Content Moderation:
Censorship and Shadow Banning

Offield, Kelly Chase
October 7, 2021

This investigation explores what "content moderation" is, otherwise known as censorship or shadow banning. Censorship occurs on tech platforms when content is removed for ideological reasons. Shadow banning is simply covert censorship that attempts to keep the censored unaware that he/she is being censored. I investigate how censorship and shadow banning occurs, expose precisely who develops these techniques, and reveal the economic motives of these programs.

Content moderation has roots in the CVE community, or *counter-violent extremist* community. Many CVE groups have developed content moderation techniques, but this article will focus on Jigsaw, a special CVE owned by Google. It is important to point out that the CVE community is separate from the technology platforms, but the Jigsaw-Google duo is an exception. With the data advantage of the Google search engine, preferential treatment on the Youtube platform, and support from much of the CVE community, Jigsaw is a formidable foe to freedom of information, privacy, and social networking. Jigsaw also produces unscientific reports of hate crimes, far-right extremism, and violent white supremacy[1] in order to market their programs.

* For a review of Jigsaw's (unscientific) disinformation campaign, click here.[1]

* For a proper debunking of hate crime claims, click here.[2][3][4][5]
* A proper debunking of far-right extremism/terrorism claims, click here.[6][7]

Now for the origin of shadow banning and a peculiar relationship between Jigsaw-Google and the New York Times.

ShadowBanning from the Source

The 1st Program

Jigsaw partnered with the New York Times on "content moderation" for the newspaper's social media pages[8] and to test Jigsaw's first shadow ban program, called Perspective API[9]. At that time, the program simply deleted or hid comments that clients (such as the New York Times) found disagreeable. A client could use the text-based AI program to highlight certain comments that the client could more conveniently browse, rather than browse through all comments to find "toxic" ones. The Client could also allow the text AI program to simply delete or hide the comments rather than be involved in the process at all. If this second option were the case, then many comments would never be viewable to other users, and some users may not realize that his/her comment was hidden. This provides obscurity to the client and prevents the likelihood of user scrutiny. As this machine-learning (or text-based or speech-based AI) was improved, it expanded to include images and video as well as comments[9]. If a group has a platform or page that they want to implement content moderation on, they can use Perspective API to sift through comments before they are published on the site or have the program listen to video to determine if it needs to be taken down[9].

To summarize, Perspective API allowed platforms or specific pages to hide or delete comments/content, and even do so before those comments are published online (such as a Youtube comment section). If content is hidden, it would not acquire more likes or comments or any other form of feedback, because other users could not see the hidden content. This program works without informing the user's who's content is hidden or deleted.

Surprisingly, marketing videos of Perspective API exist, though some have been deleted recently. Here is a link that does not work. I found this as a "case study" justifying the use of Perspective API, where the New York Times was the client. This may not be the first time that I have

followed a suspicious story of the New York Times where the news agency deleted articles or pages[10]. For a link to a marketing video that works (for now), click here.

The 2nd Program

Jigsaw has another tool called Moderator, which is also open-sourced[11] (at least in it's earlier days) so denying it's existence has that hurdle to overcome. Oddly enough, Jigsaw's Moderator was also a tool that they developed in partnership with the New York Times[12]. This is the second documented case of the New York Times social media pages being the testing grounds for shadow ban software.

Moderator is a text-based AI "that leverages Perspective to prioritize comments"[11][12]. Not only is the CVE deleting comments with Perspective API, but is also prioritizing certain comments with Moderator. It is still unclear if Moderator bipasses base algorithms of a platform to prioritize comments or if machine likes are added to these chosen comments to boost them to the top of a page or comment section.

Technically, some components of content moderation (shadow banning) are a form of censorship, such as Perspective API where opinions are suppressed. Moderator is more difficult to classify since it is preferential treatment of other opinions, rather than deleting opposing views. Nonetheless, the two together produce a result where a diversity of ideas cannot exist, and free speech is an afterthought.

The deleting and suppressing of comments, along with the preferential treatment that other comments receive, is irrefutably a result of Jigsaw and their programs. However, how do we know if Jigsaw is using these programs for the better good of all? Is the CVE truly so far-left that it suppresses innocuous opinions?

Are conservatives and the right-leaning targeted?

Jigsaw's Justifications for Shadow Banning

Jigsaw's disinformation projects[13] use "data" from the Atlantic Council’s Digital Forensic Research Lab (DFRLab). The DFRLab collects data based off of definitions of disinformation from several groups, including Facebook[14]. Contrary to the Jigsaw narrative on far-right extremism in North America, the DFRLab ranks Russia and Iran as the top sources of

disinformation campaigns. The lab's findings do not necessarily support the specific campaigns that Jigsaw wages, but as I found with dozens of Jigsaw's citations, many were misinterpreted or misused entirely[1].

