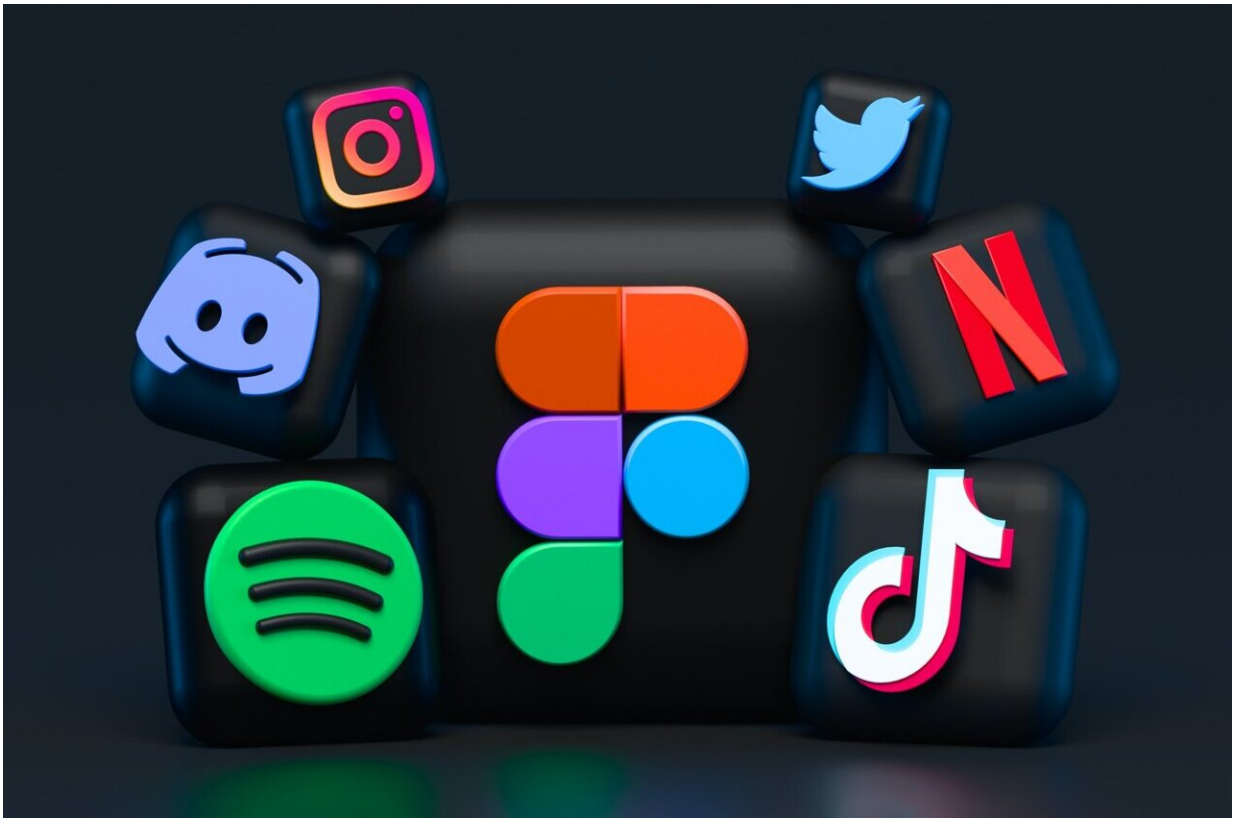
Why social media firms will struggle to follow new EU rules on illegal content

May 10 2022, by Greig Paul



Credit: Unsplash/CC0 Public Domain

Social media allowed us to connect with one another like never before. But it came with a price—it handed a megaphone to everyone, including terrorists, child abusers and hate groups. EU institutions recently reached

agreement on the Digital Services Act (DSA), which aims to "make sure that what is illegal offline is dealt with as illegal online."

The U.K. government also has an online safety bill in the works, to step up requirements for digital platforms to take down illegal material.

