CHAMBER
OF PROGRESS

July 6, 2021

# A.B. 587 Would Tip Off Criminals, Foreign Trolls, White Supremacists, and Child Predators On How to Evade Internet Removals

*Bill would enable the very hate speech it's aimed at preventing;*
*like giving bank robbers a live map of police officer locations*

Senate Judiciary Committee
California State Capitol
Room 2187
Sacramento, CA 95814

Dear Members of the Senate Judiciary Committee:

We share your commitment to promoting a healthy Internet free of hate, hoaxes, and disinformation. But **A.B. 587**, which is currently before your committee, would actually tip off criminals, foreign trolls, white supremacists, and COVID deniers on how to evade removal of their incendiary content by online platforms. In short, it would hand the bad guys a playbook to spread the kind of hate speech which the bill is intended to deter. **We urge you to reject this legislation as written.**

Our organization, the Chamber of Progress (progresschamber.org), is a new center-left tech industry coalition promoting technology's progressive future. In order to promote healthy online communities, we believe it is vitally important for online services to be transparent about their terms of service and content moderation policies.

In fact, most major online services publish regular reports[1] detailing their content removals, demonetization, and downranking of certain types of content. And all major platforms already have content policies prohibiting or restricting hate speech, radicalization, disinformation, harassment, or foreign political interference (all categories specified by this bill).[2]

But A.B. 587 goes further than that, requiring platforms to publicly share the particular terms and methods they use for rooting out hateful and incendiary content. The bill would require online

[1] *Transparency Reports*, Twitter (Jul. 2021), https://transparency.twitter.com/en/reports.html; *Transparency Reports*, Facebook, (Jul. 2021), https://transparency.fb.com/data/; *Google Transparency Report*, Google, (Jul. 2021), https://transparencyreport.google.com/?hl=en.

[2] *Rules and Policies*, Twitter (Jul. 2021), https://help.twitter.com/en/rules-and-policies#twitter-rules; *Community Standards Updates and Protections*, Facebook (Jul. 2021), https://www.facebook.com/communitystandards/introduction; *Youtube Policies*, Google, (Jul. 2021), https://support.google.com/youtube/topic/2803176?hl=en&ref_topic=6151248

platforms to publicly reveal details of how they identify, downrank, and deter incendiary and sometimes illegal content -- including requiring disclosure of:

- "Any rules or guidelines regarding how a social media company's automated content moderation systems enforce terms of service and when these systems involve human review."
- "Any training materials provided to human content moderators intended to educate them…"
- "Any rules, guidelines, product changes, and content moderator training materials that cover how the social media company would remove individual pieces of content, users, or groups that violate the terms of service, or take broader action against individual users or against groups of users that violate the terms of service."

Practically, this section of A.B. 587 would mandate the disclosure of: patterns, keywords, and phrases that automated content review systems use to flag inappropriate content; the inner workings of software that automatically detects child sexual abuse imagery; and platforms' watchlist of coded language used by extremists and hate groups to avoid raising suspicion.

To illustrate the impacts of this legislation, consider how the disclosures above would tip off bad actors in each of the following areas:

| Type of Content | What A.B. 587 Would Expose… | Leading To… |
|---|---|---|
| **Drug trafficking** | Coded language used to describe drug trafficking operations | Traffickers developing new codewords, increasing trafficking |
| **Child predators** | How automated systems detect repeat child predators online | Predators altering their behavior, increasing predation |
| **Russian election trolls** | Language, campaigns, and themes used by foreign election interference operations | Development of novel interference strategies, increasing interference |
| **Scammers** | Patterns and keywords of repeat scams | Development of new scam strategies, with increased success rate |
| **Insurrection and Election "Big Lie"** | QAnon and insurrection-related language that sparks additional review | Conspiracy theorists developing new theories and language |

| | | | |
|---|---|---|---|
| **White supremacists** | %$#! | Watchlist of coded language for future racist attacks | New language that skirts the policy line, allowing planning of future activities |
| **COVID deniers** | 🚫 | Watchlist of COVID- or vaccine-related falsehoods that trigger review | Denial communities go deeper underground with new coded language |

For example, Facebook, Twitter or YouTube's content moderation policies might automatically trigger review of any posts using the phrases "SWP" (supreme white power), "ZOG" (Zion occupied government), "hidden enemy" (describing Covid-19 deep state conspiracy theories), "army of Jesus" (a Russian election troll meme), or "KPC" ("keeping parents clueless," child predator slang). **But if A.B. 587 required the companies to publicly disclose these exact terms, bad actors would simply adjust strategies.**

Over the past few months, our organization has worked with Democratic legislators in Republican-controlled state legislatures like Texas and South Carolina to successfully defeat Republican legislation that is remarkably similar to A.B. 587. These bills aimed to increase "transparency" around content moderation based on a false perception of anti-conservative bias. For that reason it is strange to see Democratic legislators in California import an extreme Republican strategy from other states.

No major online platform wants to become a haven for hate or harassment, and their content moderation practices enable them to stay one step ahead of bad actors. They are not perfect, but they want their services to provide a positive environment for most users.

But A.B. 587's mandate of publicly exposing the details of platforms' content policies would be like **handing bank robbers a live map of where every police officer is located.** It would help bad actors script exactly how to evade being detected or punished.

We urge you to reject A.B. 587 in order to prevent hate, hoaxes, crime, and lies from spreading even further on the Internet.

Sincerely,

Adam Kovacevich

Adam Kovacevich
CEO and Founder
Chamber of Progress