REVIEW

From likes to trusts: How trust feedback reduces misinformation on social media

Authors

Gizem Ceylan
The Ohio State University, Columbus, OH, USA

Wendy Wood
University of Southern California, Los Angeles, CA, USA

Corresponding author:
Gizem Ceylan, The Ohio State University, Fisher Hall, 2100 Neil Avenue, Columbus, OH 43210, USA
Email: ceylan.7@osu.edu

Keywords

misinformation, social media, rewards, trust feedback, reinforcement learning, habits

Behavioral Science & Policy
1–11

DOI: 10.1177/23794607261423714
journals.sagepub.com/home/bsx

Abstract

By rewarding engagement over accuracy, social media platforms foster the spread of misinformation. Likes and similar engagement reactions are a central form of feedback on most platforms and shape what users share. On the platforms, users quickly learn that sharing interesting, attention-grabbing content garners positive feedback, even when it is inaccurate. With repeated exposure to these social rewards, sharing interesting content—including interesting but inaccurate content—can become a habit. We review evidence for this reward-based learning and propose a simple redesign of platform rewards: adding a Trust button so users can reward accurate, reliable posts. Experimental evidence supports this approach: Users give Trusts to accurate posts more than inaccurate posts. Then, when they receive trust feedback, users increasingly share accurate content (even when less interesting) and reduce sharing of inaccurate but highly interesting posts. Because this intervention changes incentives, it is scalable, preserves user choice, and aligns with people's stated goal of sharing accurate information. Misinformation interventions that overlook the role of social media incentives are unlikely to produce lasting results.

Despite a variety of attempts to reduce the spread of misinformation, social media remain a conduit for rapid, worldwide distribution of falsehoods.[1] Most interventions proposed to curb misinformation focus on users who post, share, and consume content on social media. Such interventions may remind users to consider accuracy before posting,[2] forewarn them about the flawed arguments typical of false information, a form of psychological inoculation against misinformation sometimes called *prebunking*,[3] or encourage them to conduct additional searches to verify accuracy.[4] However, the long-term effectiveness of such interventions is not promising. Only a few studies have examined the question of lasting impact, and these have often showed declining influence.[5] Thus, we need more enduring approaches to limiting misinformation.

To meaningfully reduce misinformation propagation across social media, we propose modifying the incentives that drive users to share inaccurate content. Our practical policy solution is simple but powerful: Add an incentive for users to post and share information that is accurate instead of merely interesting or attention grabbing. Our research has shown that this simple structural shift can maintain user engagement while reducing misinformation.[6]

Surveys indicate that users value accurate information. In our studies, almost all social media users said that spreading accurate

information is important to them, and that it was more important than sharing information that supports their own political views or that attracts attention and is widely read.[7] Most users are therefore not trying to deceive or mislead others. Instead, they are relying on social media to access news, connect with others, form relationships, establish group memberships, and reduce loneliness.[8]

Yet, users' claims regarding their desire to read and share accurate information clearly conflict with what they share on social media. They continue to spread material that others are likely to find interesting—content that is novel, attention grabbing, or likely to invoke emotional responses[9–11]—even when that information is inaccurate.[12]

What drives such contradictory behavior? Why would people say they value being truthful but still share falsehoods? The answer lies in social media's incentive structure. In a survey of users across multiple social media sites (TikTok, X, Instagram, and LinkedIn), 51% agreed that they would reshare potentially inaccurate content if they believed their network would find it engaging or entertaining.[13] Moreover, 77% of these users stated that interesting and novel posts receive attention regardless of their accuracy, and 90% were confident that sharing inaccurate but engaging content would garner likes and positive comments. Thus, it seems that user intention to share truthful information often conflicts with the rewards that social media offers for sharing attention-getting, engaging material.

This conflict helps to explain why user-focused interventions often fail, as summarized in Table 1: Even when individuals recognize that interesting information is false, sharing it with others can result in positive feedback. Algorithms further amplify attention-grabbing posts that receive multiple likes, stoking a cycle in which the most engaging content, whether accurate or inaccurate, gains even more visibility. We therefore propose that meaningfully reducing misinformation requires reconfiguring the reward system itself.

Reward Structures on Social Media

Rewards on social media were created to draw users to a site, hold their attention, and keep them coming back. Social media sites, including MySpace and Live Spaces, that failed to maintain large numbers of intensive daily users were mostly unsuccessful.[36] Among currently successful sites such as Facebook, Twitter/X, and Instagram, the greater the number of regular users, the higher the advertising revenues.[36] With many frequent users, sites can attract more advertisers and successfully deliver personalized marketing to users. Having occasional, sporadic users does not provide the same financial benefits.

