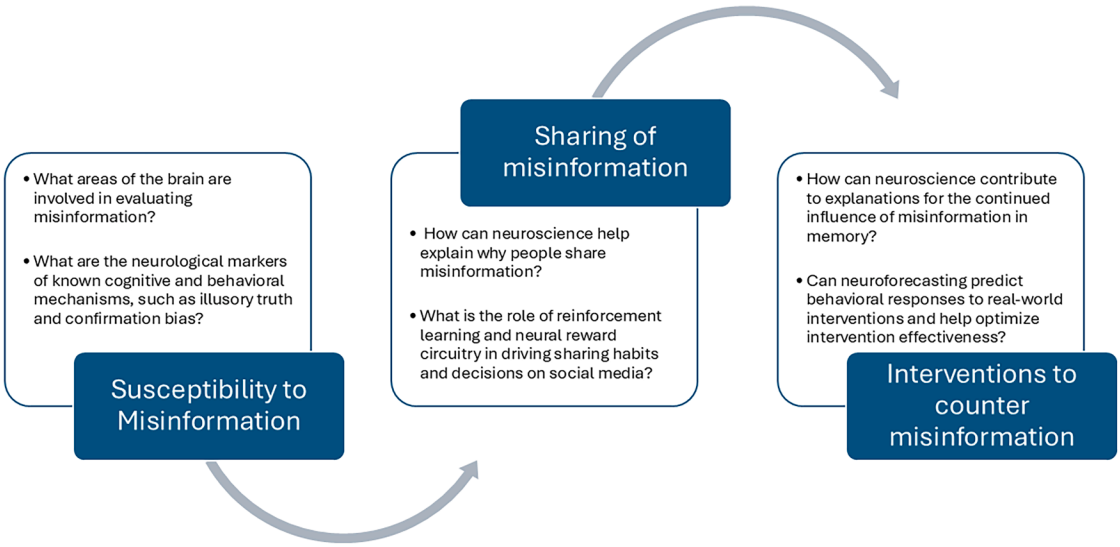

Figure 1. A research agenda for the neuroscience of misinformation

Other key mechanisms behind belief in misinformation include confirmation bias (i.e., the tendency for people to accept information that confirms their prior beliefs) and motivated cognition (the finding that our pre-existing goals, values, and identities shape basic cognitive processes including low-level perception, attention, and decision-making). Ample research has shown that pre-existing worldviews lead people to favor information congruent with those beliefs (reinforcing the initial attitude) and, in some cases, to also discount scientific evidence to the contrary. Like illusory truth, the behavioral evidence for confirmation bias is strong, but little is known about its neural underpinnings. A possible neural mechanism involves activation of the brain's dopaminergic reward system when people learn information that is valuable. This activity resembles a reward prediction error, but it occurs even when information does not increase the probability of a concrete reward (i.e., instrumental value). Non-instrumental value can include hedonic value (i.e., information containing good news rather than bad news) or cognitive value (i.e., when it supports or builds in a constructive way on people's schemas about the world).[4] Increased dopamine signaling might support engagement with confirmatory information, while reduced dopaminergic signaling, or increased noradrenergic signaling, may underlie avoidance of information that challenges one's existing beliefs. One common explanation for selective exposure is cognitive dissonance, or the psychological discomfort people feel when they learn about inconsistencies between their beliefs and behavior. For example, when misinformation is highly consistent with a person's attitude, corrections induce greater mental discomfort, which predicts continued reliance on the misinformation when making inferences. At a neural level, the anterior cingulate cortex often has been implicated in detecting cognitive conflict. Also relevant is a finding that people are less sensitive (at a neural level) to the strength of the opinions of peers when those opinions are disconfirming. More generally, susceptibility to fake news likely involves theories of how the brain deals with unexpected new information in light of established priors. Future work examining the brain response to truthful information that challenges one's sociopolitical beliefs and misinformation that reinforces one's (prior) beliefs would elucidate our understanding of how people develop worldviews that do not align with evidence.

