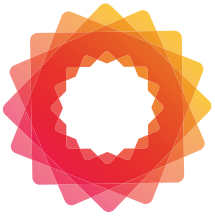
NUTS & BOLTS *of* CONTENT MODERATION

A primer for policymakers on content moderation

SEPTEMBER 2019

presented by:



Engine

CHARLES KOCH INSTITUTE

INTRODUCTION

The Internet has enabled individuals to easily consume, create, and share content with other users across the world. Indeed, we post, comment on, and share everything from political opinions, to cat videos, to product reviews every minute on platforms across the Internet. These Internet platforms rely on several things to host user content while maintaining the integrity of the platform, including their own tireless content moderation efforts as well as the current legal framework underpinning the Internet since 1996.

That legal framework, especially Section 230 of the Communications Decency Act, has come under attack from all sides in recent years. Some argue Internet platforms are doing too much content moderation, especially in ways that allegedly disadvantage specific political viewpoints. At the same time, others argue Internet platforms are doing too little content moderation and failing to keep users safe and in compliance of state and local rules. As policymakers think through whether there's a need to change this foundational Internet law, it's critical that they understand the ways in which all Internet platforms—not just the biggest two or three—rely on this framework to conduct their current content moderation practices.

In this report, and through a series of events in Washington, D.C. in the summer of 2019, Engine and the Charles Koch Institute sought to unpack the nuts and bolts of content moderation. We examined what everyday content moderation looks like for Internet platforms and the legal framework that makes that moderation possible, debunked myths about content moderation, and asked attendees to put themselves in the shoes of content moderators.

CONTENTS

GLOSSARY

Child Exploitation
The use of a child, online, for sexual purposes or to create child pornography.

Content ID
An algorithmic tool built by YouTube to assist the platform in identifying copyright-infringing content when it is uploaded by users. Copyright holders upload their protected works into a database maintained by the platform, and content uploaded by users is compared to the protected works in the database. When YouTube's algorithm detects a match, YouTube informs copyright holders who can make a claim to have the content removed from the platform or monetize the user-uploaded content.

Content Moderation
The way platforms monitor and apply a set of rules and guidelines to user-generated content to determine whether content created by their users can remain on the platform.

Digital Millennium Copyright Act (DMCA)
A law passed by Congress in 1998 that created rules, including a notice and takedown regime, dictating the way platforms must handle claims of copyright infringement over user-uploaded content.

Global Internet Forum to Counter Terrorism (GIFCT)
An initiative founded by Google, Facebook, Twitter, and Microsoft in 2017 to combat the spread of terrorist content and violent extremism propaganda online.

Hate Speech
Threatening or abusive speech that attacks a person on the basis of protected traits, such as race, ethnic origin, national origin, religion, sex, disability, or sexual orientation.

Notice and Takedown
The regime under which platforms are notified of an alleged copyright infringement in user-uploaded content and then disable access to the disputed content, or risk statutory damages under copyright law.

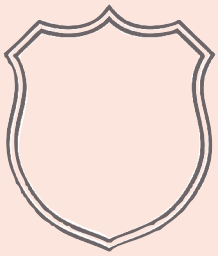
Platform
A website that hosts user-content, such as Reddit or Medium.

"Stop Enabling Sex Traffickers Act" and "Allow States and Victims to Fight Online Sex Trafficking Act" (SESTA-FOSTA)
SESTA-FOSTA is s law ostensibly designed to prevent platforms from facilitating sex trafficking by creating a carve-out in Section 230 that holds platforms legally liable for user speech if they knowingly assist, support, or facilitate illegal sex trafficking conduct.

WHAT IS SECTION 230?

Often described as the "26 words that created the Internet," Section 230 of the Communications Decency Act (47 U.S.C. § 230), passed in 1996, enabled online platforms to host user-generated content without being held legally responsible for the speech of their users. Section 230 shields websites from liability for content created and shared by users and gives platforms the ability to find and remove objectionable content without fear of legal action.