One way that social media fosters repeated and enthusiastic engagement is through rewards—likes, loves, shares, and so on. This feedback, which signals recognition from the community, serves as a powerful driver of user behavior.[37] Research suggests that receiving likes and comments boosts users' happiness, self-esteem, and overall satisfaction with social media use. Likes provide instant social validation, reinforcing the perception that a post is important and worthy of attention.[38] When users share content that generates likes, shares, and comments, they experience a physiological reward response, activating the brain's reward centers[37,39] and boosting release of the reward-associated neurotransmitter dopamine.[40] These rewards are so important that teens and young adults often delete posts that do not generate sufficient positive reactions.[41,42]

The reward systems on these platforms create a cycle of posting or sharing and then receiving feedback. More rewarding feedback leads to greater subsequent use. Specifically, users tend to post more frequently and at shorter intervals after receiving larger numbers of likes on previous posts.[43,44]

What type of content generates high levels of engagement—likes, comments, and shares—on social media? We use the umbrella term *interesting content* to describe information that captures attention, provokes emotional responses, or surprises users with unexpected details. Not all of this content is positive. Research shows that high-arousal, negative content, including misinformation, moral outrage, outgroup animosity, and incivility, is particularly likely to spread widely online. [9–11,45,46] The algorithm-driven nature of social media platforms further amplifies this type of information, as algorithms promote posts that are highly engaging so they can reach even broader audiences. This boost results in a *paradox of virality* in that the content users engage with and share is not the accurate content they claim they want to see.[11]

We argue that social media's current reward structure plays a key role in elevating interesting content. As we explain in the next section, when engaging information gets rewarded, users learn to focus on the interest value of content, often to the detriment of its accuracy and quality.

Learning to Share Misinformation

Sharing others' posts is critical to the spread of misinformation: On Facebook, 38% of views of misinformation and 65% of views of inaccurate photos take

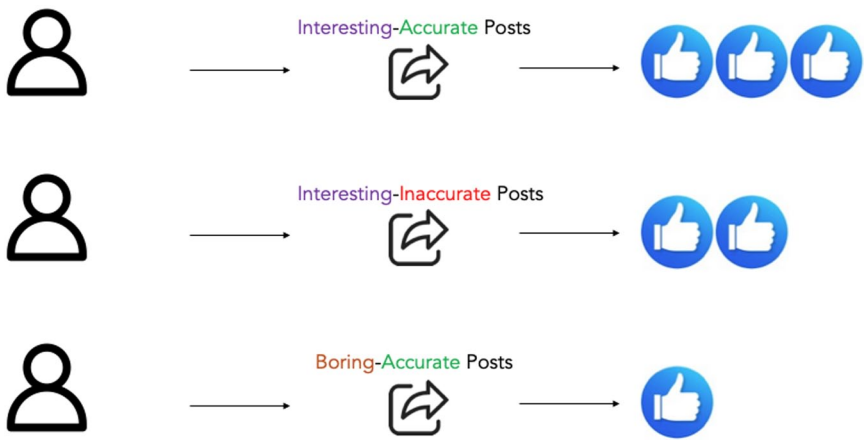
Table 1. Comparison of misinformation interventions

Intervention type		Does it increase accuracy discernment?	Does it limit spread of misinformation?	Are the effects enduring?
User interventions				
Lateral reading & verification	Users trained to evaluate information through cross-checking and search strategies[14–16]	Inconsistent	Largely untested	Largely untested
Inoculation (prebunking)	Users forewarned about flawed online argumentation[17,18]	Yes	Yes, but tested largely in self-reports	Typically decay in weeks; can be reinstated with boosters
Media literacy tips	Users provided with tips for spotting misinformation (e.g., "Check the URL")[19,20]	Yes	Yes, but tested largely in self-reports	Inconsistent
Social norm-based nudges	Users reminded about sharing norms (e.g., others share only true information)[21,22]	Yes	Inconsistent, but tested largely in self-reports	Largely untested
Debunking/ postbunking	Users provided with accurate information and refutations after exposure to misinformation[23,24]	Yes	Largely untested	Largely untested
Prompts to evaluate accuracy	Users reminded to reflect on accuracy before sharing[8,25–27]	Yes	Yes in self-reports, not in behavior	Short (stable across experimental session)
Platform interventions				
Source credibility	Platforms add source credibility ratings to posts (e.g., NewsGuard)[28]	No	Largely untested	Largely untested
Friction on social media use	Platforms add delays to slow reading and sharing[29,30]	Largely untested	Largely untested	Largely untested
Credibility tags and ratings	Platforms append accuracy indicators to posts (e.g., Community Notes)[31,32]	Yes	Largely untested	Largely untested
Content moderation	Platforms remove information, use algorithms to slow spread (virality circuit breaker), or reduce visibility of certain users (shadow banning)[33,34]	Largely untested	Yes	Long lasting
Altering reward structure	Platforms replace like rewards with trust rewards[6,7,35]	No	Yes	Long lasting; strengthens with experience

place after a mere two shares.[47] Each reshare not only amplifies a specific piece of misinformation but also initiates a positive feedback loop that promotes sharing other misinformation. Because reshares generate rewards, they effectively educate sharers on the types of content that attract others. This reward system encourages people to share misinformation through *reinforcement learning*, similar to that demonstrated in B. F. Skinner's studies of pigeons.[48] Just as pigeons learned which behaviors (such as pecking a button) would earn them food rewards, social media users learn which content will garner the social rewards of likes and comments.

Although information that is both interesting and accurate will produce these rewards, users quickly discover that interesting but inaccurate content (misinformation) also generates rewards. Posts with two positive features (interesting + accurate) will typically yield more rewards than posts with a single positive feature (interesting + inaccurate), as shown in Figure 1. And all interesting posts typically get more rewards than content that is accurate but boring, such as simple statements of fact. This reward structure teaches users that the entertainment value of information on social media frequently outweighs its accuracy.

