Session III: The Way Forward: Regulation, Self-Policing, and Other Avenues

Karen Kornbluh, "The Internet's Lost Promise and How America Can Restore It," *Foreign Affairs,* September/October 2018 Issue

John Villasenor, "Artificial intelligence, deepfakes, and the uncertain future of truth," *The Brookings Institute,* February 14, 2019

Nina Jankowicz, "Election Interference: Ensuring Law Enforcement Is Equipped to Target Those Seeking to Do Harm," *Woodrow Wilson International Center for Scholars, Kennan Institute,* June 12, 2018

Naja Bentzen, "Online disinformation and the EU's response," *European Parliamentary Research Service,* February 2019

Věra Jourová, "Code of Conduct on countering illegal hate speech online," *European Commission for Justice, Consumers and Gender Equality*, February 2019

Richard Allan, "Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?" *Facebook Newsroom,* June 27, 2017

Foreign Affairs

The Internet's Lost Promise and How America Can Restore It

by Karen Kornbluh
September/October 2018 Issue

In the United States, Russia sought to help one presidential candidate over another in the 2016 election—not only through hacking and the release of e-mails but also through an extensive information operation that included paid ads, fake social media accounts, and divisive content. In China, authorities are harnessing the power of artificial intelligence to perfect an Orwellian system of online and real-world surveillance to track citizens' every move. In Myanmar, a UN rapporteur found that Facebook had helped spread hate speech, contributing to the ethnic cleansing of Rohingya Muslims. At a time when fully half of the world's population is connected to the Internet, it is hard to escape the conclusion that the technology that promised to give power to the powerless has ended up also hurting the very people it was supposed to help.

Openness allowed the Internet to become a global network that has fostered extraordinary innovation and empowered entrepreneurs, consumers, and political organizers. But along the way, some of the openness was lost, and darker forces took root.

Today, large technology companies have come to dominate the online experience, constantly gathering users' personal data, often without their knowledge, and feeding it through proprietary algorithms to curate search results, recommendations, and news. Propagandists and extremists wishing to conceal their identities fund targeted ads and create armies of social media bots to push misleading or outright false content, robbing citizens of a basic understanding of reality. And authoritarians take advantage of technology to censor information and suppress dissent.

The United States invented the Internet, and from the beginning, it promoted its vision of an open and free Internet on the global stage. But today, U.S. leadership is largely absent as the platform is increasingly being weaponized. It's time for Washington to overcome its techno-utopian belief that the Internet can fix itself and instead take active steps to ensure that the Internet is a tool to strengthen, not undermine, democratic values.

From Hope to Disappointment

The most commonly told origin story of the Internet starts with the brilliant young entrepreneurs who invented life-changing technologies from inside their garages. In reality, the early Internet received significant help from the U.S. government. It grew out of ARPANET, the Advanced Research Projects Agency Network, a decentralized network created by the Pentagon that was designed to withstand a nuclear attack. The inventors of the Internet Protocol and the World Wide Web received government grants and support from government research labs.

Moreover, in the mid-1990s, when the Internet was beginning to enter people's homes and workplaces, the U.S. government aggressively promoted competition with the existing telecommunications network, a choice that allowed the early Internet to flourish. The Federal Communications Commission exempted Internet service providers, such as AOL, from paying the charges that long-distance carriers had to pay and implemented the Telecommunications Act of 1996 in a way that, for a few years at least, opened the regional phone companies up to competition, stimulating billions of dollars of spending on the deployment of broadband networks. When Congress passed the 1996 Communications Decency Act, it included a provision—Section 230—that largely freed certain Internet companies from liability for third-party content posted on or moving across their networks or platforms. Combined with the decentralized design of the Internet, these policies promoted a medium that allowed users to exchange information freely.

The United States proselytized its pro-openness policy framework abroad. In 1997, Washington negotiated an agreement through the World Trade Organization that committed 67 signatory countries to "procompetitive regulatory principles" when it came to telecommunications, paving the way for the global Internet. And to set the rules of the road for the Internet, it endorsed a handful of "multistakeholder" organizations, including the Internet Corporation for Assigned Names and Numbers, or ICANN (which manages the domain name system), and the Internet Engineering Task Force (which promotes technical standards). This framework promoted competition, provided new avenues for sharing information, and allowed the Internet to become a vibrant platform for free expression and innovation. The Internet seemed to be ushering in a new era of democratization and entrepreneurship. By 2011, it was being credited with causing the Arab Spring.

But by then, the Internet had changed greatly. Early on in its history, users communicated directly, and e-mail was the "killer app." With the advent of the World Wide Web, users could easily generate and share their own content. But today's digital platforms—including Amazon, Facebook, Google, and Twitter—use algorithms to organize the user experience. Social media companies earn more ad revenue the longer they can get people to spend on their platforms and the more narrowly they can target them, and so they have every incentive to gather as much data as possible and feed it into algorithms that optimize the content their users see.

At the same time, the offline world moved online. In a 2017 survey of Americans conducted by the USC-Annenberg School for Communication and Journalism, respondents admitted spending an average of 24 hours a week online. Forty percent of them said they thought the Internet plays an integral role in American politics, and 83 percent reported that they shopped online. Most of the relevant government policies were designed when the Internet was just a fringe part of people's lives, but it has come to touch nearly every aspect.

News also moved online, with more people now getting it through the Internet than from television, as did advertising. As a result, print journalism's economic model fell apart. In the past, when the future of news seemed in question, Americans publicly debated what role media should play in a democracy. Congress regulated growing forms of media, with the 1927 Radio Act and then the 1934 Communications Act requiring broadcasters to act in the public interest as a condition of their receiving licenses to use the public airwaves. Civil society joined the debate,

too. After World War II, the Commission on Freedom of the Press, led by Robert Hutchins, the president of the University of Chicago, concluded that mass media must be committed to social responsibility. And in 1967, the Carnegie Commission on Educational Television issued a report on how to bring public broadcasting to U.S. households, spurring the passage that same year of the Public Broadcasting Act, which established the Corporation for Public Broadcasting. But when the Internet took off, no such examination took place.

In short, as the Internet grew more centralized and as its role expanded, policymakers failed to keep up. When it came to updating regulations for online activities—whether the matter at hand involved political advertising or privacy—the Internet was treated as a special realm that did not need regulation. And the bad guys took notice.

Digital Dictators

In the heady days of the Arab Spring, some observers believed the Internet gave dissidents a distinct advantage over their oppressors. But the despots largely learned to use the technology for their own ends. It turned out that even though social media and other technologies can help protesters, they can also help the state.

A 2017 report by Freedom House found that Internet freedom had declined globally for the seventh year in a row as China, Russia, and some Gulf states deployed a number of sophisticated methods for restricting access to online information and to communications tools. They have blocked virtual private networks, making it harder for users to evade censorship controls, and they have done the same with encrypted messaging apps such as Telegram, robbing dissidents of the ability to organize confidentially. In the Philippines, President Rodrigo Duterte has enlisted an army of paid online followers and bots to project an atmosphere of public enthusiasm and intimidate his critics.

Sometimes, autocrats even get private companies to do their bidding. The Turkish government, in the midst of a crackdown on opposition since a failed coup attempt in 2016, has forced Facebook to remove content. (Wikipedia left the country rather than edit or remove content.) And in some countries—notably China, Iran, and Russia—governments require that citizens' data be kept in the country.

The most sophisticated effort comes from China, which, in addition to its Great Firewall, is developing a system of "social credits," which takes the idea of a credit score to its creepiest extension. The idea is to aggregate information from public and private records to assess citizens' behavior, generating a score that can be used to determine their opportunities for employment, education, housing, and travel.

The United States has struggled to respond to the online authoritarian threat. As secretary of state, Hillary Clinton championed an Internet freedom agenda to empower dissidents. The State Department devoted tens of millions of dollars to programs aimed at enhancing Internet access, fighting censorship, and creating technologies to circumvent controls. And in 2016, it established the Global Engagement Center, which was charged with coordinating efforts to counter propaganda spread by states and nonstate actors alike. But that organization has never been fully

staffed or fully funded. All the while, the tools for surveillance and control have grown more sophisticated.

Hacking Democracy

Not only has the Internet been used to strengthen authoritarian states; it has also been used to weaken democracies. As detailed in the indictments issued in February by Robert Mueller, the U.S. special prosecutor investigating Russian interference in the 2016 election, Russian operatives created fake online personas aimed at spreading false information. For example, a Twitter account by the name of @TEN_GOP purported to represent the Tennessee Republican Party and posted a steady stream of content supporting Donald Trump, the Republican nominee. In fact, it was run by the Internet Research Agency, an organization linked to the Russian government that is responsible for online influence operations. A particular goal was to depress African American turnout in order to hurt Clinton's candidacy. As an investigation by CNN found, one social media campaign called "Blacktivist" was actually a Russian troll operation; it had more "likes" on Facebook than the official Black Lives Matter page.

Those who organize disinformation campaigns on social media exploit commercial data-gathering and targeting systems. They sweep up personal data from a host of sources across different devices and categorize people by their behavior, interests, and demographics. Then, they target a given segment of users with ads and bots, which encourage users to like pages, follow accounts, and share information. In this way, disinformation campaigns weaponize digital platforms, whose algorithms seem to reward outrage because that is what keeps users engaged. As the scholar Zeynep Tufekci has found, YouTube's recommendation algorithm steers viewers toward increasingly radical and extremist videos.