THE SHIELD:



Section 230(c)(1) says "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." In practice, this means an Internet platform can host user content without being held liable if a user creates and shares illegal content. The liability protections do not extend to set actions where the platform helps the problematic content.

THE SWORD:

Section 230(c)(2), often called "the Good Samaritan" provision, protects platforms from liability when they take action "in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected." In practice, this encourages Internet platforms to engage in responsible content moderation, since the implementation of moderation doesn't strip them of their liability protections.

Combined, the provisions of Section 230 are an incredibly important set of protections for startup Internet platforms that host user content. Startup Internet platforms have changed the way users express their thoughts through blog posts, edit and share photos, make decisions about purchases, and more. These types of platforms all rely on user-generated content, and Section 230 is the legal framework that enables small and new platforms to host users' content without having to hire content moderators to review every piece of content or rely on vast legal defense funds to fight back against lawsuits over user content.

WHAT DOES 230 DO?

- It does establish a uniform regulatory regime, rather than a 50-state patchwork.
- It does prevent frivolous litigation.
- It does empower platforms to proactively monitor for objectionable content.

WHAT DOES 230 NOT DO?

- It does not immunize platforms from liability under federal criminal law.
- It does not protect a platform from liability if it develops illegal content.
- It does not apply to intellectual property law.

MYTH VS. FACT

MYTH 1:

Myth: Content moderation is easy. Harmful content is obvious. In fact, it's so straightforward that algorithms can do it well.

Fact: Moderating user content is incredibly difficult. Even when a platform has clear rules about the user speech it will and will not host, many instances of user speech fall into a gray area, where factors that are intrinsically difficult to codify—such as context, past user behavior, and an ever-evolving cultural lexicon—will help determine whether user speech violates those rules. Users will also inherently disagree about whether certain speech should be allowed at all. The fact that major Internet platforms are currently under attack simultaneously for both doing "too much" and "too little" content moderation proves that there's no one-size-fits-all that will make everyone happy.

MYTH 2:

Myth: Platforms want to keep problematic content online.

Fact: Platforms do not benefit when users create and share illegal and otherwise problematic speech. Problematic content causes users and advertisers to abandon the platform, making it bad for the platform's reputation and bad for business.

MYTH 3:

Myth: Section 230 was written as a new and specific giveaway to the tech industry.

Fact: Section 230 is centered around the idea that the entity that makes others' speech available to the public is not specifically aware of, and therefore not legally liable for, everything each speaker says. That idea far predates today's tech industry and can be found throughout U.S. legal history, including in a 1950 Supreme Court ruling, which held that a Los Angeles ordinance—stating that if you had obscene material in your store, you would be held criminally liable—was unconstitutional.

MYTH 4:

Myth: Under the law, platforms must be neutral to receive Section 230 liability protections.

Fact: Section 230 was written to encourage platforms to responsibly moderate their users' content. After a series of conflicting court decisions—including a decision against an online bulletin board that disincentivized content moderation—Congress didn't want Internet platforms being sued for user speech whether they moderated content or not. The law includes a general protection against being held liable for user speech as well as the "Good Samaritan" provision, which protects a platform against liability if it engages in "good faith" moderation of content the platform finds "objectionable." The Good Samaritan provision does not require perfect or neutral content moderation to justify Section 230 protections.

MYTH 5:

Myth: Section 230 was written when the Internet was new. Today's Internet platforms don't need it anymore.

Fact: The most ubiquitous Internet platforms are now large, global companies, but they still rely on Section 230, as do all of the small companies that offer platforms for users to share content, whether that's websites that host consumer reviews, newspapers' online comment sections, or photo-sharing apps. Section 230 ensures that all Internet platforms won't be held legally liable for the speech of their users, and that reassurance is what allows startups that host user content to create, grow, and get funding for new and innovative ideas. Without Section 230, an Internet platform could face the threat of lawsuits over any piece of user-generated content it hosts, which would require hundreds of thousands of dollars to defend against. No startup, and no investor, would choose to take on that risk.

