Social feedback mechanisms & misinformation: A neuroscience-based argument for algorithm regulation

Authors

Michael S. Cohen
Jean Decety
University of Chicago, Chicago, IL, USA

Abstract

Psychological and neuroscience research reveals how motivational systems in the brain interact with the structure of social media platforms to enhance the propagation of misinformation. Features such as likes activate the brain's reward system, encouraging people to share more emotionally engaging content, especially that which provokes outrage against opposing groups. False information is more likely to evoke outrage than is truthful information. Furthermore, exposure to misinformation increases its perceived credibility, and later corrections often fail to override its influence. To reduce impacts of false information, regulation of social media platforms should focus partly on the algorithms that prioritize content. Additionally, investing in media literacy interventions and grassroots efforts can bolster people's motivation and ability to resist misleading content.

Corresponding author:
*Michael S. Cohen, University of Chicago, 5848 S. University Ave., Chicago, IL 60637, USA
Email: mscohen@uchicago.edu*

Keywords
social media, dopamine, Digital Services Act, morality, polarization, cognition

Behavioral Science & Policy
1–9
© Behavioral Science & Policy Association 2026
Article reuse guidelines: sagepub.com/journals-permissions
DOI: 10.1177/23794607251403323
journals.sagepub.com/home/bsx

Misinformation, defined as "any information that is demonstrably false or otherwise misleading, regardless of its source or intention,"[1] is a critical global problem today because of its role in distorting reality among the public. False and misleading information is not new and has long been used in political contests, warfare, and economic competition. But social media, currently estimated to reach 5.4 billion users,[2] allows misinformation to spread much further and faster.

Research on misinformation has expanded rapidly since 2017, likely due to the assumed role of false information in the 2016 U.S. presidential election and in the Brexit referendum the same year.[3] Opinion polls suggest that a sizable number of voters from both U.S. political parties believe in conspiracy theories that spread via social media. For instance, a 2024 poll showed that 34% of Democratic voters believed the implausible conspiracy theory that Donald Trump staged his own attempted assassination.[4] Still, there are partisan asymmetries leading to a more severe problem among conservatives. Since 2016, Republican elites have spread far more false information than Democratic elites.[5] Furthermore, Republican voters are more likely to believe false information about consequential matters like public health. A January 2025 poll found that 40% of Republicans believed that it is probably or definitely true that more people have died from COVID-19 vaccines than from the virus itself, up from 25% in 2023.[6]

Similarly, 35% of Republicans in an April 2025 poll reported believing that the MMR vaccine causes autism.[7] More problematic still is that these beliefs align with policymaking by Republicans that contradicts expert consensus. Specific recent examples in U.S. federal government policy include limiting access to COVID-19 vaccines, funding research on the already-discredited belief linking vaccines to autism, and impeding development of future mRNA vaccines.[8] Similar policy changes are also happening at the state level, where legislation weakening longstanding evidence-based public health protections on vaccines, milk pasteurization, and water fluoridation has been enacted in 12 U.S. states.[9]

At the same time, the ability of researchers to study the content flowing through actual social media platforms has been hampered in the last few years by changes such as the 2023 price increase that made the application programming interface for Twitter, now known as X, unaffordable for most academic users[10] and by the shutdown of Meta's CrowdTangle tool in 2024.[11] Naturally, platform transparency can yield important insights,[12] and we do not disagree about the value of such real-world data. However, many of the studies referenced in the current review were laboratory based and relied on simulated social media environments. This approach does not require cooperation of social media platforms, allows for greater experimental control, and can enable using neuroimaging methods to identify brain mechanisms. Data from simulated environments can ideally complement real-life social media data to improve understanding of the neural, social, and cognitive mechanisms that underlie people's vulnerability to false information. We review some of these insights next and use them to help inform actionable recommendations for combatting false information.

Reward Processing

Liking on Social Media

The brain system that processes reward and motivates behavior, via interconnected regions such as the ventral striatum and ventromedial prefrontal cortex, plays a key role in processing positive feedback on social media. A series of studies showed this brain response using a simulated version of Instagram.[13,14] Participants provided some of their own Instagram posts, which were then presented in an MRI scanner alongside other images chosen by the experimenters. Participants were told that a group of their peers had chosen whether to like each image. In reality, the like count for each image was randomly assigned to be either relatively high or relatively low. When one's own images had a high number of likes, there was a broad increase in neural activity in the reward system, as well as in brain regions involved in self-relevance and processing social information. Providing likes to posts ostensibly from peers in the same task paradigm also led to reward system activation.[15] Overlap between giving and receiving likes was apparent in core reward-processing regions of the brain, further emphasizing that in the social media environment, the reward system similarly motivates both providing and receiving positive feedback.

Sharing

Reward mechanisms also play a central role when people decide to share online content. One key study in this domain demonstrated that *New York Times* news articles most likely to be shared among the general population engaged the reward system more strongly in a small group of participants tested in an MRI scanner.[16] Relationships were also observed between real-world sharing and activity in brain systems that process both one's own and others' mental states, but these activations were statistically mediated by reward processing. In other words, articles that stimulated either thinking about what other people would be thinking (mentalizing) or thinking about oneself led to sharing to the extent that this activation led to increased activity in the brain's reward circuit. These findings suggest that the neural reward response is most directly related to virality of content.