Other interesting insights can be derived from research in clinical neuroscience. For example, consider the fact that false beliefs are prevalent in individuals with psychiatric disorders and neurodegenerative conditions such as frontotemporal and Lewy body dementia. Some scholars have suggested that the study of patients with these forms of dementia could reveal insights into the shared neural mechanisms that explain the origins of misperceptions. For instance, dysfunction in frontal brain circuitry (common to such disorders) might hamper an individual's ability to discern the plausibility of claims. Overlap has also been noted between schizotypy, paranoia, and conspiratorial ideation. Lastly, recent research has used tools from neuropsychological assessment, such as the Wisconsin Card Sort Task (WCST) to assess cognitive flexibility. This line of research has consistently found that strong political ideologies are associated with cognitive inflexibility, which in turn predicts extremist beliefs and willingness to engage in violence.

The spread of misinformation

Neuroimaging work has found that social- and value-related brain signals in a small group of scanned individuals can predict population-level sharing of news media.[5] Brain regions associated with reward, mentalizing, and introspection all appear to be relevant in choices to share online content. For instance, brain responses to news articles in these regions predict real-world sharing via the *New York Times* website and on Facebook, with effects from mentalizing and introspection regions mediated by the reward response.

Decisions to share misinformation with others, like choices to share true

Please cite this article in press as: van der Linden and Cohen, The neuroscience of misinformation: A research agenda, Neuron (2025), https://doi.org/10.1016/j.neuron.2025.05.010

information, are likely influenced by brain structures involved in reward and social interaction. For example, research has shown that misinformation is often characteristically novel, shocking, and outrage provoking, which coincides with the finding that dopamine-controlled pathways are activated during information sharing and that highly salient stimuli increase dopamine levels in the medial prefrontal cortex (mPFC). Interestingly, receiving likes on social media has been associated with activity in the ventral striatum and other reward-processing brain regions and is shaped in a manner consistent with computational models of reinforcement learning.[6] The reward structure of social media is such that it often rewards extremist, emotive, and polarizing content.[1] This is relevant to habit formation insofar as habits form when people derive rewards from performing a behavior in response to a repeated cue. Research has shown that habitual sharers who automatically react to platform cues have lower information discernment and a general insensitivity to the accuracy of information.

Although promoting better information discernment generally does lead to less sharing of misinformation—implying that one reason why people share misinformation is because they find it hard to differentiate true from false news—people can also share misinformation, irrespective of its accuracy, for social or political reasons. Indeed, another important reason why people share misinformation is to signal group membership and reinforce identity-driven motivations[1]—for example, to propagate favorable narratives about the in-group or to spread derogating (mis)information about outgroups. Social media algorithms seem to especially incentivize derogating "the other side," as engagement is often driven by toxic, low-quality, emotive, and polarizing content. In recent years, research in social neuroscience has identified a network of brain regions relevant to evaluating group identity and "us" versus "them" judgments, including the amygdala (threat), fusiform gyrus (social perception), and ventral striatum (reward processing). The social neuroscience of why people share misinformation about other groups is an important area for future research.

Interventions to counter misinformation

Reactive: Debunking and fact checking

The most straightforward approach to countering misinformation is presenting factual information that debunks a false claim. Indeed, there is strong evidence that this approach can be effective in reducing inaccurate beliefs.[1,2] At the same time, debunking false information has major limitations. First, fact checking each piece of false content is a laborious process. Additionally, there are important limitations to human cognitive and emotional processing that cause debunking to be only partially effective. For instance, failure to remember a correction when there is a delay between exposure and test has been associated with increased endorsement of misinformation after a delay, a phenomenon known as belief regression. Memory failures are one possible mechanism for a broader phenomenon known as continued influence effects by which debunked information can still impact causal reasoning and other behavioral responses despite the acknowledgment of a correction.[2] Insights from cognitive neuroscience can play a valuable role toward elucidating the mechanisms behind continued influence effects.

Two theoretical explanations have drawn the most attention. The first explanation suggests that once false information has been used to create a mental model, people have difficulty forming a new mental model that does not rely on the retracted information. An alternative explanation focuses on selective retrieval, i.e., false information being more accessible in memory than retractions. This account derives from the idea that a vague sense of familiarity is sufficient for false information to influence subsequent inferences, while for a retraction to be effective, it must be bound to the original stimulus.[2] The selective retrieval explanation in particular draws upon a rich body of literature on the neural basis of memory. Specifically, it has been proposed that the hippocampus plays a critical role in binding together distinct representations in memory, while a nearby cortical structure, the PRC, is responsible for less-precise feelings of familiarity. Activity in the PRC can also distinguish between retrieval of recollection-based versus familiarity-based memories. An fMRI study found increased activity in parietal regions involved in recollection (angular gyrus and precuneus) when people successfully identified retracted information as false, relative to when consistent information was correctly identified as true. This provides evidence for recollection playing a role in rejecting retracted claims.[2]