MYTH 6:

Myth: Section 230 protects platforms from criminal liability.

Fact: Section 230 does not protect a platform from criminal liability if it violates federal law. In fact, the Justice Department shut down the notorious website Backpage.com before SESTA-FOSTA was signed into law, using existing authority to enforce federal criminal laws against sex trafficking. A website that violates federal law, or substantially contributes to the development of user content that violates federal law, can be prosecuted under federal law. A court ruled in 2007 that Roommates.com—a platform for connecting with potential roommates—violated federal housing laws by including a drop-down menu where users could express preferences about the protected characteristics of potential roommates.

MYTH 7:

Myth: Platforms are violating the First Amendment by censoring users' speech, and changing Section 230 will put an end to that.

Fact: First Amendment protections only extend to government regulation of speech. Private companies have no obligation to host any speech they choose not to. On the contrary, private companies have their own right to decide what type of speech to host; an online encyclopedia can refuse to host opinionated content just as much as a comment board about dogs can refuse to host pictures of cats.

Changing Section 230 to weaken platform liability protections will actually threaten free speech online, because it will force platforms to become legally liable for the speech of their users. If a platform has to worry about being held liable for everything its users say, it has an incentive to over-moderate any content that could result in a lawsuit. At the same time, Section 230 actually creates more and more diverse opportunities to speak online because the law enables platforms to launch without having to build a deep-pocketed legal fund to defend against lawsuits over content created by their users.

YOU MAKE THE CALL

CNN tweets an article covering a new study about "deepfake" videos and the technology being built to combat the spread of misinformation through deepfakes.

The tweet includes a clip of a deepfake video that makes it look like Democratic presidential candidate Elizabeth Warren is saying words that were attributed to her during an impersonation on Saturday Night Live.

The video clip from CNN's tweet ends up circulating throughout the conservative media. A Republican Congressman sees it but isn't aware that the clip was from a deepfake that combined Warren's features with an impersonator's words. He uses the impersonator's words to make the case against Warren as a presidential candidate.

Twitter Terms of Service

You may not do any of the following while accessing or using the Services: (iv) forge any TCP/IP packet header or any part of the header information in any email or posting, or in any way use the Services to send altered, deceptive or false source-identifying information;

ELECTION INTEGRITY POLICY

What is not a violation of this policy?
Not all false or untrue information about political events constitutes manipulation or interference in an election. In the absence of other violations, the following are generally not in violation of this policy:

- inaccurate statements about an elected official, or political party;
- organic content that is polarizing, biased, hyperpartisan, or contains controversial viewpoints expressed about elections or politics;

DID YOU TAKE IT DOWN?

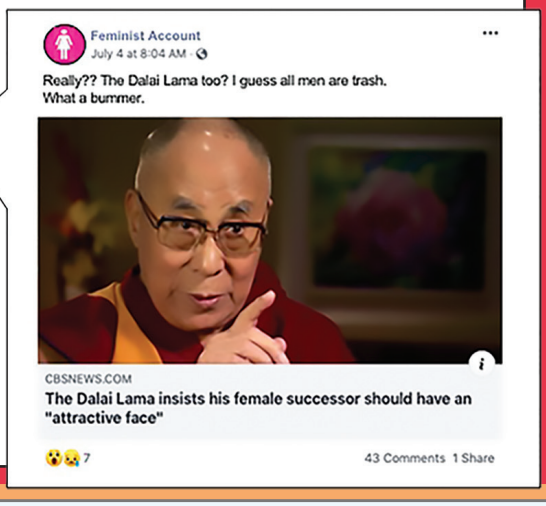
IF YOU **LEFT IT UP...**

Elizabeth Warren's campaign Twitter account pulls all of its advertising dollars from the platform. The account sends out a tweet accusing Twitter of not taking political misinformation and election interference, including from other countries, seriously. The tweet also threatens action if Warren wins in 2020.

