OCTOBER 2020

Election 2020: Content Moderation and Accountability

By Marietje Schaake and Rob Reich

AS WE APPROACH THE 2020 ELECTION IN THE UNITED STATES, content moderation on social media platforms is taking center stage. From speech issues on Facebook and Twitter to YouTube videos and TikTok brigands, the current election season is being reshaped by curation concerns about what's allowed online, what's not, upranking and downranking, and who's deciding.

Trust in the content is another major challenge as conspiracies and mis- and disinformation go viral. With billions of pieces of content posted every day, what balance should be struck between automated and human moderation? Are AI and machine learning to blame when companies miss content they promised to remove, or do we need to look to human content moderators and those sitting in their board rooms?

Content moderation is the practice of social media platforms, governments and regulators making and enforcing the rules about what content is or is not permitted on their services. This issue and others are being examined in the eight-week Stanford University course, "Technology and the 2020 Election: How Silicon Valley Technologies Affect Elections and Shape Democracy." The joint class for Stanford students and Stanford's Continuing Studies Community enrolls a cross-generational population of more than 400 students from around the world.

KEY TAKEAWAYS

- Social media companies require greater accountability and transparency about which content is removed, labeled or modified, or left online, upranked or downranked, and how these decisions are made.
- The opaque decision-making process at these private firms with global networks raises concerns of accountability, transparency, and how their commercial interests affect their stewardship of a digital civic square.
- A key governance issue to explore is whether the social media companies – and democratic lawmakers and regulators – have abdicated responsibility for content moderation.

Stanford University
Human-Centered
Artificial Intelligence

Stanford | Cyber Policy Center
Freeman Spogli Institute

With guest experts Nick Pickles and Evelyn Douek, the third class session on Oct. 7 explored platform effects, content moderation, and free speech online. Pickles is Twitter's senior director of public policy strategy and development, and Douek is a lecturer on law and doctoral candidate at Harvard Law School.

Introduction

All U.S. social media companies have content moderation standards that would violate the First Amendment, if the moderation were practiced by a government agency. Companies routinely take down nudity, hate speech, messages from suspected terrorists and sex traffickers, and more. Despite this, social media companies permit, in the name of free expression, far more than what is allowed by other countries, especially in Europe.

Most other countries have greater limits on free expression than the U.S., including for example, restrictions on Holocaust denial, incitement to violence, and hate speech. The opaque decision-making process at these private firms with global networks raises concerns of accountability, transparency, and how their commercial interests affect their stewardship of a digital civic square.

When YouTube took down a video of Marietje Schaake's 2016 parliamentary debate on banning the trade in goods used for torture and the death penalty, she searched for answers. Was this decision made by a human or AI machine? Why had this innocuous debate suddenly set off red flags at YouTube? She contacted YouTube, and her video re-appeared a few hours later. But Schaake's questions were not answered, thereby illustrating how difficult it is for a citizen user to get a credible response from social media firms about content moderation.

The debate is evolving on the roles of platforms in governing speech and moderating content. In the early days of social media, most companies were reluctant to moderate any content. Many companies referenced the First Amendment – Facebook declared it did not want to be "the arbiter of truth."

Critics such as Maria Ressa, a journalist in the Philippines, described this "an abdication of responsibility," observing how speech can get "weaponized." She also raised the issue about whether democratic lawmakers and regulators have likewise abdicated responsibility by not acting.

When the European Commission approved codes of conduct regarding terrorist content and fake news and considered the possibility of legislation, the U.S.-based tech companies took a slightly different tack on content moderation. However, this meant public authorities essentially outsourced thorny content decisions back to the very platforms themselves (whose power was considered outsized in the first place).

Then in 2016, the backlash of the Cambridge Analytica scandal spotlighted the problem of foreign interference in elections through social media platforms. And when politicians found the platforms to be an amplification mechanism for their own untruths, Twitter took a leading role in labeling misleading or hateful tweets, prompting direct attacks by President Trump, who reacted by wanting to change Section 230 of the Communications Decency Act.

In 2019, Twitter announced it would end all political ads, and while Facebook initially ruled such a move out, it has now promised a pause after the polls close. Facebook also recently announced that it will take down an unprecedented number of QAnon Pages, Groups and Instagram accounts, while still allowing so-called “organic content.” Nearly all platforms eagerly foregrounded authoritative information about COVID-19 and the administration of elections.

> *One issue are ‘content cartels,’ which can occur when the pressure for social media firms to do something leads to the creation of systems that serve the interests of the very tech platforms themselves without adequate public oversight.*

Discussion

A key issue is how social media companies are able to detect “coordinated inauthentic behavior” on their platforms. The problem is that there’s no clear definition of what constitutes this type of content – defining what is coordination and what is inauthentic is far from a value-free judgment call. The lack of clarity, transparency and definitions especially matters during election cycles.

While basic principles of content moderation standards are made public, the balance between automated and human moderation is not, the core algorithmic curation of content is opaque, and the entire operation is significantly driven by commercial incentives. Transparency and accountability issues inevitably arise.

A starting point is that there are no coordinated or global standards across companies like Twitter, Facebook, Instagram, Google, YouTube, and TikTok. We are in the nascent stage of working out the norms for acceptable online political communication. An open and public debate is urged on questions such as what constitutes an inauthentic “network” or “fake” content?