Novelty

There are various pathways by which misinformation could lead to a greater reward signal than true information, thereby motivating sharing of false content. One seminal study found that misinformation, especially in the political domain, spreads more quickly and to more people on social media than true content, based on a large sampling of Twitter data up to 2017.[17] The false content also tended to be more novel, salient, and surprising than the true content. These results can be considered alongside human and animal neuroscience research indicating that the brain's reward system is sensitive to novelty.[18,19] Thus, novelty may contribute to the spread of false information via its tendency to stimulate a reward response in the brain.

Reinforcement Learning and Habit Learning

Other mechanisms by which reward motivates behavior, specifically reinforcement learning and habit learning, also predict sharing on social media. Real-world data from Instagram show that getting more positive feedback on one post is associated with making a new post more quickly.[20] This relationship is consistent with reinforcement learning, where positive or negative feedback affects motivation for subsequent actions. In the brain, reinforcement learning is known to depend on dopamine-releasing neurons in the reward system.[21] However, this relationship does not hold for individuals with higher posting frequencies and more habitual posting styles.[22] Here, posting appears to be a habit, that is, a learned response to cues, initially motivated by reward, that persists even when the cues no longer consistently predict reward.[23] Sharing becomes habitual when repeated decisions to share content, and receiving positive social feedback for doing so, lead people to share even without positive feedback.

A series of studies using a simulated social media ecosystem further tied habitual sharing with sharing of misinformation.[24] Specifically, people who posted more frequently and/or habitually on their real Facebook accounts shared more stories overall in the simulated environment. Moreover, a higher proportion of the stories that they chose to share were false. However, when participants were given monetary incentives for sharing accurate information, the ratio of true to false stories among habitual sharers improved. Creating a new, healthier habit seemed to have helped people unlearn the habits of indiscriminate sharing that, at least in some, had been trained by real social media platforms.

Motivating Accuracy

The reward system has also been invoked in suggesting that motivation to be accurate competes with other motivations such as group identity.[25] Indeed, attention to accuracy can be improved using subtle accuracy nudges, leading people to share a higher proportion of true content.[26] Other changes in the design of simulated social media environments can also motivate a focus on accuracy. For instance, in one study, people given feedback indicating trust or distrust rather than like or dislike tended to share more accurate content in future posts.[27] Feedback from others indicating that a post is misleading can also make people less likely to share such posts, even for individuals and topics that are highly politically polarized.[28] Thus, social media platforms could be modified from their current structure to increase the motivation to be accurate.

Distinguishing Wanting From Liking

A key takeaway from this body of work is that when social media algorithms are designed to maximize engagement, such as likes and shares, they probably optimize the degree to which the social media environment stimulates the brain's reward system. These signals motivate behavior in a way that goes beyond any reasoned choices a person might make. In fact, neuroscience research has documented a distinction between wanting and liking. The mesolimbic dopamine system, often considered the core reward system, motivates animals to approach or want a certain object or response option even when they have no experience with it being pleasurable.[29] The liking system that encodes pleasure relies on a partially overlapping but neurobiologically distinct system mediated by opioid neurotransmitters.[30]

The finding that wanting (but not necessarily liking) drives behavior is considered a model for addictive behavior in both humans and animals, and indeed, this neural mechanism has been directly associated with social media addiction.[31] Here, we propose that wanting drives engagement on social media, and thus is reinforced by typical social media algorithms. Posts that generate positive emotions (liking) would only drive engagement and algorithmic prioritization to the extent that they also stimulate wanting. This distinction may help explain why accurate, uplifting, and thoughtful content that people believe should go viral (which hypothetically stimulates mostly liking) is disadvantaged relative to the provocative and incendiary content, including misinformation, that people correctly believe actually does go viral, likely by stimulating wanting.[32] How social media engages these core motivational systems of the brain is therefore likely to play a central role in the propagation of misinformation.

Socioemotional and Moral Motivations

Psychological drivers of the spread of misinformation also include moral motivations and group identity. These factors plausibly contribute to the motivational force that activates the reward system, as discussed in the previous section.

Morality

Referencing of moral topics is associated with the spread of content on social media. One key early study found that posts on Twitter about contentious political topics were more likely to be shared when they expressed greater levels of moral–emotional language: words with both moral and emotional connotations.[33] Another study found that when people were shown headlines about contentious topics, there was a bias toward sharing the headlines that agreed with their own prior point of view (dubbed *myside bias*). This effect was stronger for topics that the person defined as being of absolute moral importance and for attitudes that were more extreme, whether the headline was true or false.[34]

Other work has shown that the moral framing of a headline similarly affects sharing decisions. Specifically, participants were more likely to share headlines when the framings of those headlines matched their own values with regard to moral foundations, and this effect was particularly robust for false headlines.[35] This study primarily relied on a carefully crafted set of headline framings that the authors could vary systematically, presented in a simulated social media environment. They were then able to largely replicate key findings (though with less conceptual precision) in an accompanying analysis of real social media data.[35] Together, these studies suggest that people are highly motivated to share content that reinforces their moral values, even when it is false.