IF YOU **TOOK IT DOWN..**

The Republican member of Congress accuses Twitter of censoring conservative speech. He threatens to hold a hearing and introduce legislation to hold the tech industry accountable for their actions that have stifled conservative speech and their overall anti-conservative bias.

A popular Facebook account that posts feminist memes shares a CBS News article about the Dalai Lama saying his female successor should have an attractive face.
The account holder adds commentary when sharing the post, including the line: "I guess all men are trash."



A Democratic Congresswoman shares the account's post and adds commentary indicating that she's amplifying a constituent's feelings. She includes the line: "It's no wonder people feel that men really are trash."

Community Standards

11. Hate Speech

Policy Rationale

We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence.

We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We also provide some protections for immigration status. We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation. We separate attacks into three tiers of severity, as described below.

Do not post:

Tier 1 attacks, which target a person or group of people who share one of the above-listed characteristics or immigration status (including all subsets except those described as having carried out violent crimes or sexual offenses), where attack is defined as

- Any violent speech or support in written or visual form
- Dehumanizing speech such as reference or comparison to:
  - Insects
  - Animals that are culturally perceived as intellectually or physically inferior
  - Filth, bacteria, disease and feces

Sometimes people share content containing someone else's hate speech for the purpose of raising awareness or educating others.... ie: a breastfeeding group for women only...we allow the content but expect people to clearly indicate their intent, which helps us better understand why they shared it. Where the intention is unclear, we may remove the content.

DID YOU TAKE IT DOWN?

IF YOU **LEFT IT UP...**

The Facebook account of a popular website that posts content predominantly for men shares the Congresswoman's post that includes, "men really are trash." The account points out that the Congresswoman's post is technically in violation of Facebook's policy against hate speech and accuses the company of enforcing its rules inequitably. The post goes viral.

IF YOU **TOOK IT DOWN..**

The Congresswoman posts on Facebook in frustration that her post was taken down, accusing Facebook of failing to provide a platform for its users to discuss gender equality issues. She uses the litter emoji in her post, and the litter emoji begins trending across the platform accompanied by users' descriptions of when they feel like they were silenced on Facebook while trying to discuss gender issues.

A DAY IN THE LIFE OF SECTION 230

Our current legal framework for content moderation has led to the creation of an entire ecosystem of online platforms that users rely on every day. Thanks to Section 230, Internet platforms can host user content—and moderate it as they see fit—without having to worry that they'll be legally liable for what users say. Below is an example of the various ways a Hill staffer might interact with platforms that rely on Section 230 throughout a typical day.

7:30 a.m. Read newest post on **Reddit's** Game of Thrones subreddit after last night's finale while eating breakfast.

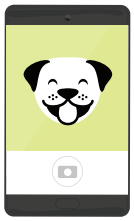
8:00 a.m. Check **Yelp** for reviews of the venue for tonight's happy hour for staffers from your lawmaker's state. Remind folks of the time and address for the happy hour in the **Facebook** group for staffers working for lawmakers in that state.

8:30 a.m. Open up **Stitcher** and search for a podcast to listen to as you commute to work. When the podcast host asks listeners to support the podcast, pull up your **Patreon** account to donate $5 to the podcast. Fire off a quick post on **Nextdoor** when you pass by a new coffee shop in the neighborhood that's opening today and offering free pastries.

9:00 a.m. Get to your desk and use **Google's** search engine to find the bill number and text of the bill on the floor this afternoon. Take to your boss's **Twitter** account to explain why she'll be voting against the bill later today.

10:30 a.m. Use **Eventbrite** to find an event this afternoon on the Hill that offers free lunch and an interesting panel discussion. Look on **LinkedIn** to learn about the background of the person you're meeting for coffee at 11 a.m.

11:45 a.m. As you wrap up coffee and head to the lunch event, check your **Snapchat** to find a photo your friend sent you of their dog, and check to see if anyone has responded to your **Craigslist** posting about an open room in your group house on Capitol Hill.