Figure 1. Social media reward structure



With repeated experience, existing reward feedback (post/share → get likes/loves/comments → post/share) trains users to continue posting and sharing the type of content that generated positive reactions in the past. In this feedback loop, information accuracy can become an afterthought.[7] This training explains why even well-intentioned users who claim to value accuracy may gradually start sharing more interesting but potentially less accurate content on social media. This link between interest-based rewards and misinformation has been largely overlooked in past research, given that most misinformation research ignores the interest value of content and focuses solely on its accuracy.[49]

We have demonstrated that social media users learn from financial rewards as well as social likes.[6,7] In our simulations of Facebook, some participants were rewarded for sharing interesting content regardless of its accuracy.[6] Others were rewarded for sharing accurate content regardless of whether it was interesting. Participants who received rewards for sharing interesting content prioritized sharing this type of material because it earned more rewards. At the same time, they also became less discriminating about content accuracy: While these participants increased their sharing of interesting and accurate information, they also shared more interesting misinformation. In contrast, participants rewarded for accuracy learned to prioritize sharing accurate content and did so whether the content was interesting or boring.

Current rewards on social media have negative effects in addition to spreading misinformation, such as fueling moral outrage. Twitter/X users posted more content expressing outrage after they had received positive social feedback for posting such material in the past.[35] Users expressed especially high outrage on days when they received more positive reactions than usual to their past outrage-filled posts.[50] Thus, social rewards shape what users post, and getting more rewards than expected—a positive reward prediction error—is an especially powerful driver of repeated behavior.

Frequency Makes Learning Stick

The more often users receive social rewards for their online behavior, the stronger the learning experience, and the more their behavior becomes shaped by others' reactions. As users continue to frequently share, post, or give likes, these behaviors eventually become automatic.[36] That is, earlier learning based on engagement rewards not only drives responses but also forges habit associations in users' memory when repeated often enough. Once these habits form, responses such as sharing are triggered automatically by context cues (e.g., posts expressing emotion) without much consideration of the consequences, including whether a post will spread falsehoods. Thus, habitual sharers display automated, learned responses rather than intended behaviors: They may spread misinformation without actually intending to deceive others, and they may express outrage without actually feeling outraged.[51]

Direct evidence that sharing becomes habitual comes from studies showing that, after initial learning, misinformation sharing persists even without continued rewards.[7] In one of our studies, after participants learned to expect rewards for

sharing misinformation, we removed this positive feedback. One might expect that these participants, who reported wanting to share accurate information, would then switch to doing just that. Yet, even in the absence of rewards, these participants continued to share more misinformation than others not initially trained to share it. Repeating a response even when outcomes change is a signature of habit formation. In other studies, sharing became more automatic as participants gained experience with the rewards: The more training they received, the faster they shared misinformation.[6]

Repeated exposure to rewards on social media leads users to develop habitual behaviors beyond spreading misinformation. For example, users with a longer history of expressing moral outrage on social media continued to share posts with moral outrage even after others did not respond with positive reactions.[35] Infrequent users, in contrast, were more sensitive to rewards and tempered their subsequent posts when others reacted less positively.

The habits people form from frequent, automatic sharing on social media are an especially strong driver of misinformation spread, even more so than users' political beliefs or critical reasoning skills.[7] Although past research suggests that conservative (compared with liberal) users and those who are less analytically oriented are particularly prone to misinformation sharing,[2,52] we found that these effects were small compared with users' sharing habits (see Figure 2).

These findings could suggest that problematic sharing patterns are limited to perpetually heavy sharers—sometimes called superspreaders—who have developed entrenched habits over time.[55] However, our experimental evidence shows that these reward-driven behaviors can form remarkably quickly with repeated use, even in a single experimental session. Social rewards are powerful motivators in the highly structured social media context, and people rapidly learn to repeat rewarded behaviors. Although individuals use social media in different ways, behaviors that are repeatedly rewarded (posting, sharing, liking, commenting, and even scrolling) soon become habitual.

Understanding habitual sharing of misinformation is crucial because people's sharing habits can undermine corrective interventions. Many past interventions to curb misinformation have missed this important point.[56] Once formed, habits become "sticky," which could explain the limited impact of individual-level interventions aimed at highlighting accuracy goals, improving media literacy and critical thinking, or reducing political bias.[3,57] These individual-focused interventions tend to have short-lived effects and require repeated prompts or "booster shots" to remain effective, given the conflicting pressures of platform incentives.[17,56,58] To the extent that misinformation sharing is habitual, reducing such sharing requires new learning experiences to break the habit.[59,60]

Addressing Misinformation by Realigning Platform Rewards

By rewarding users for sharing interesting information, social media sites enable the spread of falsehoods and partial truths. But this system is not the only way social media platforms can reward engagement. One alternative would be to add a mechanism for recognizing and rewarding users who post and share accurate information. This approach should not only reward participation but also curb the propagation of misinformation.

To examine the effects of changing rewards from interest to accuracy, we included a Trust button—similar to the current Like button—in our Facebook simulations.[6] Participants could then give and receive trusts for accurate content the same way they give and receive likes for interesting content.