Outrage

Outrage is one moral–emotional response that appears particularly relevant to the spread of misinformation. Research has shown that content from internet domains that tend to contain misinformation, including those associated with the Russian government's Internet Research Agency disinformation shop, elicited more angry reactions on Facebook and more outraged language in responses on Twitter.[36] Importantly, such expressions of outrage were associated with increased sharing of both true and false content, with the effect being even stronger for false content.

Another study found that positive social feedback on Twitter in response to moral outrage reinforced the level of outrage in subsequent posts, consistent with reinforcement learning.[37] Individuals were also more likely to express outrage when other members of their social networks did so, consistent with a separate mechanism known as *norm learning*. These authors also obtained causal evidence for the effect of norm learning using a simulated social media paradigm. Participants randomly assigned to a social network where outrage expressions were more common, then asked to choose between sharing a post with high levels of outrage or a more neutral post, were more likely to choose the outraged post compared to those in a network where others' posts did not contain outrage. Together, these findings strongly suggest that social media algorithms that are maximally optimized for engagement are likely to amplify outrage and, in turn, increase the spread of misinformation.

Group Identity

A related dimension that motivates sharing is the reinforcement of one's group identity. This motive often combines with outrage, creating a powerful driver of content sharing. One analysis of real social media posts by partisan news sources and U.S. politicians found that references to political outgroups (notably, the opposition party) were linked to increased sharing on Twitter and Facebook among both Republicans and Democrats.[38] These posts generally expressed negative emotions toward the opposing side, possibly including outrage. Another study linking individual survey data with sharing behavior on Twitter found that people who were more strongly partisan showed a greater preference for sharing news supporting their side, regardless of whether the news was true or false.[39]

Neural Mechanisms of Moral Conflict

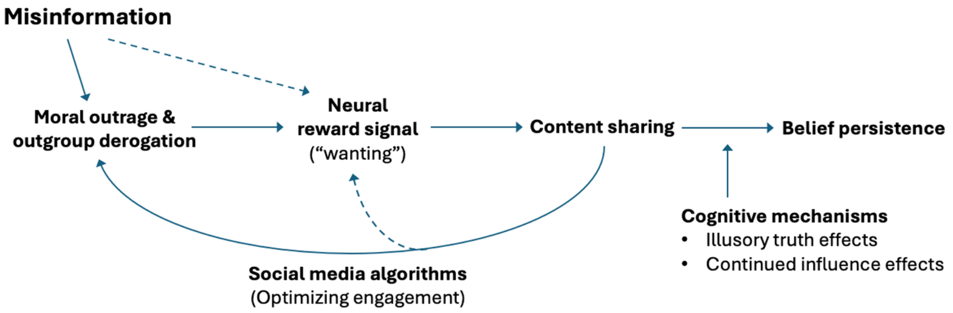
While the above studies did not directly examine brain mechanisms, there is neuroscience research showing that the reward system is stimulated when people envision making aggressive responses motivated by strong moral disagreement. One particularly interesting example is a neuroimaging study of supporters of a terrorist group affiliated with al-Qaeda, who were asked to evaluate their willingness to fight and die for a range of political causes. Higher ratings of willingness to fight and die for sacred causes were associated with increased activity in the ventromedial prefrontal cortex, a core region for reward-motivated behavior that encodes subjective value.[40] Other work has shown that even among people in the United States who were asked to consider hypothetical violent protests, activity in the reward system predicted ratings of appropriateness and moral relevance more strongly for protests on one's own side.[41]

The reward system does not seem to process evaluation of moral issues in all contexts, however. Another recent study asked participants to choose which of two groups of (peaceful) protesters they supported more. Findings showed reward system activation associated with average support for the two groups but not with the moral relevance of the decision.[42] These findings imply that it is specifically moral outrage against opposing groups that activates a reward response, not all moral content.

Algorithmic Prioritization

The findings reviewed in this section suggest a potential feedback loop whereby content that evokes moral outrage (especially false content) activates the reward system, prompting greater user engagement. This engagement drives algorithmic prioritization, which in turn reinforces the posting of similar outrage-inducing material (see Figure 1). These dynamics likely amplify the spread of misinformation and contribute to polarization more broadly. This suggests that social media platforms may need to consider reducing the reach of content that provokes moralized outrage against opposing groups. Of course, the benefits of reducing the reach of inflammatory content should be balanced against the beneficial role that this type of content can play in motivating social change and against users' apparent desire to engage with such content. Still, the risks posed by its unchecked circulation are too significant to simply ignore.

Figure 1. Hypothesized model by which social media algorithms reinforce false and polarizing content


Misinformation
Moral outrage & outgroup derogation
Neural reward signal ("wanting")
Content sharing
Belief persistence
Cognitive mechanisms
• Illusory truth effects
• Continued influence effects
Social media algorithms (Optimizing engagement)

Note. This model integrates the research reviewed in this article. Dashed lines leave open the possibility of other stimulus features besides outrage and outgroup derogation that could motivate reward signals.

