How To Prep For Calif. Social Media Content Moderation Law

By **Marc Mayer and Stacey Chuvaieva** (November 21, 2022)

Marc Mayer

Stacey Chuvaieva

On Sept. 14, California Gov. Gavin Newsom signed a new content moderation law targeting social media companies, A.B. 587.

This new law is just one among a patchwork of newly enacted social media content moderation laws reflecting an overarching desire to reduce or address the spread of harmful information in social media.

The proliferation of such laws is not surprising. Discussions about the role of social media in American democracy have been at the forefront of policy debates for several years, especially after the Capitol riots on Jan. 6, 2021.

Those discussions have triggered an increasing number of legislative initiatives — ranging from proposed changes to Section 230 of the Communications Decency Act to passage of local laws addressing content moderation.

Prior to A.B. 587, Texas, Florida and New York introduced similar laws addressing content moderation practices. Florida's S.B. 7072, signed into law May 24, 2021, restricted the ability to moderate content — i.e., deleting or banning content made available by political candidates — and imposed disclosure obligations on social media companies.

The Texas law gave state residents and the Texas attorney general the ability to sue large social media companies for unfairly banned or censored content. The Texas law also required social media companies to implement a system to let users challenge any decision taken by the company with regard to flagged or removed content.

Both laws underwent free speech challenges, with inconsistent results. Earlier this year, the U.S. Court of Appeals for the Eleventh Circuit concluded that the Florida law violated constitutional protections in NetChoice LLC v. Attorney General of Florida.

The court held that "it is substantially likely that social-media companies — even the biggest ones — are 'private actors' whose rights the First Amendment protects … [and] that their so-called 'content-moderation' decisions constitute protected exercises of editorial judgment."[1]

By contrast, in NetChoice v. Ken Paxton, the U.S. Court of Appeals for the Fifth Circuit upheld the Texas social media law in September, disagreeing with the Eleventh Circuit and holding that the law did not violate the First Amendment rights of the platforms.

New York also recently signed a law that requires social media companies to develop hateful conduct policies and reporting mechanisms for such conduct.

It is entirely possible that other states also will follow the lead, and thus it may be in the interest of all social media platforms to ensure that they have developed content moderation policies that can withstand public scrutiny.

Targeted Companies

California A.B. 587 applies to companies operating social media platforms that generate more than $100 million in gross revenue during the preceding calendar year.

Scope

The law does not provide direct instructions as to how social media companies should moderate content; instead, it nudges action through the public disclosure of information about the moderation policies employed by social media companies.

Specifically, A.B. 587 requires that affected companies publicly post their terms of service and submit semiannual reports to the California attorney general on company content moderation practices.

Twice a year, companies must submit to the attorney general a terms of service report that includes a list of prohibited content categories, a description of the company's moderation policies, and data on the flagged and actionable items. Companies that fail to submit reports timely or that submit incomplete reports may be subject to fines of up to $15,000 per violation per day.

The terms of service report shall include:

1. A copy of the current version of the terms of service, a complete and detailed description of any changes to the terms of service from to the previously filed version, if any, and a statement of how terms of service defines certain categories of content — namely, hate speech or racism, extremism, or radicalization, disinformation or misinformation, harassment or foreign political interference.

If the company's terms of service addresses the enumerated content categories, a report should also contain such definitions.

2. A description of the company's content moderation practices, including:

- Any existing policies intended to address the content categories;
- A description of how the company uses automated content moderation systems to enforce terms of service and when these systems involve human review;
- How the social media company responds to user reports of violations of the terms of service;
- How the social media company would remove or take action against content, users or groups that violate the terms of service, or what other broader action a company takes against users who violate the terms of service; and
- The languages in which the terms of service are not available, if the company offers product features in these languages.

3. Information on the flagged content categories, including all of the following:

- The total number of flagged items, actioned items and the total number of actioned items that resulted in action taken again the content, users or groups who violated the terms of service;

- The total number of actioned items of content that were removed, demonetized or deprioritized by the social media company;

- The number of times actioned items were viewed by users and shared before the action was taken;

- The number of times users appealed social media company actions taken on that platform and the number of reversals of social media company actions on appeal disaggregated by each type of action.

Structure of the Report

Information in the report should be structured in the following categories:

- The category of content;

- The type of content, such as posts, comments, messages, etc.;

- The type of content media, i.e., text, images, videos, etc.;

- How the content was flagged — by artificial intelligence, community moderator, user, etc.; and

- How the action was taken, i.e., artificial intelligence or a human moderator.

Timing and Deadlines

The new requirements of this law will go into effect Jan. 1, 2024. Thus, the first report covering activity within the third quarter of 2023 should be submitted by Jan. 1, 2024. The second report should follow by April 1, 2024, and cover activity within the fourth quarter of 2023. The third report should be submitted by Oct. 1, 2024.

Reports must be submitted electronically to the attorney general every six months: no later than April 1 of each year covering activity within the third and fourth quarters of the preceding calendar year, and no later than Oct. 1 of each year covering the first and second quarters of the current calendar year.

Publication

All reports submitted to the attorney general will be made available to the public in a searchable repository on the attorney general's website.

Getting Ready for the New Requirements

In order to comply with the new law, affected social media companies should review and clearly articulate their content moderation practices and establish tracking mechanisms to

collect data required for the reports.

Even though many big tech companies already conduct content moderation in some form, A.B. 587 requires that all companies that meet the threshold requirements develop and be ready to make available consistent content moderation practices regarding content categories.

Additionally, affected social media companies should start building internal processes for tracking and documenting any actions taken concerning such content.

A.B. 587 imposes a potentially significant new set of burdens on large social media platforms.

Prior to the law, such platforms largely could rely on the good Samaritan provisions set forth in Section 230(c) of the Communications Decency Act, which limited or barred liability for online service providers for actions "voluntarily taken in good faith" to restrict access to certain categories of content that the service provider "considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing or otherwise objectionable."

The flexibility accorded by Section 230(c) meant that service providers were free to take action on any content they deemed to be objectionable, without being required to develop consistent content moderation policies — far less publicly explain and disclose the logic behind content moderation practices.

Now, to comply with the law, social media companies must not only disclose their moderation policies, but also must track action taken concerning the content categories, and publicly report the data they have collected.

It is likely that many service providers consider this data to be quite sensitive; such as reporting the number of times users interacted with the content that was later taken down, or how long it took the platform to take any action on such content.

Moreover, the law does not impose any requirements on how quickly action must be taken on the content categories, yet the very act of reporting internal data may expose social media companies to unwanted publicity and force them to develop more stringent moderation practices.

For instance, relying fully on artificial intelligence in content moderation policies without escalating the item for human review, or harboring large amounts of content defined as disinformation or misinformation almost certainly would negatively affect public perceptions of the platform and its practices.

Another challenge posed by the law is that it does not offer any guidance on what content might constitute foreign political interference, disinformation or misinformation. Nor does it explain how social media companies might determine what information originates from foreign governments, contains false facts or intentionally misleads users.

Notwithstanding the challenges prevented by the new law, if the law holds up to judicial scrutiny, it does have the potential to usher in a new era of transparency for social media platforms.

Such transparency could facilitate the ability of members of the public and lawmakers to engage in a productive dialogue about broader public policy issues concerning content

moderation and the flow of disinformation. Whether A.B. 587 does in fact assist that dialogue remains to be seen.

Marc Mayer is a partner and Stacey Chuvaieva is an associate at Mitchell Silberberg & Knupp LLP.

[1] See NetChoice, LLC v. AG, Fla., 34 F.4th 1196 (11th Cir. 2022).