Fanning the flames: When attempts to call out misinformation backfire

Adrian Kwek

For The Straits Times

As news of the Notre-Dame fire broke, news agencies began to stream live coverage of the burning of the iconic Parisian cathedral on YouTube.

Widely disseminated screenshots showed news agency France 24's live streaming on YouTube – accompanied by a text box containing information on the Sept 11 attacks from the Encyclopaedia Britannica.

The live streams of CBS News and NBC News on YouTube were also said to be accompanied by similar text boxes. Given that the cause of the Notre-Dame fire had not yet been established, netizens responded to the association with 9/11 with derision and disbelief.

The Straits Times' website carried a Bloomberg report on the incident on Tuesday headlined "YouTube flags Notre Dame fire as 9/11 conspiracy, says system made 'wrong call'".

It turns out that as a measure against the spread of conspiratorial misinformation, YouTube had instituted a system where certain uploaded videos are labelled by algorithms as presenting conspiracy theories. This would in turn trigger a text box of potentially debunking factual information to be displayed beneath the video. Here, it appears that live streams of the Notre-Dame fire were mistakenly identified as presenting 9/11 conspiracy theories.

In this case, well-intentioned attempts to deal with perceived misinformation may have inadvertently created its own fake news event.

It is not only YouTube's algorithms that are at play here. People who read the text box information may have believed that indeed the fire was the result of a terrorist attack. Psychological factors at play influence the way viewers and readers perceive information and corrections. This has relevance for Singapore, especially given the move to introduce the Protection from Online Falsehoods and Manipulation Bill to deal with online falsehoods.

What are the psychological factors at play? Suppose you read a piece of news about a warehouse that caught fire. The news says that the fire began in a storeroom with "cans of oil paint and gas cylinders". As the fire burned, there were "oily smoke and sheets of flame", "explosions" and "toxic fumes".

As you read on, you come across a correction: "There was no paint or gas cylinders in the storage room."

Later on, when asked what the possible cause of the toxic fumes was, you reply: "The paint and the gas cylinders." Yet, when you are asked several questions later what the point of the correction message was, you reply: "That the storage room didn't have any gas or paint, but was empty."

This was an inconsistent yet common response of people who took part in American psychologist Colleen Seifert's 2002 experiment on the Continuing Influence Effect, which refers to the propensity of our minds to revert to misinformation in reasoning, even though that misinformation was supposed to have been removed by a correction.

The fact that participants were able to affirm the point of the correction message shows that they did read and understand it. Amazingly, in spite of this, they reverted to using the discredited information in subsequent reasoning about the cause of the fire.

Experts are not in agreement about what causes the continued influence of discredited information on reasoning. The causes may be complex, but research on confirmation and disconfirmation biases and backfire effects suggests that these can contribute to the Continuing Influence Effect.

Australian cognitive scientists John Cook and Stephan Lewandowsky, in their highly accessible The Debunking Handbook, list three types of backfire effects.

The Familiarity Backfire Effect refers to the entrenchment of a piece of misinformation in one's mind just from being exposed to repetitions of the misinformation. A correction backfires because it repeats the misinformation in order to discredit it.

The Overkill Backfire Effect refers to our mental preference for short and simple information as opposed to voluminous or complex information. While we ordinarily think that the more arguments we have to refute a piece of information, the less likely people will continue to use the refuted piece of information, it turns out that people may continue to use that information because it is simpler and cognitively easier to access.

How debunking agencies respond to misinformation can determine whether a piece of misinformation fades from view or gets amplified. What the continuing influence and backfire effects in turn suggest is that misinformation is retained when it is repeated, simple, emotionally significant and easily used in reasoning.

The Worldview Backfire Effect refers to our tendency to actively look for counter-arguments to corrections that are not consistent with our cherished ideologies or value systems.

How debunking agencies respond to misinformation can determine whether a piece of misinformation fades from view or gets amplified. What the continuing influence and backfire effects in turn suggest is that misinformation is retained when it is repeated, simple, emotionally significant and easily used in reasoning.

YouTube's well-intentioned attempt to counter the spread of conspiracy theories exploits the use of a viewer's own reasoning process. As reported on Wednesday in the Bloomberg Opinion article, "The terrible timing of YouTube's Notre-Dame snafu", debunking information is displayed beneath certain videos, leaving the viewer to draw his own conclusion about the veracity of the video. This measure makes an alternative debunking explanation immediately available, which competes with the misinformation in the viewer's reasoning process.

When genuine news is mislabelled as dubious content that warrants the accompaniment of debunking information, the effect can be pernicious.

An undiscerning consumer of the France 24 live stream on YouTube could experience the following: The mind looks for a cause for the Notre-Dame fire, finds none in the news, but is triggered by the text box to associate the fire with the shocking terrorist attacks of 9/11.

The path of least mental processing resistance and most emotional significance leads to the belief that terrorism is the cause of the Notre-Dame fire. What YouTube's "wrong call" has done is to generate "fake news" in the mind of this news consumer.

In Singapore, the Protection from Online Falsehoods and Manipulation Bill was tabled in Parliament on April 1. It is a prophylactic measure against the threat of online misinformation to Singapore society. Part 3, Point 11 of the Bill states that one response the Government can take against those who post online falsehoods is to issue a "Correction Direction", an order to publicise a correction notice.

Studies on the Continuing Influence Effect expose the limitation of such a response, but also suggest ways in which corrections can be run effectively. For example, using clear information and infographics is helpful. However, it remains an uphill challenge to correct the biases that led falsehoods to thrive online in the first place.

The YouTube blunder shows another way that response to misinformation can backfire – remedial responses to mislabelled misinformation can result in the accidental creation of fake news.

stopinion@sph.com.sg

• Adrian Kwek is a senior lecturer at the Singapore University of Social Sciences, and he teaches a university core course on fake news and critical thinking.