Why trusts? Trust captures users' judgment that information is credible, honest, unbiased, and reliably communicated.[61] Users often have a sense of whether content is trustworthy without extensively fact-checking it, drawing on accessible cues such as the source's reputation, the post's context, and whether others have endorsed it.[62,63]

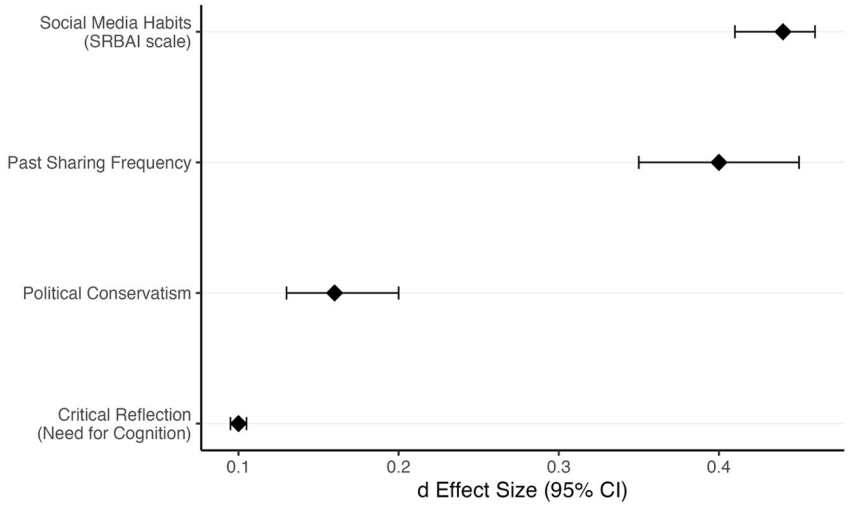
When users evaluate posts on social media, trust reactions from others can serve as both a cognitive shortcut and a social signal. As a cognitive shortcut, trusts appended to a post allow other users to accept information without extensively verifying it, much like trusting a friend's restaurant recommendation eliminates the need to research and read a number of reviews. Such cognitive shortcuts are crucial given the current information overload online.[64] As a social signal, trust feedback communicates to others that information can be believed. Information that receives high trust ratings via multiple Trust button clicks carries a social stamp of approval, indicating that it is widely perceived as credible.

Trust judgments are often made quickly and intuitively, aligning with the split-second decisions users make to engage with or scroll past content on social media.[23,65] This rapid assessment makes trusts uniquely suitable for feedback: They fit the tempo of online behavior while capturing users' intuitive sense of what is reliable.

Evidence That the Trust Button Works

Does adding a Trust button actually curb misinformation, as some have proposed?[49] Our research shows that it does by

Figure 2. Factors influencing misinformation sharing



Note. From Ceylan, Anderson, & Wood (2023), PNAS, 120(4), e2216614120. Licensed under CC BY-NC-ND 4.0. Effect size (*d*) quantifies how much each factor was associated with sharing misinformation.[7] Habits were assessed through both past frequency of sharing information on Facebook as well as self-reported assessments of how automatic users felt their posting behavior had become (the Self-Report Behavioral Automaticity Index, or SRBAI).[53] Lack of reasoning was assessed with the Need for Cognition Scale, which measures an individual's tendency to engage in and enjoy thinking.[54]

incentivizing accurate sharing behavior.[6] We first examined whether participants were more discerning of accuracy when clicking a Trust button as opposed to a Like button (see Figure 3 for an example of a post with its feedback buttons). We found that participants were more likely to give trusts to accurate posts than to inaccurate ones. The Trust button thus served as a marker of a post's accuracy.

Next, we examined whether people's political views biased their use of the Trust button. Trust buttons won't be helpful if they are simply another partisan signal, with users trusting content that aligns with their own political views regardless of its accuracy. In our study, however, truth overcame politics: Giving trusts did not substantially depend on the politics of the posts.[6] Although users tended to like content consistent with their own political views, they were less partisan when using the Trust button. Furthermore, they were 3 times more likely to trust accurate posts even when the information was inconsistent with their own politics. Trusts therefore reflected information reliability, not its likability.

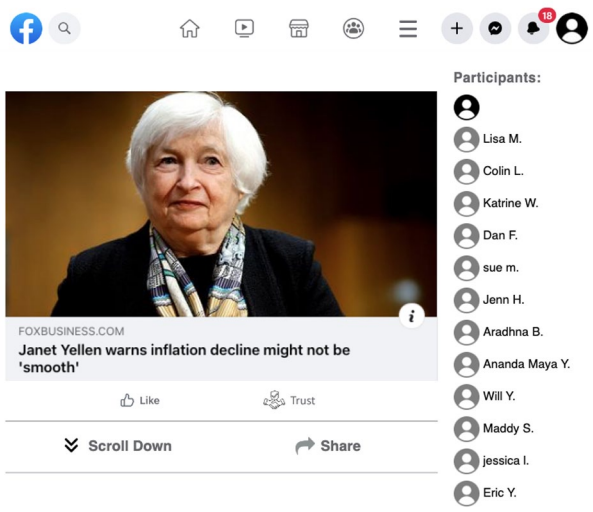
Finally, we examined how receiving trust feedback shaped users' behavior on social media. As we noted earlier, people who received likes increasingly shared interesting posts, even when inaccurate. Participants who received trusts, however, increasingly shared accurate posts—even those that were factual and boring—and they shared fewer inaccurate posts (misinformation), even when the posts were interesting. Thus, trust-based rewards promoted accuracy, even for boring content.