To be fair, the big technology companies have begun to wake up to the scale of the problem. After the consulting firm Cambridge Analytica was found to have collected the personal information of 87 million Facebook users for use in political campaigns, Mark Zuckerberg, the company's CEO, testified in Congress that Facebook would extend worldwide the controls it is implementing to satisfy the EU's General Data Protection Regulation. (But the company's removal of non-European data from European servers, which puts the information out of reach of EU regulators, raises doubts about his commitment.) Twitter has begun removing fake accounts at an accelerated rate, deleting 70 million suspicious accounts in May and June 2018. All these companies have taken steps to increase transparency when it comes to who has paid for a particular political ad.

In July, a Facebook press conference that was designed to showcase the company's progress ended up demonstrating the quandary that all the major platforms face. A CNN reporter asked how Facebook could continue to allow Infowars—a conspiracy theory site that has propagated the idea that school shootings are hoaxes and their victims "crisis actors"—to operate a page with over 900,000 followers. Company spokespeople struggled to explain in which cases false information is taken down for violations of its "community standards" and in which cases it is merely "downranked" in the Facebook news feed.

Once again, public policy hasn't kept up. There is no federal agency charged with protecting U.S. democracy in the digital age, and so the only cops on the beat are the Federal Trade Commission and the Federal Election Commission. The FTC is charged with the wide-ranging task of consumer protection and lacks sufficient staff and authority to address most of the challenges specific to the weaponization of the Internet. The Obama administration proposed an update to privacy laws that would have given the FTC more power when it comes to that issue, but Congress never took it up. And although a draft of the 2010 Dodd-Frank Wall Street Reform and Consumer Protection Act contained a provision to give the FTC rule-making authority, the provision was stripped out before the bill passed. The FEC, for its part, is perpetually stalemated along partisan lines, just as it was in 2014, when a vote regarding whether to require transparency in online political advertising ended in a deadlock. For the most part, the government has left it to individuals and digital platforms to design their own defenses, and they are falling short.

Intervention for Openness

Even though public policy played a large role in enabling the creation and growth of the Internet, a mythical, libertarian origin story arose, which fed the belief that the Internet is so open that regulation is unnecessary—indeed, that government is like Kryptonite to the Internet. Of course, this was also a convenient narrative for opponents of regulation, who fought updating offline rules to fit the online world for economic or ideological reasons. But it is critical that Washington act now to prevent the further weaponization of the Internet against democracies and individuals attempting to exercise their human rights—and to do so without sacrificing democratic values such as freedom of expression. The history of the Internet's founding offers the right model: intervention on behalf of openness.

To help tilt the balance against autocrats, the U.S. government should fully fund and staff the Global Engagement Center so that it can coordinate support for activists abroad and counter disinformation and extremist content. Washington should also continue to support the efforts that the Broadcasting Board of Governors, the federal agency that oversees Voice of America and other broadcasters, is making on this front, including developing tools that help dissidents get online and backing the fact-checking website Polygraph.info.

There are also steps that can be taken to reduce the opportunity for so-called dark money and dark data to undermine democracy. Congress should pass the Honest Ads Act, a bill proposed in October 2017 that would apply television's rules on disclosing the funding behind political advertising to the Internet. Platforms should be required to insist that entities buying political ads provide information on their donors, as well—and to verify the identity of those donors and disclose that information publicly in a sortable, searchable database. In order to deal a blow to microtargeted disinformation, Congress should borrow from Europe's General Data Protection Regulation: organizations should be required to treat political and philosophical data about users as sensitive information—so that it cannot be collected and then used to target political advertising without express permission. Users should also have more data rights, such as the ability to take their data to another platform or use it interoperably.

Digital platforms should find a way to offer users more context for the news their algorithms present. They might do so through some method of differentiating those news outlets that follow accepted journalistic practices (customs such as having a masthead, separating news from opinion, and issuing corrections) from those that do not. The platforms should be required to take down fake accounts and remove bots unless they are clearly labeled as such. The largest social media companies—Facebook, Twitter, and YouTube—need to be transparent about their content-moderation rules. Regulation might even require certain platforms to provide due process protections for users whose content is taken down. And a narrow change to Section 230 could eliminate immunity for platforms that leave up content that threatens or intentionally incites physical violence.

Of course, change must come from the top. Trump himself repeatedly refuses to acknowledge Russia's interference in the 2016 election, despite the clear findings of the intelligence community. And in May, the Trump administration's National Security Council eliminated the position of cybersecurity coordinator and handed the portfolio to a deputy with many other responsibilities. That decision should be reversed, and foreign information operations should be treated as seriously as cyberattacks are. And at the international level, Washington should promote its approach through multilateral organizations and provide technical assistance through the World Bank.

What's needed is U.S. leadership. The Internet would never have become such a transformational technology were it not for openness—a quality that was inherent in its design yet nurtured by government policies. But over time, those policies did not keep up with changes in technology or the way it was used. The victims of this lag have been those who initially benefited the most from the Internet: democracies, champions of freedom, and ordinary citizens.

It is time for them to take back the Internet. The United States is uniquely positioned to assume the lead on this task. As the promoter of the key early policies and the home to many of the largest Internet companies, only it can drive the development of a framework that ensures the openness and transparency necessary for democratic debate without harming innovation. But if the United States shirks its responsibility, it will further empower the adversaries of democracy: revisionist states, authoritarian governments, and fraudsters bent on exploiting the Internet for their own, dangerous ends.

https://www.foreignaffairs.com/articles/world/2018-08-13/internets-lost-promise

Karen Kornbluh is Senior Fellow for Digital Policy at the Council on Foreign Relations and a member of the Broadcasting Board of Governors.

The Brookings Institute

Artificial intelligence, deepfakes, and the uncertain future of truth

by John Villasenor
February 14, 2019

Deepfakes are videos that have been constructed to make a person appear to say or do something that they never said or did. With artificial intelligence-based methods for creating deepfakes becoming increasingly sophisticated and accessible, deepfakes are raising a set of challenging policy, technology, and legal issues.

Deepfakes can be used in ways that are highly disturbing. Candidates in a political campaign can be targeted by manipulated videos in which they appear to say things that could harm their chances for election. Deepfakes are also being used to place people in pornographic videos that they in fact had no part in filming.

Because they are so realistic, deepfakes can scramble our understanding of truth in multiple ways. By exploiting our inclination to trust the reliability of evidence that we see with our own eyes, they can turn fiction into apparent fact. And, as we become more attuned to the existence of deepfakes, there is also a subsequent, corollary effect: they undermine our trust in *all* videos, including those that are genuine. Truth itself becomes elusive, because we can no longer be sure of what is real and what is not.

What can be done? There's no perfect solution, but there are at least three avenues that can be used to address deepfakes: technology, legal remedies, and improved public awareness.

Deepfake Detection Technology

While AI can be used to make deepfakes, it can also be used to detect them. Creating a deepfake involves manipulation of video data—a process that leaves telltale signs that might not be discernable to a human viewer but that sufficiently sophisticated detection algorithms can aim to identify.

As research led by professor Siwei Lyu of the University at Albany has shown, face-swapping (editing one person's face onto another person's head) creates resolution inconsistencies in the composite image that can be identified using deep learning techniques. Professor Edward Delp and his colleagues at Purdue University are using neural networks to detect the inconsistencies across the multiple frames in a video sequence that often result from face-swapping. A team including researchers from UC Riverside and UC Santa Barbara has developed methods to detect "digital manipulations such as scaling, rotation, or splicing" that are commonly employed in deepfakes.

The number of researchers focusing on deepfake detection has been growing, thanks in significant part to DARPA's Media Forensics program, which is supporting the development of

"technologies for the automated assessment of the integrity of an image or video." However, regardless of how far technological approaches for combating deepfakes advance, challenges will remain.

Deepfake detection techniques will never be perfect. As a result, in the deepfakes arms race, even the best detection methods will often lag behind the most advanced creation methods. Another challenge is that technological solutions will have no impact when they aren't used. Given the distributed nature of the contemporary ecosystem for sharing content on the Internet, some deepfakes will inevitably reach their intended audience without going through detection software.

More fundamentally, will people be more likely to believe a deepfake or a detection algorithm that flags the video as fabricated? And what should people believe when different detection algorithms—or different people—render conflicting verdicts regarding whether a video is genuine?

Legal and Legislative Remedies

The legal landscape related to deepfakes is complex. Frameworks that can potentially be asserted to combat deepfakes include copyright, the right of publicity, section 43(a) of the Lanham Act, and the torts of defamation, false light, and intentional infliction of emotional distress. On the other side of the ledger are the protections conferred by the First Amendment and the "fair use" doctrine in copyright law, as well as (for social networking services and other web sites that host third-party content) section 230 of the Communications Decency Act (CDA).

It won't be easy for courts to find the right balance. Rulings that confer overly broad protection to people targeted by deepfakes risk running afoul of the First Amendment and being struck down on appeal. Rulings that are insufficiently protective of deepfake targets could leave people without a mechanism to combat deepfakes that could be extraordinary harmful. And attempts to weaken section 230 of the CDA in the name of addressing the threat posed by deepfakes would create a whole cascade of unintended and damaging consequences to the online ecosystem.

While it remains to be seen how these tensions will play out in the courts, two things are clear today: First, there is already a substantive set of legal remedies that can be used against deepfakes, and second, it's far too early to conclude that they will be insufficient.

Despite this, federal and state legislators, who are under pressure to "do something" about deepfakes, are responding with new legislative proposals. But it is very hard to draft deepfake-specific legislation that isn't problematic with respect to the First Amendment or redundant in light of existing laws.