The scale at which large social media platforms operate—they can have billions of users from across the world—presents a major challenge in policing illegal content. What is illegal in one country might be legal and protected expression in another. For example, rules around criticizing government or members of a royal family.

This gets complicated when a user posts from one country, and the post is shared and viewed in other countries. Within the U.K., there have even been situations where it was legal to print something on the front page of a newspaper in Scotland, but not England.

The DSA leaves it to EU member states to define illegal content in their own laws.

The database approach

Even where the law is clear-cut, for example someone posting controlled drugs for sale or recruiting for banned terror groups, content moderation on social media platforms faces challenges of scale.

Users make hundreds of millions of posts per day. Automation can detect known illegal content based on a fuzzy fingerprint of the file's content. But this doesn't work without a database and content must be reviewed before it's added.

In 2021, the Internet Watch Foundation investigated more reports than in their first 15 years of existence, including 252,000 that contained

child abuse: a rise of 64% year-on-year compared to 2020.

New videos and images will not be caught by a database though. While artificial intelligence can try to look for new content, it will not always get things right.

How do the social platforms compare?

In early 2020, Facebook was reported to have around 15,000 content moderators in the U.S., compared to 4,500 in 2017. TikTok claimed to have 10,000 people working on "trust and safety" (which is a bit wider than content moderation), as of late 2020. An NYU Stern School of Business report from 2020 suggested Twitter had around 1,500 moderators.

Facebook claims that in 2021, 97% of the content they flagged as hate speech was removed by AI, but we don't know what was missed, not reported, or not removed.

The DSA will make the largest social networks open up their data and information to independent researchers, which should increase transparency.

Human moderators vs tech

Reviewing violent, disturbing, racist and hateful content can be traumatic for moderators, and led to a US$52 million (£42 million) court settlement. Some social media moderators report having to review as many as 8,000 pieces of flagged content per day.

While there are emerging AI-based techniques which attempt to detect specific kinds of content, AI-based tools struggle to distinguish between

illegal and distasteful or potentially harmful (but otherwise legal) content. AI may incorrectly flag harmless content, miss harmful content, and will increase the need for human review.

Facebook's own internal studies reportedly found cases where the wrong action was taken against posts as much as "90% of the time." Users expect consistency but this is hard to deliver at scale, and moderators' decisions are subjective. Gray area cases will frustrate even the most specific and prescriptive guidelines.

Balancing act

The challenge also extends to misinformation. There is a fine line between protecting free speech and freedom of the press, and preventing deliberate dissemination of false content. The same facts can often be framed differently, something well known to anyone familiar with the long history of "spin" in politics.

Social networks often rely on users reporting harmful or illegal content, and the DSA seeks to bolster this. But an overly-automated approach to moderation might flag or even hide content that reaches a set number of reports. This means that groups of users that want to suppress content or viewpoints can weaponize mass-reporting of content.

Social media companies focus on user growth and time spent on the platform. As long as abuse isn't holding back either of these, they will likely make more money. This is why it's significant when platforms take strategic (but potentially polarizing) moves—such as removing former U.S. president Donald Trump from Twitter.

Most of the requests made by the DSA are reasonable in themselves, but will be difficult to implement at scale. Increased policing of content will lead to increased use of automation, which can't make subjective

evaluations of context. Appeals may be too slow to offer meaningful recourse if a user is wrongly given an automated ban.

If the legal penalties for getting content moderation wrong are high enough for social networks, they may be faced with little option in the short term other than to more carefully limit what users get shown. TikTok's approach to hand-picked content was widely criticized. Platform biases and "filter bubbles" are a real concern. Filter bubbles are created where content shown to you is automatically selected by an algorithm, which attempts to guess what you want to see next, based on data like what you have previously looked at. Users sometimes accuse social media companies of platform bias, or unfair moderation.

Is there a way to moderate a global megaphone? I would say the evidence points to no, at least not at scale. We will likely see the answer play out through enforcement of the DSA in court.

Provided by The Conversation