How behavioral science reduced the spread of misinformation on TikTok

By: Evelyn Gosnell, Kristen Berman, Lindsay Juarez, Richard Mathera

> We're the behavioral scientists who designed the intervention together with TikTok that reduced shares of flagged content by 24%. Here's how we did it.

Our company, Irrational Labs, is a team of behavioral scientists deeply passionate about helping companies use behavioral insights for good. So when TikTok came to us seeking expert guidance on an intervention to reduce the spread of misinformation, we immediately said yes.

In collaboration with TikTok, we designed a pair of prompts that were put on videos with "unsubstantiated content." This is a term for potentially-misleading information that fact-checkers are not able to verify. Our prompts successfully reduced shares by 24% when compared to a control group. Based on these results, this intervention is being rolled out globally across the TikTok platform over the coming weeks, starting with the US and Canada.

What secret sauce did we use?

We used a process called **Behavioral Design**. This involves generating a step-by-step map of the user experience, reviewing the relevant research, and marrying the two: You overlay insights onto the map to come up with a hypothesis-driven intervention, then test whether it works.

The science behind the intervention

After pre-testing different interventions, our team, in partnership with TikTok's product team, landed on a simple one that, in part, originated from work by leading MIT researchers on misinformation, David Rand and Gordon Pennycook.

We put a short prompt on videos with unsubstantiated content — a prompt that reminded people to think about the accuracy of the content. It looked like this:

Why would this work? Research from Rand et al shows that **people do value truth**. An accuracy prompt has been shown to work because it reminds people about their own personal value of truth — at the critical moment when they're about to share something on social media.
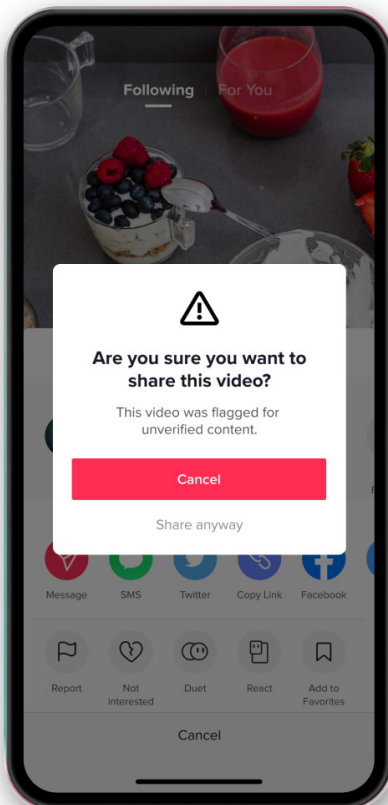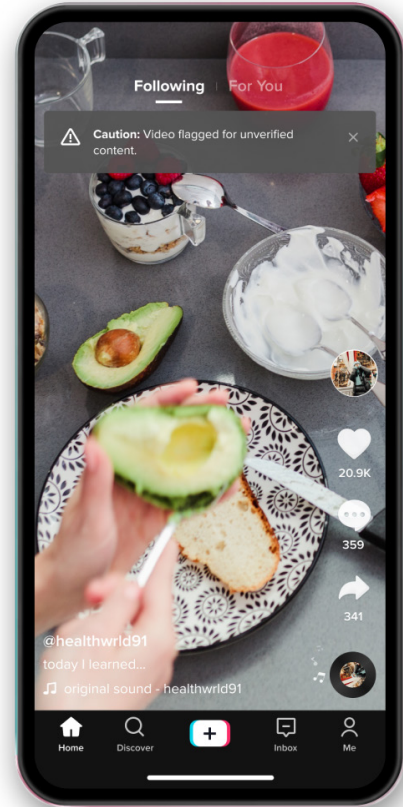
However, we still worried that this prompt would boost curiosity and thus, views. (If you see a video that says "don't watch this," you may want to watch it more.) To mitigate this, for some people we delayed the message. If after 3 seconds people were still watching and hadn't flipped to the next video, the prompt appeared.

In addition, when people went to actually share the video, we reminded them again that it was flagged as inaccurate, and asked whether they were sure they wanted to share:

Why would this work? Because humans have what are called "hot states" and "cold states." In hot states, we act on emotions. In cold states, we're more logical and deliberate. TikTok is a fast platform where users often act in hot states. By slowing people down, we hypothesized that we could decrease the sometimes overwhelming power of emotion.

A deeper dive on the results, and surprises

In addition to successfully reducing shares by 24%, the intervention also reduced likes by 7%, and views by 5%. While reducing shares is a direct metric of reducing the spread of misinformation, both likes and views are important parts of the wider context. We were heartened to see that the intervention moved the needle on them as well.

What surprised us? First, people who got the delayed prompt didn't view or share the video any less than those who saw it at the outset. Seeing the prompt right away marginally decreased the number of likes, but overall **curiosity didn't kill the cat**.

We were also intrigued to find differences in effectiveness of the intervention based on users' age. **The intervention was more likely to work on older users**, decreasing shares by 34% for users 35+, compared to a 21% reduction for users 18-24 and 25% reduction for users 25-34. In fact, when we looked at the data by looking at ages 18+, the intervention cut in half the number of shares relative to the control, with the video prompt being the main driver of this additional effect.

Where do we go from here?

We will continue to work with TikTok on new interventions. The video itself is not the only intervention point, and video prompts are only one intervention type. TikTok is deeply committed to exploring what works to reduce the spread of misinformation, and we are deeply committed to supporting them.

More broadly, we're heartened to see other platforms (like Twitter) apply research-driven insights to generate interesting interventions, and share results publicly. This level of creative thinking, willingness to try new things, and transparency all bode well for the future as we tackle this challenging problem.