For example, a (now expired) Senate bill S.3805 introduced in December 2018 would have, among other things, made it unlawful "using any means or facility of interstate or foreign commerce," to "create, with the intent to distribute, a deep fake with the intent that the distribution of the deep fake would facilitate criminal or tortious conduct under Federal, State,

local, or Tribal law." Writing at the Volokh Conspiracy regarding S.3805, USC law professor Orin Kerr observed that:

> It's already a crime to commit a crime under federal, state, local, or tribal law. It's also already a crime to 'facilitate' a crime—see 18 U.S.C. § 2 at the federal level, and state laws have their equivalents. Plus, it's already a tort to commit a tort under federal, state, local, or tribal law. This new proposed law then makes it a federal crime to either make or distribute a deepfake when the person has the intent to do the thing that is already prohibited. In effect, it mostly adds a federal criminal law hammer to conduct that is already prohibited and that could already lead to either criminal punishment or a civil suit.

State legislators in New York have considered a bill that would prohibit certain uses of a "digital replica" of a person and provide that "for the purposes of the right of publicity, a living or deceased individual's persona is personal property." Unsurprisingly, this raised concerns in the entertainment industry. As a letter from the Walt Disney Company's Vice President of Government Relations stated, "if adopted, this legislation would interfere with the right and ability of companies like ours to tell stories about real people and events. The public has an interest in those stories, and the First Amendment protects those who tell them."

Raising Public Awareness

At the end of the day, technological deepfake detection solutions, no matter how good they get, won't prevent all deepfakes from getting distributed. And legal remedies, no matter how effective they might be, are generally applied after the fact. This means they will have limited utility in addressing the potential damage that deepfakes can do, particularly given the short timescales that characterize the creation, distribution, and consumption of digital media.

As a result, improved public awareness needs to be an additional aspect of the strategy for combating deepfakes. When we see videos showing incongruous behavior, it will be important not to immediately assume that the actions depicted are real. When a high-profile suspected deepfake video is published, it will usually be possible to know within days or even hours whether there is reliable evidence that it has been fabricated. That knowledge won't stop deepfakes, but it can certainly help blunt their impact.

https://www.brookings.edu/blog/techtank/2019/02/14/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth/

John Villasenor is a nonresident senior fellow in Governance Studies and the Center for Technology Innovation at Brookings. John is also a professor of electrical engineering, public policy, and management. He addresses the intersection of technology, policy, and law.

Statement of

NINA JANKOWICZ

Woodrow Wilson International Center for Scholars, Kennan Institute

BEFORE THE UNITED STATES SENATE

COMMITTEE ON THE JUDICIARY

Concerning

"Election Interference: Ensuring Law Enforcement Is Equipped to Target Those Seeking to Do Harm"

June 12, 2018

Introduction: *Beyond Knee-Jerk Reactions*

Chairman Grassley, Ranking Member Feinstein, and distinguished Members of the Committee: thank you for having me here today. My name is Nina Jankowicz, and I am a Global Fellow within the Kennan Institute at the Woodrow Wilson International Center for Scholars, where my research focuses on Russian disinformation and influence in Eastern Europe and beyond. It is an honor to testify before you this morning on the topic of election interference in the United States and the policy solutions necessary to protect our democratic processes. It is especially heartening to see continued bipartisan interest in this topic, as it is truly one that knows no political party.

Throughout my career, I have worked on the front lines of Russia's information war. I became familiar with Russian disinformation techniques while working on Russia programming at the National Democratic Institute, a frequent target of Russian lies. As a Fulbright Public Policy Fellow in Ukraine, I advised the Ukrainian Foreign Ministry on strategic communications issues and observed the implementation of policies meant to protect Ukraine's information environment. And over the past year, I have spoken with officials countering Russian influence and disinformation across Central and Eastern Europe as I work on a book on the development of modern Russian information warfare tactics and government responses to these critical challenges to the democratic process.

These firsthand experiences and observations have led me to a conclusion that may surprise you: even if the United States Government were to acknowledge the threat posed by Russian influence campaigns today in no uncertain terms, and we were to walk out of the hearing room and secure beyond a shadow of a doubt the country's election infrastructure; even if we hermetically sealed our information environment from inauthentic users and false or misleading information, and if social media companies finally put forth a good faith effort to put users and the security of our democracy first; even then, we would *still* not successfully dispel the threat our democracy faces from malign actors' political influence operations.

If our democratic processes are to remain secure, we must think beyond knee-jerk reactions and punitive measures. The Congress and the U.S. government must put citizens at the heart of our response to disinformation and address the issues that make our society so susceptible to outside influence in the first place.

Moscow's Main Weapon: Ourselves

Over the past few months, as we've learned more about the specifics of Russia's interference in the 2016 US election, some have questioned the Russian operation's "effectiveness" or whether it is "sophisticated" enough for us to care about. Many cite the fact that the potent advertising tools used by the Internet Research Agency are indeed available to all Facebook users.

This is a line of inquiry that privileges the American or Western experience—as if we in the West are the only countries to have experienced these phenomena—and dismisses the very actual fears and societal divisions that cause some of our fellow citizens to buy into Russia's tactics. It also misses a key point: **the United States is at risk for further election interference**

today not *only* because of the social media tools that malign actors exploit, but because our society is more fractured than ever.

The 2018 Edelman Trust Barometer measured a 37% decline in trust in US institutions—government, the media, business, and NGOs—over a single year,[1] while the Pew Research Center found in December 2017 that only 18% of the population trusts the government in Washington some or most of the time.[2] This trust-deficient environment means that American citizens are looking elsewhere for information.[3] As we saw with the Internet Research Agency's social media campaigns surrounding the 2016 election, Americans of all political stripes were receptive to content of dubious origins and messages.[4] **In short, societal fractures like ours are far more valuable to malign actors than any social media targeting tool. Only solutions with citizens at their heart can truly address these fractures and ensure our society is not left vulnerable to future interference.**

European countries that have been most successful in countering malign influence in their information and electoral environments have in common one key point: their governments recognize they cannot simply fact-check or label their way out of the crises of truth that they face.

Estonia: Outreach to the Disaffected

In Estonia in 2007, the Kremlin exploited tensions between the ethnic Russian population that had remained in Estonia after the country's independence and the native Estonian population. The dominance of Kremlin-controlled Russian language media outlets in Estonia meant that the Russian population was subjected to a constant barrage of antagonizing information, claiming in its historical revisionist narrative that Estonia owed its existence to Soviet troops who "liberated" the capital, Tallinn.[5] In reality, of course, Estonia had suffered under Soviet occupation, but this mattered little to the ethnic Russians who gathered to celebrate Victory Day and other Soviet legacy holidays at the Bronze Soldier, a monument to World War II dead and tomb of the unknown soldier in central Tallinn. Crowds grew, as did tensions between Estonian nationalists and Soviet revisionists who faced off at the monument, only narrowly avoiding physical altercations.

The Estonian government eventually decided to move the statue and associated human remains from the center of Tallinn to a military cemetery on the outskirts of the city. This decision became the latest in a long line of so-called grievances the Russian population were told they had against the Estonian government. Encouraged by the Russian media, riots broke out, destroying

---

[1] Sara Fischer, "Red alert: America suffers record drop in trust; China rises," *Axios*, 22 January 2018.
[2] Pew Research Center, Public Trust in Government 1958-2017.
[3] For more on the trust gap's role in disinformation, see: Nina Jankowicz, "Our Biggest Mistake in Fighting Fake News," *The Washington Post*, 31 March 2017.
[4] For more on the Internet Research Agency's use of social media, see: Nina Jankowicz, "The Top Three Trends We Miss When Discussing Russian Ads," Alliance for Securing Democracy, German Marshall Fund, 15 May 2018.
[5] For more on the Bronze Soldier Crisis, see Kadri Liik, "The Bronze Year of Estonian-Russian Relations," International Centre for Defence Studies, 2007.

much of the center of Tallinn and killing one. Simultaneously, Estonia was hit with a wave of cyber attacks, briefly crippling the country's banking system, government services, and Internet access.

This was Moscow's first attempt to test the disinformation and influence operation tactics we are familiar with today, and social media has only strengthened the Kremlin's ability to more insidiously target and message to receptive audiences organized along societal fissures. Eleven years later, despite the ubiquity of social media, the Kremlin's messaging in Estonia is finding fewer footholds. This is partly a natural process; Russians in Estonia enjoy the economic and social benefits of residence in an EU member state, and are keenly aware of the social, political, and economic realities of life in Russia, due to frequent travel.[6] But addition to beefing up their cyber defenses and expertise, the Estonian government has made a concerted effort to conduct outreach to the ethnic Russian population in Estonia. One of Estonia's leading universities set up a Russian-language outpost in Narva, a border city that is 95% ethnically Russian. The Estonian Ministry of Culture views Russian language programming as a strategic priority,[7] and the government has established a Russian-language TV station to compete with Russian signals. Furthermore, in terms of combatting major instances of cyber attacks and disinformation, the Estonian government believes in early governmental attribution, undermining malign actors through proactive communication.

No Estonian will tell you things are perfect, but they are much better than a decade ago. Most importantly, there is recognition among Estonian government officials and the population *writ large* that these efforts will not yield results overnight, but are a generational investment that will pay dividends in the future.

Ukraine: Beyond Bans

It's not hard to imagine what further damage the Bronze Soldier might have inflicted if social media had been more ubiquitous at the time, as this is a strategy that Russia has expanded upon and pursued in Ukraine since 2014. After Ukrainians overthrew a corrupt Kremlin-aligned government, Moscow illegally annexed the Crimean peninsula and invaded Ukraine's Donbas region. It also launched a parallel assault on Ukraine's information space, flooding social media with fake news claiming the new Ukrainian government was fascist and its election unconstitutional, among many other narratives meant to discredit the post-Maidan authorities.