12:45 p.m. During the event's panel discussion, use **Wikipedia** on your phone to look up a term one of the panelists used that you've never heard before, and bookmark a post on **Medium** that another panelist recommends reading to learn more about the topic.

3:00 p.m. Post your boss's floor speech opposing the bill on the floor to **YouTube**. Find your boss's local newspaper's coverage of the floor vote and read through the newspaper's comment section on the article about your boss's vote.

5:00 p.m. As you head over to happy hour, scan through reviews on **Goodreads** before you pick a book to order on **Amazon**.

8:00 p.m. On your way home, post a group photo from happy hour on **Instagram** and look for inspiration for what to make for a potluck dinner this weekend on **Pinterest**.

Section 230 and the liability protections it provides for Internet platforms of all sizes have come under scrutiny from all sides recently as policymakers take issue with different kinds of user content and the platforms that host it.

CURRENT LAW

The first major move against Section 230 was a pair of bills signed into law in 2018 ostensibly aimed at curbing sex trafficking online. The law—the combined Fight Online Sex Trafficking Act (FOSTA) and the Stop Enabling Sex Traffickers Act (SESTA)—created an exemption to Section 230's liablity protections for Internet platforms that "knowingly" facilitate illegal sex trafficking. While the law sounds laudable, it created uncertainty for platforms that host content that could potentially be sex trafficking. As a result of the law, platforms like Craigslist removed some content to avoid potential lawsuits where they would be on the hook for content uploaded by users. At the same time, the Department of Justice was able to remove Backpage.com—a site notorious for helping users skirt the law and the site's rules about sex trafficking—before FOSTA-SESTA was signed into law.

NEW PROPOSALS

Section 230 has been a popular political punching bag. In some cases, lawmakers accuse platforms of moderating too much content and suppressing certain political views. At the same time, some lawmakers say they're frustrated that platforms aren't moderating content more aggressively, pointing to the spread of political misinformation, hate speech, and online extremism. Lawmakers have put forward several proposals to address these alleged concerns, from requiring platforms to certify that their moderation practices are politically neutral, to prohibiting the spread of manipulated political content, to deputizing platforms to enforce state and local laws.

TRADE

The protections found in Section 230 have also played an important role in recent U.S. trade policy as the country renegotiated the North American Free Trade Agreement with Canada and Mexico. The resulting trade agreement, the United States-Mexico-Canada Agreement (USMCA), has incorporated several aspects of U.S. law, including Article 19.17, which mirrors Section 230. The inclusion of liability protections for platforms will lower barriers and strengthen market access for startups. While the USMCA still needs to be ratified by Canada and a vote of Congress, if enacted, it will be the new standard for trade agreements in the digital age, serving as the blueprint for future U.S. free trade agreements, and exporting innovation-advancing platform liability protections to the world.

Engine was created in 2011 by a collection of startup CEOs, early-stage venture investors, and technology policy experts who believe that innovation and entrepreneurship are driven by small startups, competing in open, competitive markets where they can challenge dominant incumbents. We believe that entrepreneurship and innovation have stood at the core of what helps build great societies and economies, and such entrepreneurship and invention has historically been driven by small startups. Working with our ever-growing network of entrepreneurs, startups, venture capitalists, technologists, and technology policy experts across the United States, Engine ensures that the voice of the startup community is heard by policymakers at all levels of government. When startups speak, policymakers listen.

For more than five decades, Charles Koch's philanthropy has inspired bold new ideas to improve American lives. Inspired by a recognition that free people are capable of extraordinary things, the Charles Koch Institute supports educational programs and dialogue to advance these principles, challenge convention, and eliminate barriers that stifle creativity and progress. We offer educational programs, paid internships, and job placement assistance to students and professionals, and encourage civil discussion about important issues like free speech, foreign policy, and criminal justice reform. In all of our programs, we are dedicated to identifying new perspectives and ideas that help people accomplish great things for themselves and others.