Cognitive Mechanisms

Illusory Truth

A separate body of research from cognitive psychology helps explain why the effects of false information are so persistent and highlights the importance of reducing one's initial exposure. A well-known quote often attributed to Nazi propagandist Joseph Goebbels is that one can "repeat a lie often enough and it becomes the truth."[43] This phenomenon has been studied in cognitive psychology experiments dating to the 1970s,[44] and in this literature is referred to as the *illusory truth effect*. The initial work examined truth in the context of trivia questions where people did not have preexisting knowledge of the correct answers. However, illusory truth effects also occur even with statements that participants can later identify as incorrect.[45]

Prior exposure also reduces the ability to detect false political news headlines, showing the pervasiveness of illusory truth effects across various contexts. In one study that used actual fake news headlines seen on Facebook, even a single exposure increased subsequent perceptions of accuracy, both within the same session and after a week. The illusory truth effect held even for stories labeled as contested by fact-checkers or that were inconsistent with the reader's political ideology.[46]

Another study set in a simulated social media environment found that people were more likely to share false content after being previously exposed to it.[47] A statistical mediation analysis showed that prior exposure led to an increase in perceived accuracy and that this increase fully accounted for the subsequent increase in content sharing after prior exposure.[47]

These findings indicate that illusory truth effects likely contribute to the belief in and spread of misinformation both inside and outside the laboratory. However, some research has shown that making people aware in advance that certain statements may be false can reduce illusory truth effects.[48] Thus, educating people on how to approach media content more skeptically may help reduce the impact of false information.

Continued Influence Effects

Another cognitive phenomenon that makes people vulnerable to misinformation is *continued influence effects* (CIEs). CIEs describe the difficulty of fully correcting an initial impression formed by false information, even when a clear factual refutation is provided. One classic approach for studying CIEs involves telling participants that flammable chemicals contributed to the spread of a fire. Later, they are told that this information was false. Nevertheless, participants are much more likely to cite flammable chemicals as contributing to the fire if they are exposed to this incorrect information.[49]

More recent work has shown that CIEs are similarly robust when negative information about political candidates is presented but then factually refuted.[50] Emotional arousal triggered by accusations also appears to lead to prioritized processing for the accusations compared to factual refutations. This emotional prioritization may cause accusations to have more impact on decision-making about candidates than less-arousing refutations. As with illusory truth, corrections do appear to be more effective if people are suspicious of the false information before ever being exposed to it.[51] Still, once false information has spread, factual corrections are unlikely to eliminate its effects on subsequent beliefs.

Importance of Preemptive Refutation

The studies reviewed in this section suggest that once misinformation takes hold, it is difficult to fully rebut its influence, even in neutral contexts such as trivia questions or causal reasoning narratives. This persistence shows how deeply vulnerability to false information is grounded in cognition. It is crucial for the media to recognize that false narratives can exert powerful and pervasive negative effects even when accurate information is presented alongside them. News organizations should use this knowledge to adopt a more deliberate approach about what narratives they prioritize. When likely false stories are newsworthy and cannot be ignored entirely, these organizations must make clear in their own voice—and before conveying the story—that there is reason to be skeptical of what is being presented. Overall, this body of research makes clear that the spread of false content via either mass media or social media can be harmful, even if the truth is readily accessible.

Overall Policy Recommendations and Conclusions

If policymakers have a goal of reducing the spread and impact of misinformation, social media platforms cannot be left unregulated. Evidence suggests that social media executives are aware of this challenge. For example, both Facebook and Twitter/X previously implemented content moderation policies and algorithmic adjustments to label or limit the reach of false and polarizing content. However, the platforms have since reversed these measures.[52,53] There is evidence that these policies can be effective; for instance, deplatforming of influencers posting extremist content on Twitter reduced harmful posts from those who remained.[54] However, Meta CEO Mark Zuckerberg primarily justified his company's policy change with his belief that society currently values the right to free expression over regulation of harmful content.[55]

Regulation of Content Algorithms

A major theme that emerged from our model (see Figure 1) is the urgent need to regulate social media algorithms. Other reviews have drawn similar conclusions based on research from social psychology,[56,57] but our review is unique in highlighting the role of the neural reward system. That is, social media algorithms tend to prioritize content that people "want" (see Reference 29), even if it makes them unhappy and/or harms society.

Based on our framework, future research could more precisely characterize the content (especially false content) that motivates wanting in the brain despite causing harm to individuals and/or society. Regulatory bodies could then target such content to require its deprioritization in algorithmic recommendations. Recent laws, particularly the European Union's Digital Services Act, provide a good starting point for such regulation. Articles 34–35 of this law require audits of algorithms for systemic risks to society and require efforts to mitigate those risks. The law also provides legal mechanisms to compel cooperation from platforms.