Civil society groups in Ukraine launched several initiatives to separate fact from fiction, including StopFake, a fact-checking program that was one of the first defensive battalions in the modern information war. The government also undertook a series of initiatives meant to restrict access to Russian media sources, including blocking the Russian social networking sites vKontakte and Odnoklassniki in May 2017. These steps are well-intentioned and make important statements about the information environment and Ukraine's commitment to securing it, but are unlikely to change behavior in the long run. Since the 1970s, psychological research has shown

[6] See Andrew Higgins, "Two Border Cities Share Russian History—And a Sharp European Divide," *The New York Times*, 9 November 2017.

[7] See "Estonia gets creative about integrating local Russian-speakers," *The Economist*, 10 May 2018.

that repeated untruths are difficult to debunk.[8] According to a study by Columbia University's Andrew Guess, this is even more difficult on social networks such as Twitter, where "false information . . . overpowers efforts to correct it by a ratio of about three to one."[9] But simply banning access to the websites where false information proliferates is also not a cure-all; while both vKontakte and Odnoklassniki became distinctly less popular in Ukraine, they both remain among the top fifteen most accessed websites. The ban itself has inspired a great deal of criticism from media freedom advocates and fed Russian disinformation that Ukraine is treating Russian speakers unfairly.[10]

Beyond fact-checking and bans, there is a growing demand for media literacy training in Ukraine, where only 23% of the population engage in basic source cross-checking.[11] IREX, an American non-governmental organization, trained 15,000 people in critical thinking, source evaluation and emotional manipulation. As a result, IREX measured a 29% increase in participants who double-check the news they consume. Eighteen months after the end of the program, participants were 13% more likely to correctly identify and critically analyze a fake news story, 25% more likely to self-report checking multiple news sources, and 28% more likely to demonstrate sophisticated knowledge of the news media industry as compared with a control group that had not been trained.[12] Last summer, the Ukrainian Ministry of Education signed a decree prioritizing media literacy in the national curriculum. While Ukraine's battle with Russian information is far from finished, these investments in the country's future will pay dividends in years to come.

Keeping Citizens at the Heart of the American Response to Malign Influence

Citizen-based responses to election interference are not a panacea. They must work in concert with structural and punitive measures, such as securing our election infrastructure and sanctions, designed to protect our institutions. To date, however, the nascent American response has focused on reactive and short-term initiatives rather than those that are proactive and generational. In addition to the stipulations provided in the Honest Ads and Secure Elections Acts, which I support, Congress must pursue and encourage citizens-based solutions in its further interactions and work related to election protection. Below are several ideas for further Congressional exploration with these principles at heart.

Social Media Regulation

Social media companies have so far played "Whack-a-Troll" in responding to Russian disinformation and election interference: researchers uncover posts linked to Russia and social

---

[8] Lynn Hasher, David Goldstein, and Thomas Toppino, "Frequency and the Conference of Referential Validity," *Journal of Verbal Learning and Verbal Behavior,* 16, 107-112, 1977.

[9] Andrew Guess, "Fact-checking on Twitter: An examination of campaign 2014," American Press Institute, 29 April 2015.

[10] Игорь Бурдыга, "Год без "В контакте" и "Одноклассников" в Украине: действуют ли санкции?" Deutsche Welle, 16 May 2018.

[11] Erin Murrock, Joy Amulya, Mehri Druckman and Tetiana Liubyva, "Winning the war on state-sponsored propaganda," IREX, 2018.

[12] *Ibid.*

media firms apologize and remove the content.[13] Unfortunately for both social media users and our democracy, it is extraordinarily easy to create and deploy fake accounts. Furthermore, there is ample evidence of information sharing and even Russian government funding of "alternative" media and fringe organizations abroad. Comparatively, trolls and bots are the least of our worries; how do we account for and stem the amplification of content that looks authentic, but has links to a malign source? These are complex challenges, but educating and empowering social media users will ameliorate them.

- To begin, social media platforms should be required to obtain **informed and meaningful consent from users to terms of service**. Most users have no idea what they are buying into when they sign up to share pictures of their dogs, chat with their friends, or follow the news. This ignorance, as well as emotion, is what Russia exploits through its online influence campaigns. All too often, users are incentivized to blindly click through terms of service that allow their data to be shared with advertisers, be they malign foreign actors or commercial entities. Users should understand the level of microtargeting to which they are being subjected, and understand its costs.
- To that end, **terms of service should be written in plain English and clearly define what content is permissible on platforms**. While platforms have been quick to use Section 230 of the Communications Decency Act to absolve themselves of responsibility for content posted on their platforms, if they are committed to supporting the democratic process, they should consider updating their terms of service to reflect whether disinformation is permissible. Those definitions should be actively enforced, whether they apply to hate speech or disinformation. It would be costly and almost certainly require human content reviewers and the establishment of a complaints and appeals process, but civil discourse and democracy are priceless.
- The steps social media companies have taken to increase advertising transparency are steps in the right direction, but blanket bans and restrictions on political ads are already being clumsily enforced.[14] One potential solution is for platforms or a third party to **establish a online advertising code of conduct that could inform a register of trusted advertisers**, akin to a Better Business Bureau.
- Social media companies have near ubiquitous access to Americans' lives; they should **embrace their role as educators**. Facebook recently announced plans to include media literacy modules at the top of users' news feeds and has taken out full page ads in national newspapers, while Twitter has participated in small-scale media literacy programs. Both platforms should focus on practices that encourage behavior change, rather than simply raising awareness.

---

[13] For more on social media's response to foreign and homegrown disinformation, see Nina Jankowicz, "Russian Trolls are Only Part of the Problem," *The New York Times*, 25 January 2018.

[14] Sean Guillory, a Russian history scholar at the University of Pittsburgh with a popular educational podcast about Russia and US foreign policy, was recently denied ads for his podcast because of their "political nature." See Sean Blumenthal, "Facebook's New Ad Disclosures Are Meant to Fight Russian Trolls. A Russian History Podcaster is Paying the Price," *Huffington Post*, 8 June 2018.

Investing in Skills to Support the Democratic Process

The investments that will best protect American democracy for generations to come are decidedly low-tech. They focus not only on empowering Americans to be more savvy consumers of information on and offline, but increasing investments in our collective understanding of civics, as well as in building and repairing critical thought and civil discourse.

- Citizens-based solutions to fighting election interference should be **wider than simply teaching social media users how to recognize online fakes and fact-check**. They should include investments in civics; citizens who better understand how government works are less likely to buy into the falsehoods and conspiracies harmful to democracy. Furthermore, they should be tied to broad-based efforts to increase critical thinking skills and preserve civil discourse. This would assist people in sampling a range of viewpoints to inform their daily lives and the criticism that is healthy for any democracy, while developing greater immunity to conspiratorial versions of the truth. Finally, to avoid politicizing these efforts, they should not be couched in the language of influence operations or a direct response to Russian tactics, but simply an investment in America's future.
- As the United States continues to mount its response to election interference and online influence campaigns, Congress should **encourage cooperation and coordination across government**, particularly between the national security community and departments of education at both the national and state levels. This will promote citizens-based solutions within policy communities that are sometimes detached from the daily concerns of their fellow Americans.
- Finally, these solutions **need not be limited to the halls of schools and universities; adults should also be a target audience** for these skills-building programs. For instance, the United States could launch training programs on digital media literacy and foreign influence for government employees, as countries such as the Czech Republic have done.

Though the issue of malign political influence seems novel and insurmountable, it is one with which our country has always struggled. Even Thomas Jefferson had similar worries, but he, too, recognized the value of investing in American citizens as a holistic response to building a more secure democracy, writing in 1820: "I know of no safe depository of the ultimate powers of the society but the people themselves; and if we think them not enlightened enough to exercise their control with a wholesome discretion, the remedy is not to take it from them, but to inform their discretion by education."

Moscow will continue to attempt to influence our democracy, as it has for decades, and now that the Kremlin has written the playbook for how to do so, other bad actors will undoubtedly imitate Russian tactics. To prepare for these future attacks on democracy—and indeed, even attacks from within—we must think beyond Russia to the key actors in the democratic process: the American people.

European Parliament

Online disinformation and the EU's response

The visibility of disinformation as a tool to undermine democracies increased in the context of Russia's hybrid war against Ukraine. It gained notoriety as a global challenge during the UK referendum on EU membership as well as the United States presidential election campaign in 2016. The European Union and the European Parliament are stepping up efforts to tackle online disinformation ahead of the May 2019 European elections.

A global phenomenon with growing visibility

The phenomenon of false, misleading news stories is at least as old as the printing press. However, social media and their personalisation tools have accelerated the spread of rumours, hoaxes and conspiracy theories. The phenomenon gained global visibility during the 2016 US presidential election, when viral false news or 'junk news' across the political spectrum received more engagement on Facebook (FB) than real news. Research has shown that Russian accounts posted over 45 000 Brexit messages in the last 48 hours of the campaign. According to the Collins Dictionary, which chose 'fake news' as its word of the year for 2017, the term has seen an unprecedented increase in usage, of 365 % since 2016.