The DFRLab released two articles[15][16] of the Capitol riots where far-right extremism was characterized as being highly networked, and reaching millions of "sympathizers" – a claim I have thoroughly debunked. This claim sets up the CVE community neatly, so that they can tackle this problem with their software.

The DFRLab articles examined institutional outlets like Twitter, which were implied to be safer as opposed to the radical and dangerous "Parler, Gab, MeWe, Zello, and Telegram"[15][16]. How convenient for the CVE community, who want total network control. As with Jigsaw's narrative about the white supremacy issue, Jigsaw's narrative about disinformation leads the CVE to the conclusion that the world needs more network control from Google and other dominant platforms, as well as more authoritarian methods by the CVE community.

"The migration reiterates that the challenge of online extremism is not limited to any one platform but rather an entire, largely unregulated ecosystem with very few barriers to engage or disseminate content.”[16] This quote from the DFRLab highlights my claim about Jigsaw's and the CVE's conclusions; note their use of the term "unregulated".

Furthermore, the DFRLab uses the appeal to authority fallacy without releasing data: "The team at the Atlantic Council’s Digital Forensic Research Lab has conducted exhaustive research"[16]. This statement is the only proof that the lab produces for their claims, and it would appear that a single appeal to authority is all that Jigsaw requires to implement shadow ban methods on millions. Youtube also commits the same logical fallacy when the platform defends their news bar by characterizing those pages as "authoritative" sources[17]. It is important to point out that users do not have the option to choose which pages show up in the Youtube news bar.

Conclusion

If it is not obvious already, the CVE community is using a small minority (a few thousand annually for hate crimes in the US; and about a hundred or less annually for terrorism globally) of sick, solitary individuals to characterize everyone that dissents. Jigsaw uses a mere 35 individuals to justify the CVE's methods[1][12], and no empirical evidence is produced or cited. It is carefully planned wording to use "millions of sympathizers" when talking about these

issues. The CVEs need to rely on storytelling because data, quantitative analysis, and statistics will only disprove the CVE claims about far-right extremism and white supremacy.
They are not talking about protecting the world from extremists. They are talking about controlling people – innocent, normal people.

References

[1] Kelly C Offield. "Grand Manipulation: Google Edition". The ARKA Journal. https://advocate-for-rights-and-knowledge-of-americans-arka.ghost.io/grand-manipulation/

[2] Kelly C Offield. "Progressive Movements: How Unscientific and Harmful Are They?". The ARKA Journal. https://advocate-for-rights-and-knowledge-of-americans-arka.ghost.io/progressive-movements/

[3] https://ucr.fbi.gov/hate-crime/2019/topic-pages/offenders

[4] https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/violent-crime

[5] https://statisticalatlas.com/United-States/Race-and-Ethnicity

[6] Kelly C Offield. "Grand Manipulation: The Extremist Myth." The ARKA Journal. https://advocate-for-rights-and-knowledge-of-americans-arka.ghost.io/grand-manipulation-myths-of-the-software-developers/

[7] Institute for Economics and Peace. "Global Terrorism Index 2019 Measuring the Impact of Terrorism."https://www.visionofhumanity.org/wp-content/uploads/2020/11/GTI-2019-web.pdf

[8] Jigsaw partnership with New York Times. https://www.youtube.com/jigsaw

[9] Jigsaw. "Perspective API" https://www.perspectiveapi.com/how-it-works/

[10] Kelly C Offield. "Security State, Their Servants, and a Convenient Cyber Breech" The ARKA Journal. https://advocate-for-rights-and-knowledge-of-americans-arka.ghost.io/frightening-unions-and/

[11] Github. "Moderator" https://github.com/conversationai/conversationai-moderator

[12] Jigsaw. "Toxicity: Case Studies" https://jigsaw.google.com/the-current/toxicity/case-studies/

[13] Jigsaw. "Disinformation" https://jigsaw.google.com/the-current/disinformation/dataviz/

[14] DFRLab. "Dichotomies of Disinformation" Github. https://github.com/DFRLab/Dichotomies-of-Disinformation

[15] Atlantic Council's Digital Forensic Research Lab. "What's next for the insurrectionists" https://www.atlanticcouncil.org/content-series/fastthinking/fast-thinking-whats-next-for-the-insurrectionists/

[16] Atlantic Council's Digital Forensic Research Lab. "How the Capitol Riot was Coordinated Online" [https://www.atlanticcouncil.org/content-series/fastthinking/fast-thinking-how-the-capitol-riot-was-coordinated-online/

[17] Youtube. "Greater Transparency for Users Around." https://blog.youtube/news-and-events/greater-transparency-for-users-around/