In contrast to Europe, it seems unlikely that federal regulations addressing these issues will be enacted in the United States in the near term. Republican politicians and voters consistently express concern that regulation against misinformation targets them asymmetrically.[58,59] However, in a study that applied systematic and unbiased determinations of problematic content, those who supported Donald Trump shared more such content, demonstrating that asymmetries in enforcement are unavoidable.[60] Still, the perception of bias is a hurdle that must be overcome.

U.S. regulations against false information will also need to account for the strong Constitutional protections against government regulation of speech in the First Amendment. Some regulation of harmful speech has been allowed by the U.S. Supreme Court, such as of speech causing "imminent lawless action,"[61,62] but not of speech that is simply false.[63] It is unclear whether the contribution of false social media content to extreme political polarization constitutes sufficient harm to allow regulation. One advantage of regulating algorithmic prioritization is that such regulation would likely face fewer First Amendment obstacles than regulating specific false content.

Another difficult issue is efforts by the current U.S. government to impede other countries' enforcement of laws regulating social media platforms.[64–66] We believe the scientific evidence reviewed here supports the premise that social media can create systemic risks and urge European regulators to be undeterred from enforcing their laws. Despite the considerable power wielded when U.S.-based social media companies align with U.S. governmental resistance to regulation, democracies cannot afford to allow unchecked exploitation of humans' psychological vulnerabilities.

Strengthening Individual Resistance

Beyond government regulation, policymakers can also support efforts to increase citizens' capacity and motivation to resist false and manipulative content. One promising approach toward this goal is psychological inoculation.[67] These interventions include games and videos that expose people to manipulation techniques in an exaggerated, humorous context. The goal is to strengthen psychological defenses against such content in the real world. Other work has reviewed a wider range of interventions beyond inoculation that can improve resilience to misinformation on an individual level.[68]

Another approach to motivating resistance is using emotional appeals. One recent example is the case of Paloma Shemirani, a young British woman who refused conventional cancer treatments and later died. Paloma's mother was an avid conspiracy theorist who persuaded her daughter to pursue ineffective alternative treatments rather than the chemotherapy that doctors estimated would have had an 80% chance of success.[69] Grassroots activism could unite loved ones of those harmed by disinformation to encourage them to share their stories publicly.

Concluding Comments

The psychological vulnerabilities we discussed here are not inherently harmful. In many contexts, they promote prosocial behavior and cooperation. The danger arises when they interact with the modern social media environment, which tends to sway people toward beliefs that can pose serious personal risks as well as risks to others.

In a world facing a complex set of serious threats, including those from climate change, pandemic diseases, terrorism, and autocratic regimes gaining influence, leaders must be able to come together to solve problems. However, solutions are made more difficult when social media amplifies false and polarizing information, as such environments can contribute to false beliefs taking hold among wide swaths of the electorate. Given that shifts in public opinion tend to influence policymaking,[70] representatives may feel pressure to act based on these false beliefs and against making the compromises between groups that are necessary to solve problems.

To limit these harmful impacts, we recommend that governments implement and enforce regulations of social media platforms in ways that are informed by psychological science. At the same time, efforts must be made to enhance people's recognition of and resistance to manipulative content. Without such measures, societies remain highly vulnerable to destabilization.

Summary of Major Scientific Points & Policy Recommendations

Scientific Points

- Liking and sharing content on social media is associated with activation of the brain's reward system.
- Misinformation and polarizing content may stimulate a greater reward response by activating outrage, novelty, moral values, and group identity.
- Content that stimulates a reward response motivates behavior but is not necessarily pleasurable.
- These dynamics lead people to engage more with false and polarizing content, which would cause algorithms optimized purely for engagement to prioritize such content.
- Once people are exposed to false information, this information can have subtle but persistent impacts even when they are unaware that it is still affecting their beliefs.

Policy Recommendations

- Limiting the spread of false information preemptively is important.
- Social media algorithms should be subject to regulation to limit the spread of false and polarizing content. The European Union's Digital Services Act provides a useful model for this type of regulation.
- Investing in public service messaging to enhance citizens' motivation and capacity to resist manipulation by online content is also important.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Michael S. Cohen iD https://orcid.org/0000-0002-0317-7050
Jean Decety iD https://orcid.org/0000-0002-6165-9891