Online disinformation as an instrument of malign influence

When designed to deceive users for political purposes, digital gossip falls under 'disinformation' – the dissemination of verifiably false or misleading information which non-state and state actors can use to intentionally deceive the public and cause public harm. The Kremlin continues its disinformation campaigns in its ongoing hybrid war against Ukraine, and is applying them in its 'holistic' information warfare against the West. Pro-Kremlin information campaigns boost Moscow's narrative of a morally decayed EU on the brink of collapse, and seek to exploit divisions in Western societies. In November 2017, British Prime Minister Theresa May accused Russia of 'weaponising information', and a February 2018 report by UK communications agency 89up.org found Russian pro-Brexit social media interference worth up to €4.6 million during the campaign. In August 2017, the US imposed fresh sanctions on Russia over its interference in the 2016 election. Following the nerve-gas attack on a former Russian spy, Sergei Skripal, and his daughter on UK soil in March 2018, the US imposed new sanctions, including on 16 Russian entities and individuals linked to the Internet Research Agency (a Russian 'troll factory' spreading disruptive content via social media) indicted by Special Counsel Robert Mueller for their role in election-meddling operations. The European Commission and the Vice-President of the Commission / High Representative of the Union for Foreign Affairs and Security Policy (HR) responded with a June 2018 joint communication on boosting resilience against hybrid threats, emphasising strategic communications as a priority.

Online platforms and their role in spreading disinformation

Whereas US tech giants had previously played down the volume of content purchased by Russian actors during the 2016 US presidential election campaign, FB, Google and Twitter told US lawmakers in November 2017 that pro-Kremlin actors bought and published divisive ads aimed at influencing both liberals and conservatives. FB said Russia-backed posts reached up to 126 million Americans during and after the 2016 election. The March 2018 disclosure that user data from 87 million FB users – including that of 2.7 million EU citizens – had been improperly shared with the controversial political consultancy company Cambridge Analytica (which used the data to micro-target and mobilise voters in the US and the UK) further increased the focus on the role of online platforms, not only in spreading, but also in monetising disinformation. In April 2018, FB CEO Mark Zuckerberg told the US Congress that tens of thousands of fake accounts were deleted to prevent election interference in 2017. He explained that Russian accounts primarily used ads to influence views on issues rather than promoting specific candidates or political messaging. In May 2018, Zuckerberg dodged questions about data protection, fake news and election security, posed by Members of the European Parliament (MEPs) in Brussels. Confidential emails from Zuckerberg, published in December 2018 – suggesting that FB secretly gave some companies access to users' friends' data – cast further doubt about FB's ethics.

This is a further updated edition of an 'at a glance' note published in May 2018.

Author: Naja Bentzen, Members' Research Service

PE 620.230 – February 2019

EN

http://www.europarl.europa.eu/RegData/etudes/ATAG/2018/620230/EPRS_ATA(2018)620230_EN.pdf

EU steps up anti-disinformation efforts to protect democracy

The FB data breach disclosure reignited the ongoing debate on the role of online platforms in the spread of conspiracy theories, disinformation and false news. In its June 2017 resolution on online platforms and the digital single market, the European Parliament had already called on the Commission to analyse the legal framework with regard to 'fake news', and to look into the possibility of legislative intervention to limit the dissemination of fake content. President Jean-Claude Juncker tasked Mariya Gabriel, Commissioner for the Digital Economy and Society, to look into the democratic challenges that online platforms create as regards the spread of fake information, as well as to reflect on possible action at EU level. In October 2017, the Commission launched a public consultation on fake news and online disinformation. It also set up a high-level expert group (HLEG) representing academia, online platforms, news media and civil society. The Commission's April 2018 communication on 'Tackling online disinformation: a European approach' took recommendations of the HLEG into account and proposed an EU-wide Code of Practice – signed by the online platforms – to ensure transparency by explaining how algorithms select news, as well as improving the visibility and accessibility of reliable news. The communication also recommended support for an independent network of fact-checkers as well as actions to boost quality journalism and media literacy.

Coordinating the response to disinformation ahead of the European elections

Responding to the June 2018 call by the European Council to protect the EU's democratic systems and 'combat disinformation, including in the context of the upcoming European elections', the Commission and the HR in December 2018 presented an 'action plan against disinformation' with specific proposals for a coordinated European response. The action plan builds on existing Commission initiatives as well as the work of the East StratCom Task Force, set up in 2015 under the European External Action Service (EEAS, see below). The action plan focuses on four main areas:

Improved detection. Strategic Communication Task Forces and the EU Hybrid Fusion Cell in the EEAS, as well as the EU delegations in the Neighbourhood countries will receive additional specialised staff and data analysis tools. The EEAS's budget for strategic communication to address and raise awareness about disinformation is planned to more than double, from €1.9 million in 2018 to €5 million in 2019.

Coordinated response. A dedicated Rapid Alert System will be set up among the EU institutions and Member States to facilitate data sharing and to provide alerts on disinformation threats in real time.

Online platforms and industry. The signatories of the EU-wide Code of Practice on Disinformation (signed on 26 September 2018) are urged to swiftly and effectively implement the commitments, focusing on actions that are urgent for the European elections. This includes deleting fake accounts, labelling messaging activities by 'bots' and cooperating with fact-checkers and researchers to detect disinformation and make fact-checked content more visible.

Raising awareness and empowering citizens. In addition to targeted awareness campaigns, the EU institutions and Member States will promote media literacy as well as support national teams of independent fact-checkers and researchers to detect and expose disinformation on social networks.

The EU's 'myth-busters' and the European Parliament

In 2015, the European Council asked the HR to prepare an action plan on strategic communication to address Russia's ongoing disinformation campaigns. As a first step, the East StratCom Task Force was set up in September 2015 under the EEAS. Since then, the team has collected more than 4 000 disinformation stories, which it has analysed, debunked and published on euvsdisinfo.eu as well as on its Twitter account, @EUvsDisinfo. The team also communicates EU policies in the Neighbourhood. Two other teams are focusing on the EU's Southern Neighbourhood and the Western Balkans. The European Parliament (EP), in its 23 November 2016 resolution on EU strategic communication to counteract propaganda, called for the East StratCom Task Force to be reinforced. In January 2018, the task force received its first budget of €1.1 million, initiated by Parliament.

On 22 January 2019, the EP Committee on Foreign Affairs (AFET) adopted a draft recommendation to the Council and the Vice-President of the Commission/HR (rapporteur: Anna E. Fotyga, EPP, Poland) calling for strategic communication to become a matter of high priority in the EU. Highlighting the Cambridge Analytica breach, it calls for legislation to safeguard future election campaigns from undue influence. It also invites Member States which have not already done so to second national experts to the teams. Parliament is expected to vote during its March I 2019 plenary part-session.

European Commission

Code of Conduct on countering illegal hate speech online

Fourth evaluation confirms self-regulation works

Factsheet | **February 2019**

Věra Jourová
Commissioner for Justice, Consumers and Gender Equality

Directorate-General for Justice and Consumers

The fourth evaluation on the *Code of Conduct on Countering Illegal Hate Speech Online* confirms continuous progress on the swift removal of illegal hate speech. While the fight against hate speech needs to continue and be further strengthened, the Code is delivering on its key commitments. It proves to be an effective tool to face this challenge.

Today, all IT companies fully meet the target of reviewing the majority of the notifications within **24 hours**, reaching an average of **89%**. These results also include Instagram and Google+ which joined in 2018. This is a significant increase from when the Code was launched back in 2016 (40% within 24 hours).

On average, IT companies **are removing 72 % of the illegal hate speech notified** to them. This is estimated to be satisfactory removal rates, as some of the content flagged by users could relate to content that is not illegal. In order to protect freedom of speech only content deemed illegal should be removed.

Key figures

1. Notifications of illegal hate speech

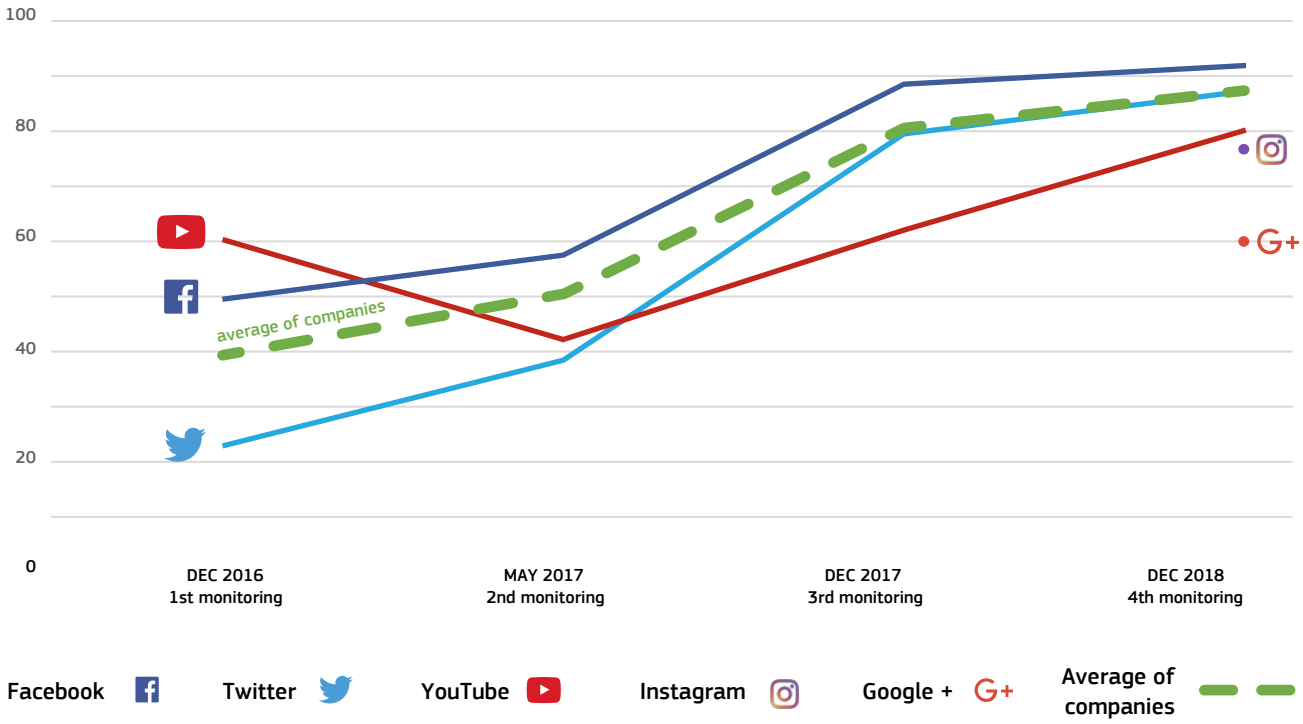
- 39 organisations from 26 Member States (all except Luxembourg and Denmark) sent notifications relating to hate speech deemed illegal to the IT companies during a period of 6 weeks (5 November to 14 December 2018). In order to establish trends, this exercise used the same methodology as the previous monitoring rounds (see Annex).
- A total of 4392 notifications were submitted to the IT companies taking part in the Code of Conduct. This represents a steady increase compared to the previous exercises.
- 2748 notifications were submitted through the reporting channels available to general users, while 1644 were submitted through specific channels available only to trusted flaggers/reporters.