A special advantage of this intervention is that it gets stronger with experience, as people have more opportunities to learn. For example, participants in one experiment chose whether or not to share 16 or 80 posts with others.[6] Recipients of the posts gave trusts when participants shared accurate information. After receiving trust feedback on 80 posts, participants shared more accurate information and less misinformation than those who received feedback on only 16 posts. These findings suggest that trust feedback not only promotes discernment in the short term but can produce stronger effects as people continue to be rewarded for trustworthy content in the long run.

Trust as a Scalable & Timely Solution for Information Quality on Social Media

By providing nuanced, scalable, and socially grounded signals of information quality, adding a Trust button offers a superior alternative to other platform interventions, as

Figure 3. A trust button allows users to provide feedback on perceived accuracy



shown in Table 1. Credibility tags, fact checks, and other ratings can confuse users, leaving them uncertain about a post's accuracy. For example, Facebook, Instagram, and Threads recently added a Community Notes feature so users can provide context to posts. However, this feedback does not necessarily help interpret accuracy.[66] Even more troubling, using content moderation to remove false information from a social media site may be seen as a threat to freedom of speech.[67] In contrast, trust ratings explicitly acknowledge the subjective nature of information evaluation while still providing useful measures of how others judge content reliability. This distinction is particularly important for politically charged topics where objective fact-checking can be challenging.

The Trust button leverages the wisdom of crowds by aggregating judgments that effectively identify misinformation at scale. Research shows that crowd ratings strongly correlate with professional fact-checker assessments.[68] In many cases, aggregated crowd judgments more reliably assess accuracy than those of individual fact-checkers or institutional efforts such as Facebook's Oversight Board.[69] Although the Community Notes feature also relies on collective judgments, the Trust button offers distinct advantages. Community Notes are only posted for consensual judgments.[70] In contrast, the Trust button enables immediate, individual feedback that aggregates into dynamic, scalable signals of reliability. This real-time, self-reinforcing mechanism allows accurate content to gain prominence through social validation, ultimately reshaping what content gets shared and amplified.

As another advantage, the Trust button builds on the existing reward architecture of social media platforms—likes, shares, and reactions—without requiring new content verification or labeling infrastructures. The Trust button does not change the structure of social feedback but instead alters its meaning. Whereas likes signal that content is entertaining or emotionally resonant, trusts signal that content is reliable. By introducing this distinction, platforms can leverage the trust reward's power, allowing accurate but mundane information to receive appropriate recognition while preventing sensational misinformation from gaining credibility through viral spread.

The Trust button also equips users to navigate uncertainty about whether content is true or false. When individuals are unsure, cognitive biases such as judging repeated information as more accurate and thus more shareable often emerge.[65,71] By aggregating judgments of trustworthiness from the broader community, the Trust button reduces this uncertainty and offers a real-time accuracy signal, ultimately curbing misinformation spread.

The rise of AI-generated misinformation makes it increasingly important for social media platforms to adopt mechanisms like Trust buttons that signal reliability. Creating interesting but false content with AI, including deepfakes,[72] is now fast and cheap.[73] Unless accuracy is actively incentivized, this shift threatens to overwhelm platforms with misinformation. The Trust button addresses this risk in three ways: First, it reduces the incentives to produce or share misinformation. If more people reward accurate content and avoid inaccurate content, the potential reach and impact of falsehoods diminish considerably. Second, trusts leverage collective human judgment to identify subtle errors in AI-generated falsehoods that individual fact-checkers or algorithms may miss.[74] Third, trusts create reputational incentives for honest contributors to share verified content. By restructuring incentives to value truth alongside interest, platforms can limit the spread of AI-generated falsehoods.

In addition, the Trust button addresses rising public concern about misinformation. Trusts offer policymakers a science-based intervention that technology companies can easily implement to foster a more transparent, safe, and trustworthy online environment.[75] Importantly, this solution helps platforms protect freedom of speech while also addressing users' growing concerns about information accuracy. According to a recent Pew Research Center poll, 40% of Americans identify inaccuracy as the feature they

most dislike about getting their news on social media, an increase of 9% since 2018.[76] Thus, using a Trust button to curb misinformation would align with public demand for healthier information ecosystems. At the same time, this intervention would help social media companies mitigate reputational risks by demonstrating their proactive commitment to information integrity.

Finally, social media managers and company executives may worry that introducing a Trust button will reduce user engagement. However, our findings suggest this intervention sustains engagement and even enhances it. Positive user experiences, especially those grounded in trustworthy content, may be key to long-term platform vitality. As recent analyses suggest, mainstream social media suffers from declining user satisfaction and engagement, likely driven in part by misinformation and polarization.[77] By realigning incentives toward accuracy rather than interest, platforms can cultivate both engagement and information quality, rather than compromising one for the other.

Reddit as a Case Study for Changing Reward Structures

Although Globig et al. first proposed the idea of a Trust button to address misinformation in 2023,[49] major social media platforms have yet to adopt it. This fact raises a fair question: If the Trust button works, and we now understand how it influences sharing, why hasn't it been implemented?

We can only speculate on the reasons behind this reluctance. Until recently, political will has not been sufficient to create and enforce public policies for reducing misinformation on social media. In fact, politicians and the platforms themselves may actually benefit from social media structures that spread falsehoods and rile up their political base.[78] In addition, social media researchers have largely promoted individual-level solutions (again, see Table 1), which have demonstrated limited long-term success at combatting the current reward structures that promote misinformation on social media. As long as the major scientific outlets continue to favor individual-level approaches, social media companies will make little progress in controlling online misinformation.