References

1. van der Linden, S., Albarracín, D., Fazio, L., Freelon, D., Roozenbeek, J., Swire-Thompson, B., & 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. Kepios. (n.d.). *Global social media statistics*. Datareportal. Retrieved October 8, 2025, from https://datareportal.com/social-media-users
3. Camargo, C. Q., & Simon, F. M. (2022, September 20). Mis- and disinformation studies are too big to fail: Six suggestions for the field's future. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-106
4. Kerr, A. (2024, July 18). BlueAnon rears its head: One-third of Dems believe conspiracy theory that Trump staged assassination attempt. *Washington Free Beacon*. https://freebeacon.com/democrats/blueanon-rears-its-head/
5. Lasser, J., Aroyehun, S. T., Simchon, A., Carrella, F., Garcia, D., & Lewandowsky, S. (2022). Social media sharing of low-quality news sources by political elites. *PNAS Nexus*, *1*(4), Article pgac186. https://doi.org/10.1093/pnasnexus/pgac186
6. Kearney, A., Sparks, G., Hamel, L., Montalvo, J., III, Valdes, I., & Kirzinger, A. (2025, January 28). *KFF Tracking Poll on Health Information and Trust: January 2025*. Kaiser Family Foundation. https://www.kff.org/health-information-trust/poll-finding/kff-tracking-poll-on-health-information-and-trust-january-2025/
7. Montero, A., Sparks, G., Montalvo, J. III, Kirzinger, A., & Hamel, L. (2025, April 23). *KFF Tracking Poll on Health Information and Trust: The public's views on measles outbreaks and misinformation*. Kaiser Family Foundation. https://www.kff.org/health-information-trust/kff-tracking-poll-on-health-information-and-trust-the-publics-views-on-measles-outbreaks-and-misinformation/
8. Santhanam, L. (2025, September 26). 12 ways RFK Jr. has undercut vaccine confidence as health secretary. *PBS News*. https://www.pbs.org/newshour/health/12-ways-rfk-jr-has-undercut-vaccine-confidence-as-health-secretary
9. Smith, M. R., & Ungar, L. (2025, October 22). Anti-science bills hit statehouses, stripping away public health protections built over a century. *The Associated Press*. https://www.ap.org/news-highlights/spotlights/2025/anti-science-bills-hit-statehouses-stripping-away-public-health-protections-built-over-a-century/
10. Gotfredsen, S. G. (2023, December 6). Q&A: What happened to academic research on Twitter? *Columbia Journalism Review*. https://www.cjr.org/tow_center/qa-what-happened-to-academic-research-on-twitter.php
11. Gotfredsen, S. G., & Dowling, K. (2024, July 9). Meta is getting rid of CrowdTangle—and its replacement isn't as transparent or accessible. *Columbia Journalism Review*. https://www.cjr.org/tow_center/meta-is-getting-rid-of-crowdtangle.php
12. Pasquetto, I. V., Swire-Thompson, B., Amazeen, M. A., Benevenuto, F., Brashier, N. M., Bond, R. M., Bozarth, L. C., Budak, C., Ecker, U. K. H., Fazio, L. K., Ferrara, E., Flanagin, A. J., Flammini, A., Freelon, D., Grinberg, N., Hertwig, R., Jamieson, K. H., Joseph, K., Jones, J. J., . . .Yang, K. C. (2020, December 9). Tackling misinformation: What researchers could do with social media data. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-49
13. Sherman, L. E., Payton, A. A., Hernandez, L. M., Greenfield, P. M., & Dapretto, M. (2016). The power of the like in adolescence: Effects of peer influence on neural and behavioral responses to social media. *Psychological Science*, *27*(7), 1027–1035. https://doi.org/10.1177/0956797616645673
14. Sherman, L. E., Greenfield, P. M., Hernandez, L. M., & Dapretto, M. (2017). Peer influence via Instagram: Effects on brain and behavior in adolescence and young adulthood. *Child Development*, *89*(1), 37–47. https://doi.org/10.1111/cdev.12838
15. Sherman, L. E., Hernandez, L. M., Greenfield, P. M., & Dapretto, M. (2018). What the brain 'likes': Neural correlates of providing feedback on social media. *Social Cognitive and Affective Neuroscience*, *13*(7), 699–707. https://doi.org/10.1093/scan/nsy051