- Facebook received the largest amount of notifications (1882), followed by Twitter (1314) and YouTube (889). This breakdown is similar to previous exercises. Instagram (279) and Google+ (28), which have joined the Code of conduct in early 2018, were tested too. Microsoft did not receive any notification.
- In addition to flagging the content to IT companies, the organisations taking part in the monitoring exercise submitted 503 cases of hate speech to the police, public prosecutor's bodies or other national authorities.

2. Time of assessment of notifications

- In **88.9% of the cases** the IT companies assessed the notifications **in less than 24 hours**, an additional 6.5% in less than 48 hours, 3.9% in less than a week and in 0.7% of cases it took more than a week.
- Facebook assessed the notifications in less than 24 hours in 92.6% of the cases and 5.1% in less than 48 hours. The corresponding figures for YouTube are 83.8% and 7.9% and for Twitter 88.3% and 7.3%, respectively. Instagram's performance is positive, 77.4% of notifications were assessed in less than 24 hours, while Google+ did so in 60% of the cases[1].
- The target of reviewing the notifications within one day is fully met by all the IT companies and there has been additional progress compared to the previous monitoring exercise (81.7%).

Rate of notifications reviewed within 24 hours since the launch of the Code of Conduct

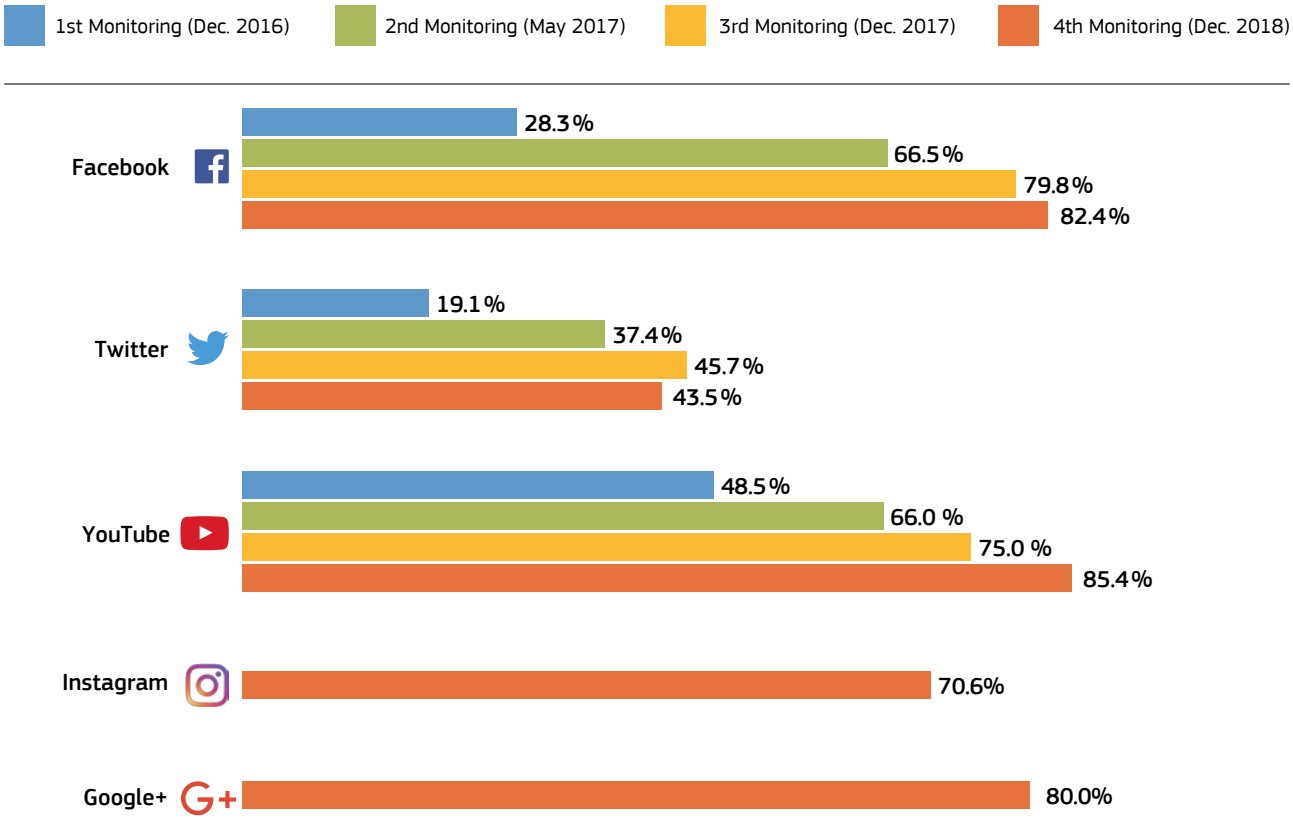


[1] The figures for Google+ are based on a significantly lower number of cases compared to the other IT companies.

3. Removal rates

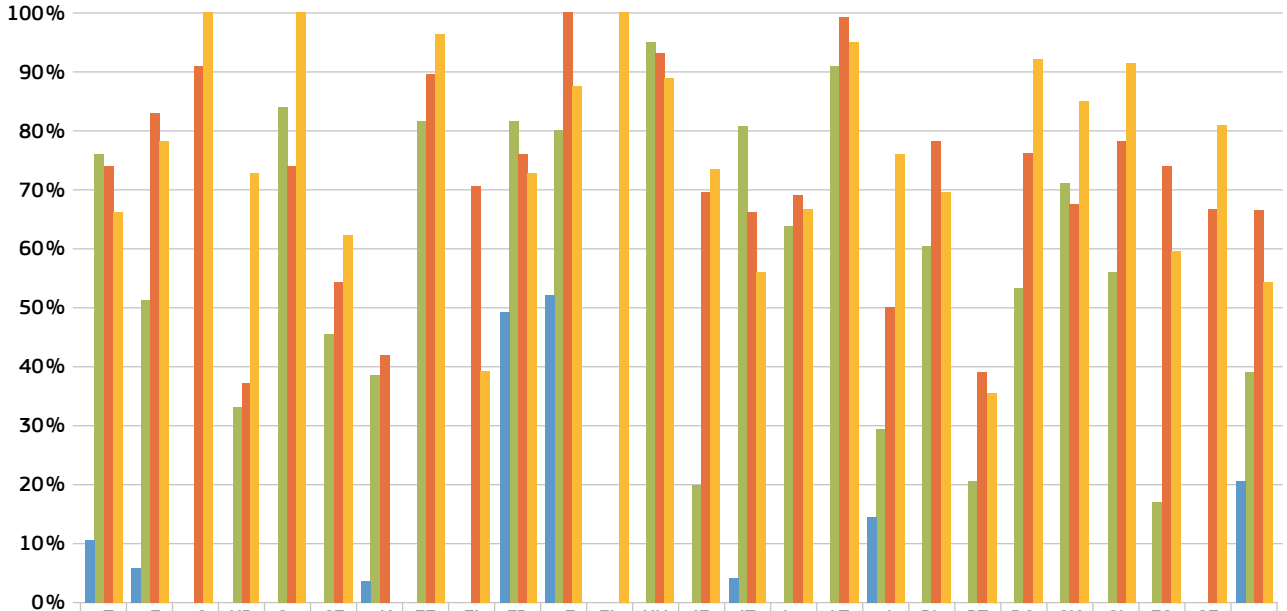
- Overall, **IT companies removed 71.7% of the content** notified to them, while 28.3% remained online. This represents a small increase compared to the 70% one year ago.
- YouTube removed 85.4% of the content[2], Facebook 82.4% and Twitter 43.5%. Both Facebook and, especially, YouTube made further progress on removals when compared to last year. Twitter, while remaining in the same range as in the last monitoring, has slightly decreased its performance. Google+ removed 80.0% of the content and Instagram 70.6%.
- Removal rates varied depending on the severity of hateful content. On average, 85.5% of content calling for murder or violence against specific groups was removed, while content using defamatory words or pictures to name certain groups was removed in 58.5 % of the cases. This suggest that the reviewers assess the content scrupulously and with full regard to protected speech.
- The divergence in removal rates of content reported using trusted reported channels as compared to channels available to all user was only 4.8%. This difference was more than twice as high in December 2017 (10.5%).

