



All CCS and Cross-sectoral | Media & Press | Video Games & Multi-media | All Europe Countries | International | Featured Stories

The dynamics that polarise us on social media are about to get worse

Meta's shift to community-driven fact-checking threatens to amplify social media polarization by exploiting psychological biases that favor extreme views, potentially transforming online platforms into breeding grounds for misinformation and group conflict.

[Colin M. Fisher, UCL](#)

Meta founder and CEO Mark Zuckerberg has announced big changes in how the company addresses misinformation across Facebook, Instagram and Threads. Instead of relying on independent third-party factcheckers, Meta will now emulate Elon Musk's X (formerly Twitter) in using "community notes". These crowdsourced contributions allow users to flag content they believe is questionable.

Zuckerberg claimed these changes promote "free expression". But [some experts](#) worry he's bowing to right-wing political pressure, and will effectively allow a deluge of hate speech and lies to spread on Meta platforms.

Research on the group dynamics of social media suggests those experts have a point.

At first glance, community notes might seem democratic, reflecting values of free speech and collective decisions. Crowdsourced systems such as Wikipedia, Metaculus and PredictIt, though imperfect, often succeed at harnessing the [wisdom of crowds](#) — where the collective judgement of many can sometimes outperform even experts.

[Research shows](#) that diverse groups that pool independent

judgements and estimates can be surprisingly effective at discerning the truth. However, wise crowds seldom have to contend with social media algorithms.

Many people rely on platforms such as Facebook for their news, risking exposure to misinformation and biased sources. Relying on social media users to police information accuracy could further polarise platforms and amplify extreme voices.

Two group-based tendencies — our psychological need to sort ourselves and others into groups — are of particular concern: in-group/out-group bias and acrophily (love of extremes).

Ingroup/outgroup bias

Humans are biased in how they evaluate information. People are more likely to trust and remember information from their in-group — those who share their identities — while distrusting information from perceived out-groups. This bias leads to echo chambers, where like-minded people reinforce shared beliefs, regardless of accuracy.

It may feel rational to trust family, friends or colleagues over strangers. But in-group sources often hold [similar perspectives and experiences](#), offering little new information. Out-group members, on the other hand, [are more likely to provide diverse viewpoints](#). This diversity is [critical](#) to the wisdom of crowds.

But too much disagreement between groups can prevent community fact-checking from even occurring. Many community notes on X (formerly Twitter), such as those related to COVID vaccines, were likely [never shown publicly](#) because users disagreed with one another. The benefit of third-party factchecking was to provide an objective outside source, rather than needing widespread agreement from users across a network.

Worse, such systems are vulnerable to manipulation by well organised groups with political agendas. For instance, Chinese nationalists [reportedly](#) mounted a campaign to edit Wikipedia entries related to China-Taiwan relations to be more favourable to China.

Political polarisation and acrophily

Indeed, politics intensifies these dynamics. In the US, [political identity increasingly dominates](#) how people define their social groups.

Political groups are motivated to define “the truth” in ways that advantage them and disadvantage their political opponents. It’s easy to see how organised efforts to spread politically motivated lies and discredit inconvenient truths could corrupt the wisdom of crowds in Meta’s community notes.

Social media accelerates this problem through a phenomenon called acrophily, or a preference for the extreme. [Research shows](#) that people tend to engage with posts slightly more extreme than their own views.

Extreme and negative views get more attention online, driving social media communities apart. [evan_huang/Shutterstock](#)

These increasingly extreme posts are more likely to be negative than positive. Psychologists have known for decades that [bad is more engaging than good](#). We are hardwired to pay more attention to negative experiences and information than positive ones.

On social media, this means negative posts – about violence, disasters and crises – get more attention, often at the expense of more neutral or positive content.

Those who express these extreme, negative views gain status within their groups, attracting more followers and amplifying their influence. Over time, people come to think of these slightly more extreme negative views as normal, slowly moving their own views toward the poles.

[A recent study](#) of 2.7 million posts on Facebook and Twitter found that messages containing words such as “hate”, “attack” and “destroy” were shared and liked at higher rates than almost any other content. This suggests that social media isn’t just amplifying extreme views — it’s [fostering a culture of out-group hate](#) that undermines the collaboration and trust needed for a system like community notes to work.

The path forward

The combination of negativity bias, in-group/out-group bias and acrophily supercharges one of the greatest challenges of our time: polarisation. Through polarisation, extreme views become normalised, eroding the potential for shared understanding across group divides.

The best solutions, which I examine in [my forthcoming book, The Collective Edge](#), start with diversifying our information sources. First, people need to engage with — and collaborate across — different groups to break down barriers of mistrust.

Second, they must seek information from multiple, reliable news and information outlets, not just social media.

However, social media algorithms often work against these solutions, creating echo chambers and trapping people's attention. For community notes to work, these algorithms would need to prioritise diverse, reliable sources of information.

While community notes could theoretically harness the wisdom of crowds, their success depends on overcoming these psychological vulnerabilities. Perhaps increased awareness of these biases can help us design better systems — or empower users to use community notes to promote dialogue across divides. Only then can platforms move closer to solving the misinformation problem.

[Colin M. Fisher](#), Associate Professor of Organisations and Innovation and Author of "The Collective Edge: Unlocking the Secret Power of Groups", [UCL](#)

Image Creator: Richard Drew , Credit: AP, CC BY 4.0

This article is republished from [The Conversation](#) under a Creative Commons licence. Read the [original article](#).