16. Scholz, C., Baek, E. C., O'Donnell, M. B., & Falk, E. B. (2017). A neural model of valuation and information virality. *Proceedings of the National Academy of Sciences*, *114*(11), 2881–2886. https://doi.org/10.1073/pnas.1615259114
17. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559
18. Lisman, J. E., & Grace, A. A. (2005). The hippocampal-VTA loop: Controlling the entry of information into long-term memory. *Neuron*, *46*(5), 703–713. https://doi.org/10.1016/j.neuron.2005.05.002
19. Bunzeck, N., Dayan, P., Dolan, R. J., & Duzel, E. (2010). A common mechanism for adaptive scaling of reward and novelty. *Human Brain Mapping*, *31*(9), 1380–1394. https://doi.org/10.1002/hbm.20939
20. Lindström, B., Bellander, M., Schultner, D. T., Chang, A., Tobler, P. N., & Amodio, D. M. (2021). A computational reward learning account of social media engagement. *Nature Communications*, *12*, Article 1311. https://doi.org/10.1038/s41467-020-19607-x
21. Averbeck, B., & O'Doherty, J. P. (2021). Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology*, *47*, 147–162. https://doi.org/10.1038/s41386-021-01108-0
22. Anderson, I. A., & Wood, W. (2023). Social motivations' limited influence on habitual behavior: Tests from social media engagement. *Motivation Science*, *9*(2), 107–119. https://doi.org/10.1037/mot0000292
23. Pauli, W. M., Cockburn, J., Pool, E. R., Pérez, O. D., & O'Doherty, J. P. (2018). Computational approaches to habits in a model-free world. *Current Opinion in Behavioral Sciences*, *20*, 104–109. https://doi.org/10.1016/j.cobeha.2017.12.001
24. Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, *120*(4), Article e2216614120. https://doi.org/10.1073/pnas.2216614120
25. Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, *22*(3), 213–224. https://doi.org/10.1016/j.tics.2018.01.004
26. Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*, 590–595. https://doi.org/10.1038/s41586-021-03344-2
27. Globig, L. K., Holtz, N., & Sharot, T. (2023). Changing the incentive structure of social media platforms to halt the spread of misinformation. *eLife*, *12*, Article e85767. https://doi.org/10.7554/eLife.85767
28. Pretus, C., Javeed, A. M., Hughes, D., Hackenburg, K., Tsakiris, M., Vilarroya, O., & Van Bavel, J. J. (2024). The misleading count: An identity-based intervention to counter partisan misinformation sharing. *Philosophical Transactions of the Royal Society B*, *379*, Article 20230040. https://doi.org/10.1098/rstb.2023.0040
29. Berridge, K. C. (2023). Separating desire from prediction of outcome value. *Trends in Cognitive Sciences*, *27*(10), 932–946. https://doi.org/10.1016/j.tics.2023.07.007
30. Stark, E., Berridge, K. C., & Kringelbach, M. L. (2022). The neurobiology of liking. In M. Skov & M. Nadal (Eds.), *The Routledge international handbook of neuroaesthetics* (pp. 63–70). Routledge.
31. He, Q., Turel, O., Brevers, D., & Bechara, A. (2017). Excess social media use in normal populations is associated with amygdala-striatal but not with prefrontal morphology. *Psychiatry Research: Neuroimaging*, *269*, 31–35. https://doi.org/10.1016/j.pscychresns.2017.09.003
32. Rathje, S., Robertson, C., Brady, W. J., & Van Bavel, J. J. (2024). People think that social media platforms do (but should not) amplify divisive content. *Perspectives on Psychological Science*, *19*(5), 781–795. https://doi.org/10.1177/17456916231190392
33. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114
34. Marie, A., Altay, S., & Strickland, B. (2023). Moralization and extremism robustly amplify myside sharing. *PNAS Nexus*, *2*(4), Article pgad078. https://doi.org/10.1093/pnasnexus/pgad078
35. Abdurahman, S., Reimer, N. K., Golazizian, P., Baek, E., Shen, Y., Trager, J., Lulla, R., Kaplan, J., Parkinson, C., & Dehghani, M. (2025). Targeting audiences' moral values shapes misinformation sharing. *Journal of Experimental Psychology: General*, *154*(4), 935–957. https://doi.org/10.1037/xge0001714
36. McLoughlin, K. L., Brady, W. J., Goolsbee, A., Kaiser, B., Klonick, K., & Crockett, M. J. (2024). Misinformation exploits outrage to spread