Removals per IT Company



[2] YouTube has also limited the features of an additional 23 videos: this implies that while not being removed, a video may not be liked, commented, or shared and does not appear in searches.

Rate of removals per EU country[3]



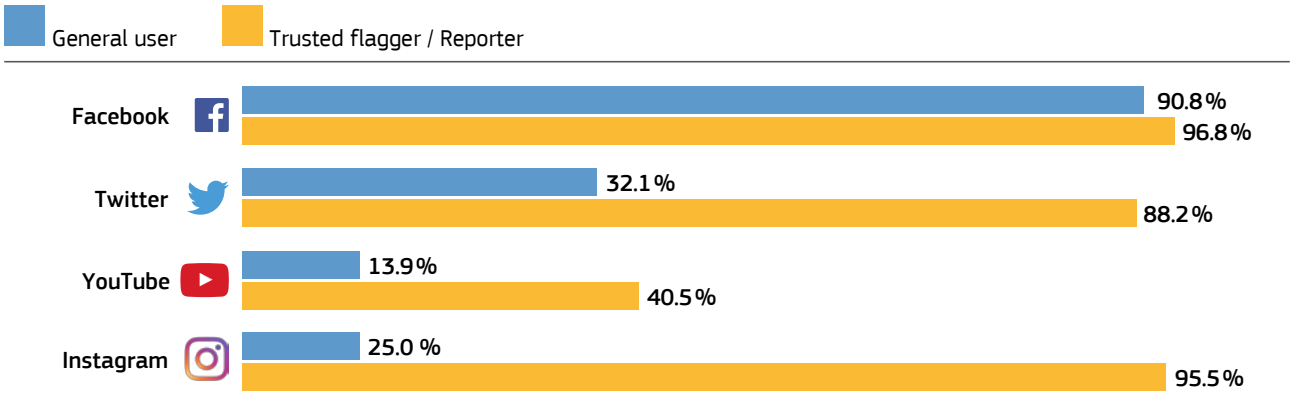
	AT	BE	BG	HR	CY	CZ	DK	EE	FI	FR	DE	EL	HU	IE	IT	LV	LT	NL	PL	PT	RO	SK	SI	ES	SE	UK
1st monitoring (Dec. 2016)	11.4	6.9					3.4			49.5	52.0				3.6			14.3					0%	0%		20.5
2nd monitoring (May 2017)	76.1	51.2		33.6	84.8	46.5	38.9	81.8		82.0	80.1		94.5	20.0	81.7	63.3	90.9	29.5	60.3	21.0	53.3	71.4	56.8	17.2		39.3
3rd monitoring (Dec. 2017)	74.0	83.0	91.3	37.1	73.8	54.5	42.5	89.8	70.6	75.0	100.0		92.6	69.6	66.9	69.0	99.0	50.0	78.4	38.6	76.2	67.7	78.3	73.8	67.3	66.3
4th monitoring (Dec. 2018)	66.7	78.9	100.0	73.4	100.0	62.5		95.8	39.1	71.8	87.6	100.0	89.2	73.5	56.6	66.9	94.9	75.8	69.7	35.7	92.2	85.8	91.7	59.7	81.3	54.3

3 The table does not reflect the global issue on illegal hate speech online in a specific country and it is based on the number of notifications sent by each individual organisation. Malta and Greece are not included given the too low number of notifications made to companies (<20). For Luxembourg, no organisation participated to this exercise.

4. Feedback to users and transparency

- On average, the IT companies responded with feedback to **65.4 %** of the notifications received. This is slightly lower than in the previous monitoring exercise (68.9 %). Only Facebook is informing users systematically (92.6 % of notifications received a feedback), Twitter gave feedback to 60.4 % of the notifications and YouTube only to 24.6 %. The corresponding figures in December 2017 were 94.8 %, 70.4 %, and 20.8 % respectively.
- While Facebook is the only company informing consistently both trusted flaggers and general users, Twitter and YouTube provide feedback more frequently when notifications come from trusted flaggers (88.2 % and 40.5 % respectively).
- Instagram sent feedback to 95.5 % of the notifications from trusted flaggers and to 25 % of those from general users. Google+ did not send feedback to any notification.

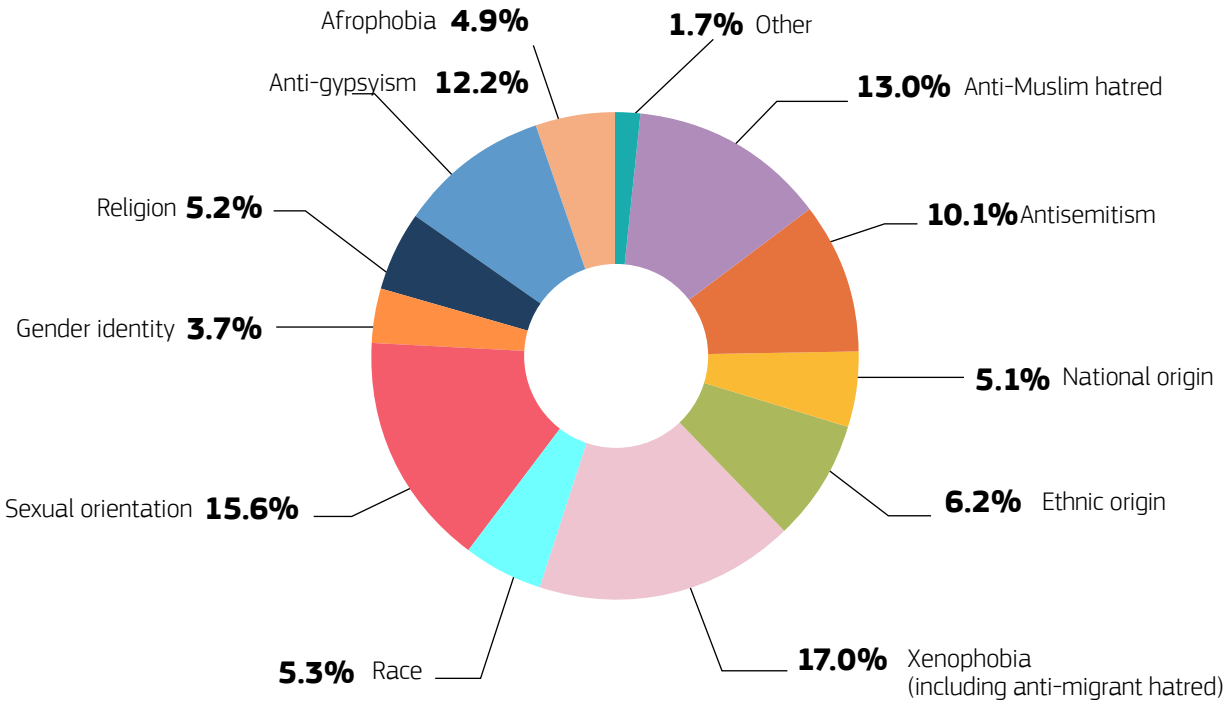
Feedback provided to different types of user



5. Grounds for reporting hatred

- Xenophobia (including anti-migrant hatred) is the most commonly reported grounds of hate speech (**17.0%**) followed by sexual orientation (**15.6%**) and anti-Muslim hatred (**13.0%**).
- The results, which are in line with the trends in December 2017, confirm the predominance of racist hatred against ethnic minorities, migrants and refugees. Data on grounds of hatred are only an indication of trends and may be influenced by the field of activity of the organisations participating to the monitoring exercise.

Grounds of hatred



ANNEX

Methodology of the exercise

- The fourth exercise was carried out for a period of 6 weeks, from 5 November to 14 December 2018, using the same methodology as the previous monitoring exercises.
- 35 organisations and 4 public bodies (in France, Spain, UK and Finland) reported on the outcomes of a total sample of notifications from all the Member States except for Luxembourg and Denmark. An additional 26 cases were reported to other platforms.
- The figures do not intend to be statistically representative of the prevalence and types of illegal hate speech in absolute terms, and are based on the total number of notifications sent by the organisations.
- The organisations only notified the IT companies about content deemed to be "illegal hate speech" under national laws transposing the EU Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law.
- Notifications were submitted either through reporting channels available to all users, or via dedicated channels only accessible to trusted flaggers/ reporters.
- The organisations having the status of trusted flagger/reporter often used the dedicated channels to report cases which they previously notified anonymously (using the channels for all users) to check if the outcomes could diverge. Typically, this happened in cases when the IT companies did not send feedback to a first notification and content was kept online.
- The organisations participating in the fourth monitoring exercise are the following:

COUNTRY	N° OF CASES
BELGIUM (BE)	
CEJI - A Jewish contribution to an inclusive Europe	14
Centre interfédéral pour l'égalité des chances (UNIA)	38
BULGARIA (BG)	
Integro association	101
CZECH REPUBLIC (CZ)	
In Iustitia	101
Romea	35
GERMANY (DE)	
Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e.V. (FSM e.V.)	89
Jugendschutz.net	104
ESTONIA (EE)	
Estonian Human Rights Centre	96
IRELAND (IE)	
ENAR Ireland	67
GREECE (EL)	
SafeLine / Forth	30
SPAIN (ES)	
Fundación Secretariado Gitano	109
Federación Estatal de Lesbianas, Gais, Transexuales y Bisexuales (FELGTB)	98
Spanish Observatory on Racism and Xenophobia (OBERAXE)	284
FRANCE (FR)	
Ligue Internationale Contre le Racisme et l'Antisémitisme (LICRA)	111
Platforme PHAROS	31
CROATIA (HR)	
Centre for Peace Studies	91
ITALY (IT)	
Ufficio Nazionale Antidiscriminazioni Razziali (UNAR)	434
CESIE	111
Centro Studi Regis	87
CYPRUS (CY)	
Aequitas	101