One issue is the arrival of what Douek calls “content cartels,” which can occur when the pressure for social media firms to do something leads to the creation of systems that serve the interests of the very tech platforms themselves without adequate public oversight.

While moderation is increasingly performed through AI-driven tools – reflecting a desire to take action faster to prevent harm – how can one ensure that some advanced technologies used to scan large volumes of texts, images and video's aren't exclusive to a few large companies with the resources to devote to the problem? An increasing reliance automation might cement the competitive advantages of large and established companies. Then again, so too might a reliance on human moderation, which requires potentially thousands of employees. How can we preserve a competitive marketplace?

Another issue is how to establish a governance framework or a "guardrails" approach across major companies. However, applying standard approaches across platforms, large and small, and on a global scale, may be incredibly difficult, especially when "closed doors" and a lack of transparency characterize decision-making at the social media companies. And, some say that if these companies were more transparent with their content moderation policies, it would be easier for bad actors to game the system.

About 50 percent of the content Twitter removes is done through automation, and the other 50 percent through removal by human means. On the latter, for example, the video of the Christchurch massacre in New Zealand made instant headlines, but social media companies took a long time to contain the spread of the grotesque and hateful content. Still, the video can be found on many platforms.

Social media companies should aspire to enact policies that are tightly-defined, clearly communicated, and enforced in transparent ways so users and policymakers are well aware of the rules of the road. And democratic governments should assume greater responsibility for establishing frameworks of transparency and accountability.

Content moderation isn't just about individual pieces of content. It's often about patterns of behavior, and examining how an account behaves – and what is the best intervention among a range of actions that could be taken. Beyond removing content, it can be demoted, demonetized, or made more difficult to share and contribute to virality. The speed or accuracy of content moderation is an issue, as artificial intelligence is increasingly used but is not always successful in identifying troubling content, especially when it is unusual or novel. Additionally, context is often critical in determining if content is satire or hateful, or a user is legitimate or impersonating someone else.

Meanwhile, misinformation issues involving the coronavirus pandemic has further escalated the issue of content moderation. Platforms recently have taken assertive approaches in removing such content and boosting trusted information from sources like the World Health Organization.

A clear need exists for real-time scientific research on the effects of new technology and content moderation on society and the political debate. This type of research could lead to better understanding and evidence-based policymaking. In Europe, for example, Facebook has invited scrutiny because the company failed to adequately address concerns brought to their attention. From their impact on election outcomes, to spreading of conspiracies and hate speech, their message has been that regulation would stifle innovation. Only recently they changed direction and have asked for regulations, while still hiring large numbers of lobbyists.

Final Thoughts

As social media platforms grow in terms of users, they inevitably face pressure to moderate content or risk public backlash when controversial posts arise. While some at the outset may try to differentiate themselves by adopting policies with minimal intervention, these policies often converge towards industry standards.

The rising threat of foreign and domestic interference in U.S. elections since the 2016 election comes replete with new tactics, tools, and vulnerabilities.

In terms of best practices, social media companies should aspire to enact policies that are tightly-defined, clearly communicated, and enforced in transparent ways so users and public policymakers are well aware of the rules of the road.

However, for the public interest to be strengthened, along with oversight and accountability, democratic governments need to act. Both companies and governments are advised to create clarity to ensure that data access for outside research is properly valued and guaranteed.

The rising threat of foreign and domestic interference in U.S. elections since the 2016 election comes replete with new tactics, tools, and vulnerabilities. As a result, questions continue to arise about how to hold social media companies accountable for their impact on political and civic discourse.

ELECTION 2020: CONTENT MODERATION AND ACCOUNTABILITY

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI), applies rigorous analysis and research to pressing policy questions on artificial intelligence, particularly human-centered AI technologies and applications. For further information, please contact HAI-Policy@stanford.edu.

The Cyber Policy Center at the Freeman Spogli Institute for International Studies is Stanford University's premier center for the interdisciplinary study of issues at the nexus of technology, governance and public policy.

The views expressed in this issue brief reflect the views of the authors.

Rob Reich is a professor of political science at Stanford University. He is also the associate director of Stanford's Institute on Human-Centered Artificial Intelligence, director of the Center for Ethics in Society, and faculty co-director of Stanford's Center on Philanthropy and Civil Society. His new book, *Digital Technology and Democratic Theory*, will be published in December.

Marietje Schaake is the international policy director at Stanford's Cyber Policy Center and international policy fellow at Stanford's Institute for Human-Centered Artificial Intelligence. President of the Cyber Peace Institute, Marietje served between 2009 and 2019 in the European Parliament, focusing on trade, foreign affairs and technology policies.

Stanford University
Human-Centered
Artificial Intelligence

Stanford | Cyber Policy Center
Freeman Spogli Institute

Stanford HAI: Cordura Hall,
210 Panama Street, Stanford, CA 94305-1234
T 650.725.4537 **F** 650.123.4567
E HAI-Policy@stanford.edu **hai.stanford.edu**

Stanford Cyber Policy Center: Encina Hall
616 Jane Stanford Way, Stanford, CA 94305-6055
T 650.724.6814 **E** cyber-center@stanford.edu
stanfordcyber.org