As we noted above, the possibility that shifting buttons might decrease user engagement has not emerged in any of our experiments thus far. Granted, we have mainly tested the Trust button in controlled studies rather than large-scale platform rollouts. Additional testing in a variety of contexts is, of course, a crucial step in scaling any intervention.

In the meantime, Reddit offers a compelling example of a platform that has implemented novel rewards. Reddit's upvote/downvote system allows users to evaluate content based on whether it is helpful, informative, or meaningful—not just entertaining. Accumulated karma points (derived from upvotes and downvotes) shape users' reputations and access to community participation, a mechanism that could skew incentives toward contribution quality rather than potential virality.[79] While upvotes are not explicitly trust signals, they function as community-based indicators of value and relevance, closer to trust than to pure popularity.

This community-first architecture appears to be paying off. According to the Neely Social Media Index,[77] Reddit was the only major platform to show an increase in active users from 2023 to 2025; on an aggregated basis, Facebook, Instagram, TikTok, YouTube, and Snapchat all plateaued or declined. Over the same period, the percentage of users who reported having a negative experience on Reddit dropped by 39%, and concerns about misinformation on the site were significantly lower than for other platforms. Although Reddit still faces issues of polarization, misinformation appears to be less of a reason for negative experiences on the site.

The potential gains are monetary as well as experiential. In 2024, Reddit generated $1.3 billion in revenue, reached 91 million daily active users, and had a $10 billion pre-IPO valuation—its most financially successful year since 2012.[80] In sum, Reddit's numbers imply that reward systems centered on content quality rather than entertainment value alone can scale, support user well-being, and create tangible economic value. Its performance offers a powerful, real-world proof of concept for platforms and policymakers considering implementing structural interventions like Trust buttons.

Conclusion

Social media platforms reward engagement over accuracy, a reward structure that helps to explain why even well-intentioned users spread misinformation. Although partisan actors and authoritarian regimes deliberately spread falsehoods,[56,65,81] ordinary users with no intent to deceive propagate most misinformation.[12,82]

We propose a simple shift in reward structure that will curb the flow of user-generated misinformation and improve the quality of information on social media. Instead of the current rewards for interesting, emotional, and sensational information, we propose shifting rewards to recognizing accuracy by including a Trust button. As we showed in the present article, this easily implementable platform redesign allows users to meet their dual goals of sharing accurate information that is recognized for its worth while continuing to gain social support and connect with

others online. Successful social media platforms that incorporate trust incentives in the future will be giving users what they want: ready access to information they can trust.

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

Gizem Ceylan https://orcid.org/0000-0002-3876-9312
Wendy Wood https://orcid.org/0000-0002-6117-558X

References

1. World Economic Forum. (2024). *Global risks report 2024*. https://www.weforum.org/publications/global-risks-report-2024/
2. Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011
3. Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1–16. https://doi.org/10.1037/xap0000315
4. Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, *118*(37), Article e2104235118. https://doi.org/10.1073/pnas.2104235118
5. Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Barzilai, S., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fasio, 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
6. Ceylan, G., & Wood, W. (2026). *Social approval or truth: How social feedback trains users to share interesting misinformation*. Unpublished manuscript, Department of Marketing and Logistics, The Ohio State University.
7. 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
8. Bayer, J. B., & LaRose, R. (2018). Technology habits: Progress, problems, and prospects. In B. Verplanken (Ed.), *The psychology of habit: Theory, mechanisms, change, and contexts* (pp. 111–130). Springer International Publishing.
9. Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205. https://doi.org/10.1509/jmr.10.0353
10. Milkman, K. L., & Berger, J. (2014). The science of sharing and the sharing of science. *Proceedings of the National Academy of Sciences*, *111*(Suppl. 4), 13642–13649. https://doi.org/10.1073/pnas.1317511111
11. Rathje, S., & Van Bavel, J. J. (2025). The psychology of virality. *Trends in Cognitive Sciences*, *29*(1), 914–927. https://doi.org/10.1016/j.tics.2025.06.014
12. 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
13. Ceylan, G. (2026). *How learning about social norms shape misinformation sharing*. Unpublished manuscript, Department of Marketing and Logistics, The Ohio State University.
14. Aslett, K., Sanderson, Z., Godel, W., Persily, N., Nagler, J., & Tucker, J. A. (2024). Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, *625*(7995), 548–556. https://doi.org/10.1038/s41586-023-06883-y
15. McGrew, S. (2024). Teaching lateral reading: Interventions to help people read like fact checkers. *Current Opinion in Psychology*, *55*, Article 101737. https://doi.org/10.1016/j.copsyc.2023.101737
16. Wineburg, S., Breakstone, J., McGrew, S., Smith, M. D., & Ortega, T. (2022). Lateral reading on the open internet: A district-wide field study in high school government classes. *Journal of Educational Psychology*, *114*(5), 893–909. https://doi.org/10.1037/edu0000740
17. Maertens, R., Roozenbeek, J., Simons, J. S., Lewandowsky, S., Maturo, V., Goldberg, B., Xu, R., & van der Linden, S. (2025). Psychological booster shots targeting memory increase long-term resistance against misinformation. *Nature Communications*, *16*, Article 2062. https://doi.org/10.1038/s41467-025-57205-x
18. Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, *8*(34), Article eabo6254. https://doi.org/10.1126/sciadv.abo6254
19. Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117
20. Guess, A. M., McGregor, S., Pennycook, G., & Rand, D. (2024). *Unbundling digital media literacy tips: Results from two experiments*. OSF. https://doi.org/10.31234/osf.io/u34fp
21. Andı, S., & Akesson, J. (2021). Nudging away false news: Evidence from a social norms experiment. *Digital Journalism*, *9*(1), 106–125. https://doi.org/10.1080/21670811.2020.1847674
22. Butler, L. H., Prike, T., & Ecker, U. K. H. (2024). Nudge-based misinformation interventions are effective in information environments with low misinformation prevalence. *Scientific Reports*, *14*, Article 11495. https://doi.org/10.1038/s41598-024-62286-7
23. Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*, 13–29. https://doi.org/10.1038/s44159-021-00006-y
24. Chan, M.-p. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579
25. Fazio, L. (2020, February 10). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-009
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*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2
27. Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, *13*, Article 2333. https://doi.org/10.1038/s41467-022-30073-5
28. Aslett, K., Guess, A. M., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). News credibility labels have limited average effects on news

