Platform-level interventions to reduce the spread of misinformation & hateful content online

THE BEHAVIOURAL INSIGHTS TEAM

The UK's recent riots provide a window for improving online safety

In the wake of the UK's recent riots, social media platforms are under fire for perpetuating misinformation, disinformation, and hateful content. This note outlines behavioural factors that point to the need for platform and system-wide interventions to combat misinformation.

Underpinning this note is analysis from a survey of 2,000+ UK adults, conducted between 16 August 2024 - 19 August 2024.

Social media platforms are hotbeds for false information. 74% of social media users in our survey reported seeing some false information in the week prior on at least one of the social media platforms they use.

The way platforms are designed can not only facilitate but amplify the flow of false information and hateful content. Algorithms designed to maximise user engagement show users content that aligns with existing beliefs. This is often sensational and emotionally charged, creating echo chambers in which false information thrives.

The 'paradox of passivity' means passive consumption of content could be more harmful than people think. Many social media users consume media passively, scrolling and absorbing content without critical consciousness, active engagement, or fact-checking. Over time, passively-absorbed content accumulates and - when combined with *confirmation bias, availability bias*, and our tendency to believe something is true the more it is repeated (*illusory truth effect*) - the risk that users begin to believe false information increases.

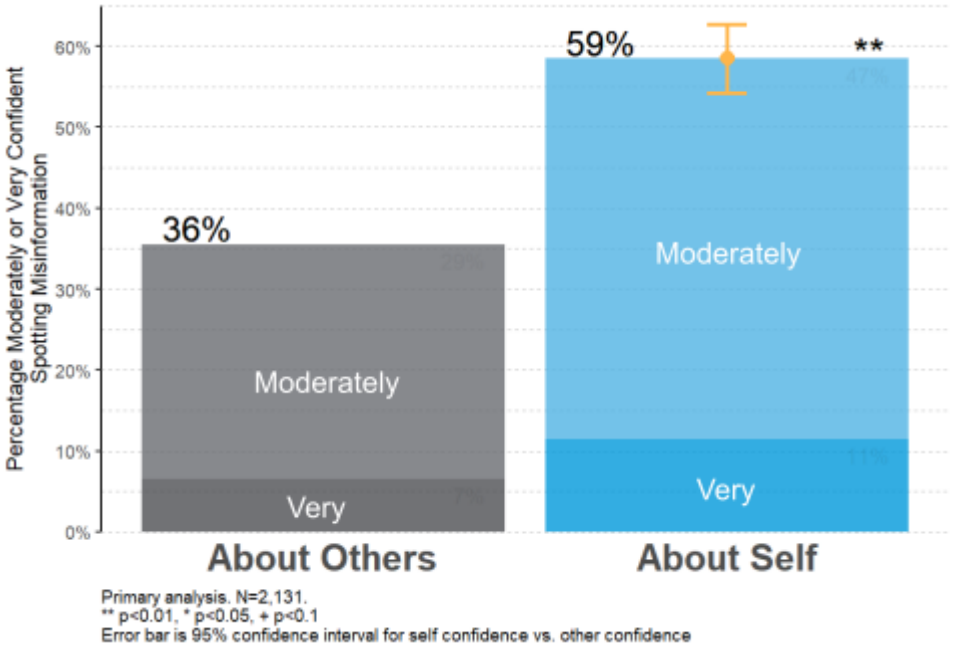
Recommendation 1: Platforms should use evidence-based prompts to reduce the spread of false information

Many social media platforms add prompts to flag misleading posts or to convey additional information or context, such as whether a post has been fact-checked. However, these types of 'debunking' interventions have not been consistently found to be effective at changing beliefs or reducing the spread of misinformation. Platforms should use evidence-based interventions in prompts and flags on posts, including using a prompt as a 'timely moment' to signpost to evidence-based inoculation interventions (where users are trained in the techniques used to convey false information, such as the Bad News Game), or using accuracy prompts (which encourage users to reflect before sharing a post). Platforms should be encouraged to test the effectiveness of the specific interventions their platform is using to combat the spread of false information

and other sensitive content, and to publish the results of these tests publicly. This would not only help to hold platforms to account, but also contribute to the evidence base on what works in this space.

Users are overconfident about their own abilities. Driving is the oft-given example of *illusory superiority* (also known as the better-than-average effect), with a 1980s study finding that 93% of American drivers rated themselves as being safer than average- a statistical impossibility.

Illusory superiority extends to misinformation online. In our recent survey of 2000+ UK adults, half our sample were randomised to answer how well they thought *they* could spot false information on social media platforms. The other half of the sample was asked how well they thought *others* could spot false information. We found a statistically significant difference between perceptions of self and perceptions of others. 59% of people reported that they could detect false information either moderately well or very well. However, only 36% of people thought that others would be able to do the same. This is highly suggestive of overconfidence, and is consistent with findings from Ofcom, where 69% of respondents felt confident in identifying misinformation, but only 22% were able to correctly identify a genuine post. Overconfidence may be a crucial factor for explaining how false and low-quality information spreads via social media, with the overconfident most susceptible to believing false information and spreading it further.