online. *Science*, *386*(6725), 991–996. https://doi.org/10.1126/science.adl2829

37. Brady, W. J., McLoughlin, K. L., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(3), Article eabe5641. https://doi.org/10.1126/sciadv.abe5641
38. Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), Article e2024292118. https://doi.org/10.1073/pnas.2024292118
39. Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*, *115*(3), 999–1015. https://doi.org/10.1017/S0003055421000290
40. Hamid, N., Pretus, C., Atran, S., Crockett, M. J., Ginges, J., Sheikh, H., Tobeña, A., Carmona, S., Gómez, A., Davis, R., & Vilarroya, O. (2019). Neuroimaging ‘will to fight’ for sacred values: An empirical case study with supporters of an Al Qaeda associate. *Royal Society Open Science*, *6*(6), Article 181585. https://doi.org/10.1098/rsos.181585
41. Workman, C. I., Yoder, K. J., & Decety, J. (2020). The dark side of morality—Neural mechanisms underpinning moral convictions and support for violence. *AJOB Neuroscience*, *11*(4), 269–284. https://doi.org/10.1080/21507740.2020.1811798
42. Cao, Q., Cohen, M. S., Bakkour, A., Leong, Y. C., & Decety, J. (2025). Moral conviction interacts with metacognitive ability in modulating neural activity during sociopolitical decision-making. *Cognitive, Affective, & Behavioral Neuroscience*, *25*(2), 291–310. https://doi.org/10.3758/s13415-024-01243-3
43. Stafford, T. (2016, October 26). How liars create the ‘illusion of truth’. *BBC*. https://www.bbc.com/future/article/20161026-how-liars-create-the-illusion-of-truth
44. Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*, 107–112. https://doi.org/10.1016/S0022-5371(77)80012-1
45. Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*(5), 993–1002. https://doi.org/10.1037/xge0000098
46. Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465
47. Vellani, V., Zheng, S., Ercelik, D., & Sharot, T. (2023). The illusory truth effect leads to the spread of misinformation. *Cognition*, *236*, Article 105421. https://doi.org/10.1016/j.cognition.2023.105421
48. Jalbert, M., Schwarz, N., & Newman, E. (2020). Only half of what I’ll tell you is true: Expecting to encounter falsehoods reduces illusory truth. *Journal of Applied Research in Memory and Cognition*, *9*(4), 602–613. https://doi.org/10.1016/j.jarmac.2020.08.010
49. 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. https://doi.org/10.1037//0278-7393.20.6.1420
50. Cohen, M. S., Halewicz, V., Yildirim, E., & Kable, J. W. (2024). Continued influence of false accusations in forming impressions of political candidates. *PNAS Nexus*, *3*(11), Article pgae490. https://doi.org/10.1093/pnasnexus/pgae490
51. Lewandowsky, S., Stritzke, W. G. K., Oberauer, K., & Morales, M. (2005). Memory for fact, fiction, and misinformation: The Iraq War 2003. *Psychological Science*, *16*(3), 190–195. https://doi.org/10.1111/j.0956-7976.2005.00802.x
52. Kopp, A. (2024, October 28). *Two years after the takeover: Four key policy changes of X under Musk*. Zenodo. https://doi.org/10.5281/zenodo.14040407
53. Corse, A., Bobrowsky, M., & Horwitz, J. (2025, January 7). Social-media companies decide content moderation is trending down. *The Wall Street Journal*. https://www.wsj.com/tech/social-media-companies-decide-content-moderation-is-trending-down-25380d25
54. Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (2021, October). Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), Article 381. https://doi.org/10.1145/3479525
55. Nix, N., Oremus, W., & Gregg, A. (2025, January 7). Meta ends fact-checking, drawing praise from Trump. *The Washington Post*. https://www.washingtonpost.com/technology/2025/01/07/meta-factchecking-zuckerberg/
56. McLoughin, K. L., & Brady, W. J. (2024). Human–algorithm interactions help explain the spread of misinformation. *Current Opinion in Psychology*, *56*, Article 101770. https://doi.org/10.1016/j.copsyc.2023.101770
57. Brady, W. J., Jackson, J. C., Lindström, B., & Crockett, M. J. (2023). Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, *27*(10), 947–960. https://doi.org/10.1016/j.tics.2023.06.008
58. Walker, M., & Gottfried, J. (2019, June 27). *Republicans far more likely than Democrats to say fact-checkers tend to favor one side*. Pew Research Center. https://www.pewresearch.org/short-reads/2019/06/27/republicans-far-more-likely-than-democrats-to-say-fact-checkers-tend-to-favor-one-side/
59. Barrett, P. M., & Sims, J. G. (2021). *False accusation: The unfounded claim that social media companies censor conservatives*. NYU|Stern Center for Business and Human Rights. https://bhr.stern.nyu.edu/wp-content/uploads/2024/02/NYUFalseAccusation_2.pdf
60. Mosleh, M., Yang, Q., Zaman, T., Pennycook, G., & Rand, D. G. (2024). Differences in misinformation sharing can lead to politically asymmetric sanctions. *Nature*, *634*, 609–616. https://doi.org/10.1038/s41586-024-07942-8
61. *Brandenburg v. Ohio*, 395 U.S. 444 (1969). https://supreme.justia.com/cases/federal/us/395/444/
62. *Chaplinsky v. New Hampshire*, 315 U.S. 568 (1942). https://supreme.justia.com/cases/federal/us/315/568/
63. *United States v. Alvarez*, 567 U.S. 709 (2012). https://supreme.justia.com/cases/federal/us/567/709/
64. Michaels, D., Gordon, M. R., & Mackrael, K. (2025, May 15). Trump administration targets Europe’s digital laws as a threat to basic rights and U.S. business. *The Wall Street Journal*. https://www.wsj.com/politics/policy/trump-administration-targets-europes-digital-laws-as-a-threat-to-basic-rights-and-u-s-business-20db1016
65. Kroet, C. (2025, September 3). US Congress hearing set to tease tensions with EU over digital rulebook. *EuroNews*. https://www.euronews.com/next/2025/09/03/us-congress-hearing-set-to-tease-tensions-with-eu-over-digital-rulebook
66. Rupar, A. (2025, April 22). *Nina Jankowicz to the European Parliament: “Before I describe the details of Russia’s recent online influence campaigns, I would like to* [Online forum post]. Threads. Retrieved from https://www.threads.com/@aaron.rupar/post/DIwPdBHp3zs/
67. van der Linden, S. (2024). Chapter One—Countering misinformation through psychological inoculation. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. 69, pp. 1–58). Academic Press. https://doi.org/10.1016/bs.aesp.2023.11.001
68. Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Barzilai, S., Basol, M., Berinski, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R., . . . Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, *8*, 1044–1052. https://doi.org/10.1038/s41562-024-01881-0
69. Spring, M. (2025, June 22). Our sister died of cancer because of our mum’s conspiracy theories. *BBC*. https://www.bbc.com/news/articles/crenzwyvpn1o
70. Caughey, D., & Warshaw, C. (2018). Policy preferences and policy change: Dynamic responsiveness in the American states, 1936–2014. *American Political Science Review*, *112*, 249–266. https://doi.org/10.1017/S0003055417000533