However, continued influence effects are not only present when people explicitly infer the cause of an event. Other work has found these effects when derogatory information about political candidates is presented and subsequently corrected.[7] Specifically, mock candidates targeted by false accusations that were refuted were judged more negatively than those that were never targeted. Derogatory information is hypothesized to induce lasting negative judgments because an accusation associates a negative affective response with the targeted candidate, while a retraction only engages cognitive processing. This explanation necessarily draws on neuroscience evidence. Specifically, a memory boost for accusation stimuli was retained after a 2-day delay, while a memory boost for refutation stimuli was apparent after a short delay but not after this longer delay. Prior neuroscience work has shown that memory modulation via the amygdala is stronger at a delay compared to an immediate test due to enhanced consolidation. Furthermore, an fMRI study using the same paradigm found that brain regions typically associated with socioemotional processing (e.g., lateral orbitofrontal cortex, left temporoparietal junction) were activated in response to candidates accused of misconduct, and this increase in activity was not reliably reduced when the accusation had been refuted. Further studies informed by social, affective, and cognitive neuroscience will be needed to help establish why people fail to adjust misperceptions in light of corrections.

Proactive: Prebunking and psychological inoculation

Given the limitations of debunking, one of the more promising approaches proposed in recent years to address the problem of misinformation is "psychological inoculation" or "prebunking," which seeks to prevent people from encoding misinformation in the first place.[8]

Please cite this article in press as: van der Linden and Cohen, The neuroscience of misinformation: A research agenda, Neuron (2025), https://doi.org/10.1016/j.neuron.2025.05.010

Consistent with the metaphor of biological inoculation, people are exposed to weakened doses of manipulation techniques that underpin popular misinformation (e.g., fearmongering, scapegoating) in a controlled environment, aiming to train recognition of and resistance to such techniques when they are encountered in their full form later in the real world. These methods are a modernization of techniques originally developed by social psychologists in the 1960s and include entertaining gamified interventions (e.g., "Bad News") as well as high-production-value videos for social media. These preemptive interventions have been adopted by public health authorities such as the World Health Organization (WHO) and scaled to social media by technology companies such as Google. For example, in one real-world field study, prebunking videos were placed in the ad spaces on YouTube, reaching hundreds of millions of people.[1,8] Although research has uncovered key mechanisms behind inoculation's effectiveness—including memory and motivation—this area of research has generally not drawn upon insights from neuroscience. We propose several ways in which neuroscience methods could be used to better understand and enhance the effectiveness of these interventions. Naturalistic neuroimaging methods could help measure engagement as a precursor of intervention effectiveness. The degree to which interventions change the brain's response to information intended to mislead could also be measured to demonstrate effectiveness. Finally, following recent work showing that better memory for interventions enhances their long-term effectiveness, using emotion to enhance consolidation of memory for interventions could help make their impact more persistent.

Measuring brain activity to optimize inoculation follows rich recent literatures in neuroforecasting and neuroeconomics.[5,9] Brain responses within small, scanned samples can predict choice preferences and effectiveness of persuasive messaging among the larger population more precisely than behavioral measures.[5,9] These "brain-as-predictor" studies point to the involvement of rapid implicit processes that are best assessed using neuroscience methods. For instance, the response to antismoking public service announcements in brain regions, including the mPFC, predicts how effectively those advertisements motivate interest in actually quitting smoking among larger populations. Reward responses, e.g., to crowdfunding campaigns and to YouTube videos, also reliably predict popularity in the real world.[9] These findings in the aggregate imply that reward and mentalizing responses in the brain can be useful for prediction of preferences outside of a scanned sample. Accordingly, we can assume that the degree to which an inoculation intervention reduces the response to false information in reward and mentalizing brain systems (or increases the response in these brain regions toward targets of derogatory false information) would predict the intervention's real-world effectiveness.