Overconfidence likely limits the effectiveness of prompts and other media literacy interventions like inoculation. While interventions such as inoculation have been found to be effective, they rely on people to voluntarily take them. This means they will suffer from selection bias, wherein the overconfident are least likely to take up the

intervention. In a recent Turing study, despite over 90% of social media users reporting encountering misinformation online, just 7% of people report using self-help resources to counteract misinformation. Users also generally believe they should not be held responsible for misinformation, with the majority of our respondents (59%) believing the onus should rest on the platforms themselves.

Recommendation 2: Reduce the spread of false information by setting content controls to show reduced harmful content by default

Reducing the overall amount of legal but harmful content (sometimes referred to as sensitive content) that users see can help to account for overconfidence and the paradox of passivity. Evidence shows that initial content control choices are sticky, meaning that users are less likely to deviate from initial choices in terms of the content that is viewed.

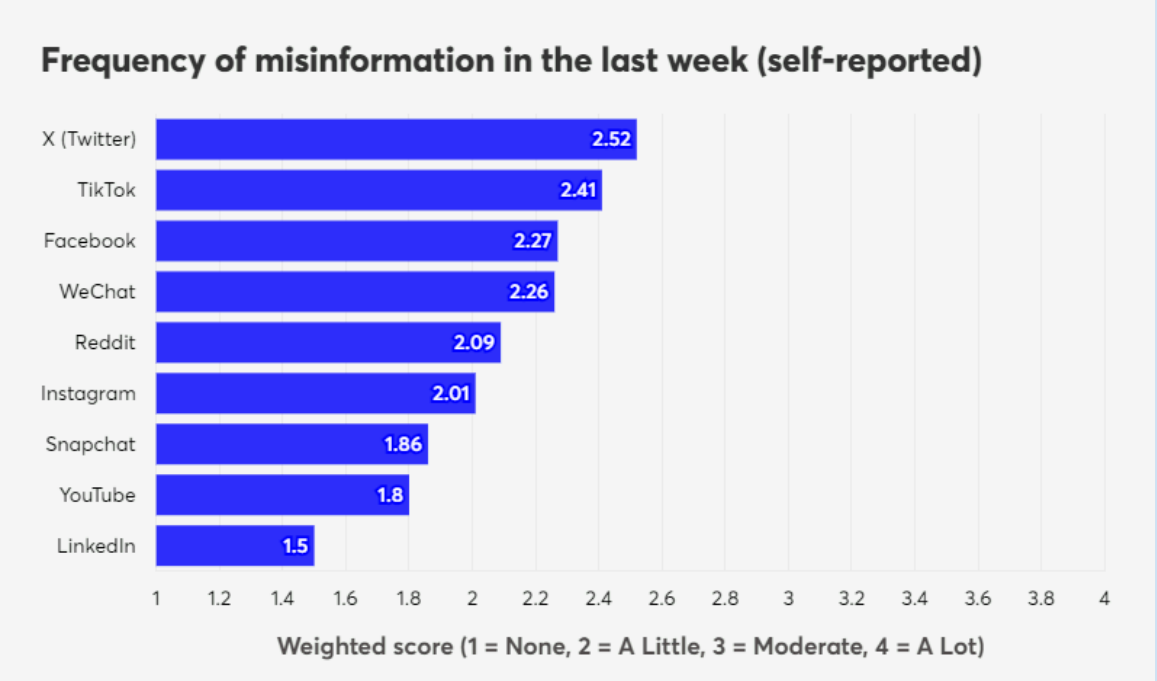
Presenting content controls with the option to show reduced sensitive content as the default is a low-cost way to reduce the circulation of legal but harmful content (which includes false information). Our recent trial with Ofcom demonstrated the impact of pre-selection: there was an 18pp difference between those who selected "Reduced sensitive content" in the forced choice when no option was pre-selected (24%), compared to when "Reduced sensitive content" was pre-selected (42%). This intervention would give users easy access to content controls, while also reducing overall societal exposure to legal but harmful content.

There are misaligned incentives between social media platforms and users. Social media algorithms prioritise engagement, often promoting controversial content. While users say they dislike misinformation, they are reluctant to pay for its removal. Our survey found that 82.5% of respondents would not pay for premium, misinformation-free versions of the platforms they currently use. Users can (and do) switch platforms, as was demonstrated by some high profile users on X following misinformation about riots. However, these are usually one-off events.

Recommendation 3: Publish platform rankings to shift user choice and drive continuous improvements to platforms

Increasing transparency may be one way of addressing such market failures. Ranking platforms based on the prevalence of false or hateful content could mobilise consumers to switch platforms, effectively "deshrouding" the issue of misinformation. In our survey, 52% of respondents would be moderately or very likely to switch to a higher-ranked platform with less false information.

In our survey, we asked participants how frequently they saw misinformation in the last week. We used these results to create a ranking of social media platforms [1]:



Such a ranking could be combined with other objective measures of the prevalence of false information and presented to social media users and tested experimentally to see if it would encourage a change in the platforms individuals use.

[1] Note, this ranking was not significance tested and excludes those that were unsure about seeing false information. There are significantly fewer users of WeChat (73) than other platforms, therefore caution should be taken when interpreting results.