THE NATIONAL INTEREST

Trial and Error May be the Secret to Social Media Content Moderation

Until the general population develops stronger antibodies against ideological pathogens, the best course may be to focus on light-touch user interface interventions that hamper the spread of disinformation without removing the content entirely.

October 20, 2020

James Pethokoukis

Content moderation is hard. We saw another example of that this week with the donnybrook over how Facebook and Twitter handled that New York Post story about Hunter Biden. Of course, no moderation would leave us with an internet that was a sewer, especially social media.

But maybe that is a good analogy. The development and expansion of cities, hubs of commerce and culture, was tremendously important for the advance of human civilization. But they were also diseased death traps until public sanitation. In the age of the internet, another innovation that fosters human connection, the dangerous pathogens are violent political movements or crazy conspiracy theories. As Cato Institute analyst Julian Sanchez told a congressional hearing in September, "The social media platforms on which these pathogens spread find themselves in the unenviable position of attempting, by trial and error, to discover how one builds a functional sewer system."

And that is exactly what we saw happen with Facebook and Twitter last week. Trial and error. Twitter CEO Jack Dorsey has already admitted the company made a mistake by blocking the spread of the Hunter Biden story without explanatory context. Certainly more errors will be made. Indeed, conservatives used to believe that trial and error, especially through the market process, would usually yield far better results than top-down instruction or rule-making by government central planners. Thus their caution and humility about regulation, particularly in emerging industries. (Friedrich Hayek wrote a bit about trial and error, about the evolutionary experiment that brought us the institutions of modern civilization.) More from Sanchez:

Until the general population develops stronger antibodies against ideological pathogens, the best course may be to focus on light-touch user interface interventions that hamper the spread of disinformation without removing the content entirely. Fact checks are one strategy already implemented by many platforms. Another is the disabling of single-click sharing of content flagged as false or misleading, and the adjustment of content-recommending algorithms to ensure that such content is not foisted upon users who do not willingly seek it out. The memetic

equivalent of a face mask, this does not prevent transmission by a determined user, but it does at least reduce the rate of casual or unthinking transmission. Another possibility is to visually distinguish content articles from reputable news outlets. In the pre-Internet era, the difference between a New York Times or Wall Street Journal report and a mimeographed screed could be seen at a glance, independently of the differences in style and content. On a Facebook feed, everything looks more or less identical. This strategy is only viable, of course, to the extent platforms can resist both political and user pressure to give their imprimatur to unreliable information sources with large constituencies. … If we expect inherently risk-averse businesses to be proactive about curating content to stanch the spread of extremist rhetoric and disinformation, they must be confident they are free to muddle through the process of developing adaptive responses—and, inevitably, to make many mistakes along the way—without incurring ruinous legal consequences as a result.