World view

By Tali Sharot

To quell misinformation, use carrots — not just sticks

Social-media platforms should reward users for reliable, accurate and trustworthy posts.

In 2020 alone, social-media shares, likes and similar interactions with misleading online news doubled to 17% of all engagements. This staggering growth has consequences: polarization, violent extremism, racism and resistance to climate action and vaccines. Social-media companies have taken some steps to combat misinformation by using warnings and 'sticks', such as removing a few virulent spreaders of falsities and flagging misleading content. Facebook and Instagram users can report concerning posts, and Twitter prompts users to read articles before retweeting them.

How social-media companies should revamp their recommendation algorithms to quell misinformation is being discussed, but something is missing from the conversation: how to improve what users want to post and spread. Right now, users lack clear, quick incentives for reliability. Social-media platforms need to offer 'carrots' for truth.

As a neuroscientist who studies motivation and decision making, I have seen how even trivial rewards strongly influence behaviour. Most readers have felt an ego boost when their post received 'likes'. Such engagement also results in followers, which can help people secure lucrative deals.

Thus, if a certain type of content generates high engagement, people will post more content like it. Here is the conundrum: fake news generates more retweets and likes than do reliable posts, spreading 6–20 times faster. This is largely because such content captures attention and confirms existing beliefs. What's more, people share information even when they do not trust it. In one experiment (G. Pennycook *et al. Nature* https://doi.org/10.1038/s41586-021-03344-2; 2021), 40% of users who were shown fake news articles congruent with their political affiliation would consider sharing them, even though only 20% thought they were accurate.

At the moment, users are rewarded when their post appeals to the masses — even if it's of poor quality. What would happen if users were rewarded for reliability and accuracy? A system that explicitly provides visible rewards for reliability has never, to my knowledge, been introduced by any major social-media platform. Such a system would work with the natural human tendency to select actions that lead to the greatest reward. It could thus both reinforce user behaviour that generates trustworthy material and signal to others that the post is dependable.

Reward systems have been successfully implemented before. In Sweden, drivers were offered prizes for obeying the speed limit, and average speed was reduced by 22%. In South Africa, a health-insurance company offered clients points whenever they purchased fruits and vegetables in the supermarket, visited the gym or attended a medical screening. Points could be exchanged for items, and were made visible to participants and their social circles. Participants' behaviour changed enough to reduce hospital visits.

> **"What would happen if users were rewarded for reliability and accuracy?"**

A challenge to implementing such a system on social media is how to assess the reliability of posts. One option would be to add a 'trust' button and display the number of 'trust' clicks a post receives. Although care would need to be taken to prevent gaming, such a feature would provide an engagement route for users, and would thus be compatible with the business model of social-media companies. And it would prime users to consider reliability. The study mentioned above found that prompting users to consider the veracity of a single statement made them less likely to consider sharing fake headlines.

There are positive examples of user-based assessments. In Amazon's reviewer ranking, reviewers rated as helpful are invited by Amazon to join the Amazon Vine programme, which offers perks for reviews. However, a social-media user's assessment will be driven partially by confirmation bias. Encouragingly, a recent paper reports that the average reliability assessment from a large group of people is highly correlated with that of professional fact checkers, albeit only if the group is politically balanced (J. Allen *et al.* Preprint at https://doi.org/fz4j; 2020). The quality of Wikipedia demonstrates that it is possible to create an environment in which reliability assessments are effectively outsourced to users, as long as there is oversight by trusted users or fact-checkers. Social-media companies already employ fact-checkers, who could be given the ability to disable the trust button on misleading posts and provide a 'gold star' for reliable ones. An automated evaluation that assesses content by, for instance, cross-referencing it with substantiated web sources could be used for this purpose as well.

However reliability is assessed, those who consistently rank high could be awarded with a 'reliable user' badge. (The current blue checkmark on Twitter simply verifies that an account holder of an 'account of public interest' is authentic.)

Some might argue that the power of carrots to promote accurate information will be inadequate against engagement-optimizing algorithms and human tendencies that promote misinformation. But it is an approach that should be tried. Right now, sticks are used very narrowly. In the 2 weeks before the US election, only 0.2% of all election-related tweets were labelled as misleading. By contrast, carrots — trust ratings, gold stars and 'reliable user' badges — can apply to most users.

The exact features of a carrot system need to be carefully developed and thoroughly tested. We need the combined expertise of network scientists, computer scientists, psychologist and economists, among others. To create a healthy information ecosystem, we need to add carrots.

Tali Sharot is a professor of cognitive neuroscience at University College London, where she directs the Affective Brain Lab.
e-mail: t.sharot@ucl.ac.uk

MICHAEL LIONSTAR