A further relevant advance is the use of naturalistic neuroimaging methods. Intersubject correlations (ISCs) can be used to identify the degree to which brain regions are engaged by video or auditory stimuli. One recent study showed that ISCs in response to Islamic State (ISIS) terrorist propaganda videos predicted ratings of persuasiveness.[10] A double dissociation in brain activity was also observed, such that ISCs in the nucleus accumbens predicted persuasiveness of videos with messaging expected to evoke rewards to oneself, while ISCs in mentalizing regions (especially the dorsal mPFC) predicted persuasiveness of videos with messaging that evoked benefits to one's community. These results imply that similar methods could be applied to evaluate mechanisms of action to optimize inoculation interventions.

Some emerging work has focused on resistance to persuasion, which is a key goal of inoculation interventions. This mechanism is likely mediated by activation of the dorsolateral prefrontal cortex (dlPFC) and other regions in the brain's frontoparietal control network. Activity in the dlPFC has been associated with resistance to antidrug persuasive messages among those at risk of drug use, i.e., who are motivated to rationalize against the messaging. We might expect inoculation to stimulate similar mechanisms toward resisting persuasion when people are exposed to false content, providing another possible dimension along which to optimize these emerging psychological interventions. Lastly, inoculation interventions could be combined with other ways of combatting misinformation that have shown some promise, such as chatbots, promoting actively open minded thinking, accuracy nudges, social norm incentives, and media literacy training.[1] Understanding brain responses to these interventions will help optimize their implementation and effectiveness.

Conclusion

We have shown that cognitive, social, affective, and behavioral neuroscientific findings have a crucial role to play in elucidating the brain mechanisms behind why people are susceptible to misinformation, how people are conditioned online to share misinformation with others, and, finally, how neuroscience can help inform intervention efficacy by examining brain responses that predict real-world engagement with interventions and their targets. Neuroscience also has an important role to play in furthering our understanding of how social media algorithms leverage affective, false, and extremist content to manipulate the brain's reward system and keep people engaged on their platforms. Improved understanding of these neural mechanisms could help guide the development of relevant policies to hold technology companies accountable for the spread of misinformation. Our hope is that consideration of the brain mechanisms that draw people to false information and those activated by debunking and prebunking misinformation will enhance efforts to combat the miasma of propaganda that threatens the future of our democracies.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. van der Linden, S., Albarracín, D., Fazio, L., Freelon, D., Roozenbeek, J., Swire-Thompson, B., and Van Bavel, J. (2023). Using Psychological Science to Understand and Fight Health Misinformation: An APA Consensus Statement (American Psychological Association). https://www.apa.org/pubs/reports/misinformation-consensus-statement.pdf.

2. Ecker, U.K.H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L.K., Brashier, N., Kendeou, P., Vraga, E.K., Amazeen, M.A., and Amazeen, M.A. (2022). The psychological drivers of misinformation belief and its resistance to correction. Nat. Rev. Psychol. *1*, 13–29.

3. Wang, W.-C., Brashier, N.M., Wing, E.A., Marsh, E.J., and Cabeza, R. (2016). On known unknowns: Fluency and the neural mechanisms of illusory truth. J. Cogn. Neurosci. *28*, 739–746.

4. Sharot, T., and Sunstein, C.R. (2020). How people decide what they want to know. Nat. Hum. Behav. *4*, 14–19.

5. Falk, E., and Scholz, C. (2018). Persuasion, influence, and value: Perspectives from communication and social neuroscience. Annu. Rev. Psychol. *69*, 329–356.

6. Lindström, B., Bellander, M., Schultner, D.T., Chang, A., Tobler, P.N., and Amodio, D.M. (2021). A computational reward learning account of social media engagement. Nat. Commun. *12*, 1311.

7. Cohen, M.S., Halewicz, V., Yildirim, E., and Kable, J.W. (2024). Continued influence of false accusations in forming impressions of political candidates. PNAS Nexus *3*, 490.

8. van der Linden, S. (2024). Chapter One - Countering misinformation through psychological inoculation. In Advances in Experimental Social Psychology, B. Gawronski, ed. (Elsevier), pp. 1–58.

9. Knutson, B., and Genevsky, A. (2018). Neuroforecasting Aggregate Choice. Curr. Dir. Psychol. Sci. *27*, 110–115.

10. Cohen, M.S., Leong, Y.C., Ruby, K., Pape, R. A., and Decety, J. (2024). Intersubject correlations in reward and mentalizing brain circuits separately predict persuasiveness of two types of ISIS video propaganda. Sci. Rep. *14*, 13455.