diet quality and fail to reduce misperceptions. *Science Advances*, *8*(18), Article eabl3844. https://doi.org/10.1126/sciadv.abl3844
29. Fendt, M., Holford, D. L., & Lewandowsky, S. (2024). *Friction against fiction: Adding 'grit' to boost psychological inoculation against misinformation*. OSF. https://doi.org/10.31234/osf.io/dzbm7
30. Jahn, L., Rendsvig, R. K., Flammini, A., Menczer, F., & Hendricks, V. F. (2023). *Friction interventions to curb the spread of misinformation on social media*. arXiv. https://doi.org/10.48550/arXiv.2307.11498
31. Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, *42*(4), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0
32. Drolsbach, C. P., Solovev, K., & Pröllochs, N. (2024). Community notes increase trust in fact-checking on social media. *PNAS Nexus*, *3*(7), Article pgae217. https://doi.org/10.1093/pnasnexus/pgae217
33. Bak-Coleman, J. B., Kennedy, I., Wack, M., Beers, A., Schafer, J. S., Spiro, E. S., Starbird, K., & West, J. D. (2022). Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, *6*(10), 1372–1380. https://doi.org/10.1038/s41562-022-01388-6
34. Chen, Y.-S., & Zaman, T. (2024). Shaping opinions in social networks with shadow banning. *PLoS ONE*, *19*(3), Article e0299977. https://doi.org/10.1371/journal.pone.0299977
35. Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33). https://doi.org/10.1126/sciadv.abe5641
36. Anderson, I. A., & Wood, W. (2021). Habits and the electronic herd: The psychology behind social media's successes and failures. *Consumer Psychology Review*, *4*(1), 83–99. https://doi.org/10.1002/arcp.1063
37. Meshi, D., Tamir, D. I., & Heekeren, H. R. (2015). The emerging neuroscience of social media. *Trends in Cognitive Sciences*, *19*(12), 771–782. https://doi.org/10.1016/j.tics.2015.09.004
38. Zell, A. L., & Moeller, L. (2018). Are you happy for me . . . on Facebook? The potential importance of "likes" and comments. *Computers in Human Behavior*, *78*, 26–33. https://doi.org/10.1016/j.chb.2017.08.050
39. Cohen, M. S., & Decety, J. (2026). Social feedback mechanisms & misinformation: A neuroscience-based argument for algorithm regulation. *Behavioral Science & Policy*. https://doi.org/10.1177/23794607251403323
40. 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
41. Contrera, J. (2016, May 25). This is what it's like to grow up in the age of likes, lols and longing. *The Washington Post*. https://www.washingtonpost.com/sf/style/2016/05/25/13-right-now-this-is-what-its-like-to-grow-up-in-the-age-of-likes-lols-and-longing/
42. Jargon, J. (2020, February 18). Teens are deleting Instagrams almost as fast as they post them. *The Wall Street Journal*. https://www.wsj.com/articles/teens-are-deleting-instagrams-almost-as-fast-as-they-post-them-11582021801?
43. 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
44. 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
45. 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
46. Frimer, J. A., Aujla, H., Feinberg, M., Skitka, L. J., Aquino, K., Eichstaedt, J. C., & Willer, R. (2023). Incivility is rising among American politicians on twitter. *Social Psychological and Personality Science*, *14*(2), 259–269. https://doi.org/10.1177/19485506221083811
47. Kantrowitz, A. (2021, November 4). *The case to reform the share button, according to Facebook's own research*. Big Technology. https://www.bigtechnology.com/p/the-case-to-reform-the-share-button
48. Skinner, B. F. (1991). *The behavior of organisms: An experimental analysis*. BF Skinner Foundation. https://www.bfskinner.org/wp-content/uploads/2016/02/BoO.pdf
49. 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.sa2
50. Perez, O. D., & Dickinson, A. (2020). A theory of actions and habits: The interaction of rate correlation and contiguity systems in free-operant behavior. *Psychological Review*, *127*(6), 945–971. https://doi.org/10.1037/rev0000201
51. Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*, 769–771. https://doi.org/10.1038/s41562-017-0213-3
52. Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bies, A. M., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, *7*(10), Article 201199. https://doi.org/10.1098/rsos.201199
53. Gardner, B., Abraham, C., Lally, P., & De Bruijn, G.-J. (2012). Towards parsimony in habit measurement: Testing the convergent and predictive validity of an automaticity subscale of the Self-Report Habit Index. *International Journal of Behavioral Nutrition and Physical Activity*, *9*, Article 102. https://doi.org/10.1186/1479-5868-9-102
54. Lins De Holanda Coelho, G., Hanel, P. H. P., & Wolf, L. J. (2020). The very efficient assessment of need for cognition: Developing a six-item version. *Assessment*, *27*(8), 1870–1885. https://doi.org/10.1177/1073191118793208
55. Guess, A. M., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), Article eaau4586. https://doi.org/10.1126/sciadv.aau4586
56. Roozenbeek, J., Remshard, M., & Kyrychenko, Y. (2024). Beyond the headlines: On the efficacy and effectiveness of misinformation interventions. *Advances in Psychology*, *2*, Article e24569. https://doi.org/10.56296/aip00019
57. Celadin, T., Panizza, F., & Capraro, V. (2024). Promoting civil discourse on social media using nudges: A tournament of seven interventions. *PNAS Nexus*, *3*(10), Article pgae380. https://doi.org/10.1093/pnasnexus/pgae380
58. Sasaki, S., Kurokawa, H., & Ohtake, F. (2021). Effective but fragile? Responses to repeated nudge-based messages for preventing the spread of COVID-19 infection. *The Japanese Economic Review*, *72*(3), 371–408. https://doi.org/10.1007/s42973-021-00076-w
59. Gardner, B., Rebar, A. L., de Wit, S., & Lally, P. (2024). What is habit and how can it be used to change real-world behaviour? Narrowing the theory-reality gap. *Social and Personality Psychology Compass*, *18*(6), Article e12975. https://doi.org/10.1111/spc3.12975
60. Wood, W. (2024). Habits, goals, and effective behavior change. *Current Directions in Psychological Science*, *33*(4), 226–232. https://doi.org/10.1177/09637214241246480
61. Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion; psychological studies of opinion change*. Yale University Press.
62. Cacioppo, J. T., Petty, R. E., Kao, C. F., & Rodriguez, R. (1986). Central and peripheral routes to persuasion: An individual difference perspective. *Journal of Personality and Social Psychology*, *51*(5), 1032–1043. https://doi.org/10.1037/0022-3514.51.5.1032
63. Schwarz, N., & Jalbert, M. (2020). When (fake) news feels true: Intuitions of truth and the acceptance and correction of misinformation. In R. Greifeneder, M. Jaffe, E. Newman, & N. Schwarz (Eds.), *The psychology of fake news: Accepting, sharing, and correcting misinformation* (pp. 73–89). Routledge. https://doi.org/10.4324/9780429295379-7
64. Fu, S., Li, H., Liu, Y., Pirkkalainen, H., & Salo, M. (2020). Social media overload, exhaustion, and use discontinuance: Examining the

effects of information overload, system feature overload, and social overload. *Information Processing & Management*, *57*(6), Article 102307. https://doi.org/10.1016/j.ipm.2020.102307
65. Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018
66. *How community notes work on Facebook*. (2025). Facebook. https://www.facebook.com/help/1416832942629495.
67. Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, *120*(7), Article e2210666120. https://doi.org/10.1073/pnas.2210666120
68. Martel, C., Allen, J., Pennycook, G., & Rand, D. G. (2024). Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*, *19*(2), 477–488. https://doi.org/10.1177/17456916231190388
69. Van Alstyne, M. W. (2020). Proposal: A market for truth to address false ads on social media. *Communications of the ACM*, *63*(7), 23–25. https://doi.org/10.1145/3401724
70. Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M. B. F., Coleman, K., & Baxter, J. (2022). *Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation*. arXiv. https://doi.org/10.48550/arXiv.2210.15723
71. 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
72. *The Economist*. (2024, May 1). Disinformation is on the rise. How does it work?
73. *The Economist*. (2024, May 1). Producing fake information is getting easier.
74. Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023). Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Article 436. https://doi.org/10.1145/3544548.3581318
75. European Commission. (2024). *The 2022 code of practice on disinformation*. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation
76. Wang, L., & Forman-Katz, N. (2024, February 7). *Many Americans find value in getting news on social media, but concerns about inaccuracy have risen*. Pew Research Center. https://www.pewresearch.org/short-reads/2024/02/07/many-americans-find-value-in-getting-news-on-social-media-but-concerns-about-inaccuracy-have-risen/
77. Xing, E., Liu, Y., Iyer, R., & Fast, N. (2025, July 17). Reddit in transition: A two-year examination of social media experiences and evolution. *Designing Tomorrow*. https://psychoftech.substack.com/p/reddit-in-transition-a-two-year-examination
78. Wynn-Williams, S. (2025). *Careless people: A cautionary tale of power, greed, and lost idealism*. Flatiron Books.
79. Reddit Recommended. (2023, June 30). *How to gain karma on Reddit*. LinkedIn. https://www.linkedin.com/pulse/how-gain-karma-reddit-reddit-recommended/
80. Curry, D. (2026, January 7). Reddit revenue and usage statistics (2026). Business of Apps. https://www.businessofapps.com/data/reddit-statistics/
81. Simonov, A., & Rao, J. (2022). Demand for online news under government control: Evidence from Russia. *Journal of Political Economy*, *130*(2), 259–309. https://doi.org/10.1086/717351
82. Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, *10*, Article 7. https://doi.org/10.1038/s41467-018-07761-2