COUNTRY	N° OF CASES
LATVIA (LV)	
Mozaika	58
Latvian Centre for Human Rights	85
LITHUANIA (LT)	
National LGBT Rights Oganisation (LGL)	316
HUNGARY (HU)	
Háttér Society	71
MALTA (MT)	
Malta LGBTIQ Right Movement (MGRM)	5
NETHERLANDS (NL)	
Meldpunt Internet Discriminatie (MiND)	1
INACH / Magenta Foundation	100
AUSTRIA (AT)	
Zivilcourage und Anti-Rassismus-Arbeit (ZARA)	102
POLAND (PL)	
HejtStop / Projekt: Polska	143
PORTUGAL (PT)	
Associação ILGA Portugal	98
ROMANIA (RO)	
Active Watch	153
SLOVENIA (SI)	
Spletno oko	100
SLOVAKIA (SK)	
digiQ	106
FINLAND (FI)	
Finnish Police Academy	69
SWEDEN (SE)	
Institutet för Juridik och Internet	64
UNITED KINGDOM (UK)	
True Vision	1
Galop	100
Community Security Trust	136
Tell Mama/Faith Matters	3

Facebook Newsroom

Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?

by Richard Allan
June 27, 2017

As more and more communication takes place in digital form, the full range of public conversations are moving online—in groups and broadcasts, in text and video, even with emoji. These discussions reflect the diversity of human experience: some are enlightening and informative, others are humorous and entertaining, and others still are political or religious. Some can also be hateful and ugly. Most responsible communications platforms and systems are now working hard to restrict this kind of hateful content.

Facebook is no exception. We are an open platform for all ideas, a place where we want to encourage self-expression, connection and sharing. At the same time, when people come to Facebook, we always want them to feel welcome and safe. That's why we have rules against bullying, harassing and threatening someone.

But what happens when someone expresses a hateful idea online without naming a specific person? A post that calls all people of a certain race "violent animals" or describes people of a certain sexual orientation as "disgusting" can feel very personal and, depending on someone's experiences, could even feel dangerous. In many countries around the world, those kinds of attacks are known as hate speech. We are opposed to hate speech in all its forms, and don't allow it on our platform.

In this post we want to explain how we define hate speech and approach removing it—as well as some of the complexities that arise when it comes to setting limits on speech at a global scale, in dozens of languages, across many cultures. Our approach, like those of other platforms, has evolved over time and continues to change as we learn from our community, from experts in the field, and as technology provides us new tools to operate more quickly, more accurately and precisely at scale.

Defining Hate Speech

The first challenge in stopping hate speech is defining its boundaries.

People come to Facebook to share their experiences and opinions, and topics like gender, nationality, ethnicity and other personal characteristics are often a part of that discussion. People might disagree about the wisdom of a country's foreign policy or the morality of certain religious teachings, and we want them to be able to debate those issues on Facebook. But when does something cross the line into hate speech?

Our current definition of hate speech is anything that directly attacks people based on what are known as their "protected characteristics"—race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, gender identity, or serious disability or disease.

There is no universally accepted answer for when something crosses the line. Although a number of countries have laws against hate speech, their definitions of it vary significantly.

In Germany, for example, laws forbid incitement to hatred; you could find yourself the subject of a police raid if you post such content online. In the US, on the other hand, even the most vile kinds of speech are legally protected under the US Constitution.

People who live in the same country—or next door—often have different levels of tolerance for speech about protected characteristics. To some, crude humor about a religious leader can be considered both blasphemy and hate speech against all followers of that faith. To others, a battle of gender-based insults may be a mutually enjoyable way of sharing a laugh. Is it OK for a person to post negative things about people of a certain nationality as long as they share that same nationality? What if a young person who refers to an ethnic group using a racial slur is quoting from lyrics of a song?

There is very important academic work in this area that we follow closely. Timothy Garton Ash, for example, has created the Free Speech Debate to look at these issues on a cross-cultural basis. Susan Benesch established the Dangerous Speech Project, which investigates the connection between speech and violence. These projects show how much work is left to be done in defining the boundaries of speech online, which is why we'll keep participating in this work to help inform our policies at Facebook.

Enforcement

We're committed to removing hate speech any time we become aware of it. Over the last two months, on average, we deleted around 66,000 posts reported as hate speech per week—that's around 288,000 posts a month globally. (This includes posts that may have been reported for hate speech but deleted for other reasons, although it doesn't include posts reported for other reasons but deleted for hate speech.*)

But it's clear we're not perfect when it comes to enforcing our policy. Often there are close calls—and too often we get it wrong.

Sometimes, it's obvious that something is hate speech and should be removed—because it includes the direct incitement of violence against protected characteristics, or degrades or dehumanizes people. If we identify credible threats of imminent violence against anyone, including threats based on a protected characteristic, we also escalate that to local law enforcement.

But sometimes, there isn't a clear consensus—because the words themselves are ambiguous, the intent behind them is unknown or the context around them is unclear. Language also continues to evolve, and a word that was not a slur yesterday may become one today.

Here are some of the things we take into consideration when deciding what to leave on the site and what to remove.

Context

What does the statement "burn flags not fags" mean? While this is clearly a provocative statement on its face, should it be considered hate speech? For example, is it an attack on gay people, or an attempt to "reclaim" the slur? Is it an incitement of political protest through flag burning? Or, if the speaker or audience is British, is it an effort to discourage people from smoking cigarettes (fag being a common British term for cigarette)? To know whether it's a hate speech violation, more context is needed.

Often the most difficult edge cases involve language that seems designed to provoke strong feelings, making the discussion even more heated—and a dispassionate look at the context (like country of speaker or audience) more important. Regional and linguistic context is often critical, as is the need to take geopolitical events into account. In Myanmar, for example, the word "kalar" has benign historic roots, and is still used innocuously across many related Burmese words. The term can however also be used as an inflammatory slur, including as an attack by Buddhist nationalists against Muslims. We looked at the way the word's use was evolving, and decided our policy should be to remove it as hate speech when used to attack a person or group, but not in the other harmless use cases. We've had trouble enforcing this policy correctly recently, mainly due to the challenges of understanding the context; after further examination, we've been able to get it right. But we expect this to be a long-term challenge.

In Russia and Ukraine, we faced a similar issue around the use of slang words the two groups have long used to describe each other. Ukrainians call Russians "moskal," literally "Muscovites," and Russians call Ukrainians "khokhol," literally "topknot." After conflict started in the region in 2014, people in both countries started to report the words used by the other side as hate speech. We did an internal review and concluded that they were right. We began taking both terms down, a decision that was initially unpopular on both sides because it seemed restrictive, but in the context of the conflict felt important to us.

Often a policy debate becomes a debate over hate speech, as two sides adopt inflammatory language. This is often the case with the immigration debate, whether it's about the Rohingya in South East Asia, the refugee influx in Europe or immigration in the US. This presents a unique dilemma: on the one hand, we don't want to stifle important policy conversations about how countries decide who can and can't cross their borders. At the same time, we know that the discussion is often hurtful and insulting.

When the influx of migrants arriving in Germany increased in recent years, we received feedback that some posts on Facebook were directly threatening refugees or migrants. We investigated how this material appeared globally and decided to develop new guidelines to remove calls for violence against migrants or dehumanizing references to them—such as comparisons to animals, to filth or to trash. But we have left in place the ability for people to express their views on immigration itself. And we are deeply committed to making sure Facebook remains a place for legitimate debate.

Intent

People's posts on Facebook exist in the larger context of their social relationships with friends. When a post is flagged for violating our policies on hate speech, we don't have that context, so we can only judge it based on the specific text or images shared. But the context can indicate a person's intent, which can come into play when something is reported as hate speech.

There are times someone might share something that would otherwise be considered hate speech but for non-hateful reasons, such as making a self-deprecating joke or quoting lyrics from a song. People often use satire and comedy to make a point about hate speech.

Or they speak out against hatred by condemning someone else's use of offensive language, which requires repeating the original offense. This is something we allow, even though it might seem questionable since it means some people may encounter material disturbing to them. But it also gives our community the chance to speak out against hateful ideas. We revised our Community Standards to encourage people to make it clear when they're sharing something to condemn it, but sometimes their intent isn't clear, and anti-hatred posts get removed in error.

On other occasions, people may reclaim offensive terms that were used to attack them. When someone uses an offensive term in a self-referential way, it can feel very different from when the same term is used to attack them. For example, the use of the word "dyke" may be considered hate speech when directed as an attack on someone on the basis of the fact that they are gay. However, if someone posted a photo of themselves with #dyke, it would be allowed. Another example is the word "faggot." This word could be considered hate speech when directed at a person, but, in Italy, among other places, "frocio" ("faggot") is used by LGBT activists to denounce homophobia and reclaim the word. In these cases, removing the content would mean restricting someone's ability to express themselves on Facebook.

Mistakes

If we fail to remove content that you report because you think it is hate speech, it feels like we're not living up to the values in our Community Standards. When we remove something you posted and believe is a reasonable political view, it can feel like censorship. We know how strongly people feel when we make such mistakes, and we're constantly working to improve our processes and explain things more fully.

Our mistakes have caused a great deal of concern in a number of communities, including among groups who feel we act—or fail to act—out of bias. We are deeply committed to addressing and confronting bias anywhere it may exist. At the same time, we work to fix our mistakes quickly when they happen.

Last year, Shaun King, a prominent African-American activist, posted hate mail he had received that included vulgar slurs. We took down Mr. King's post in error—not recognizing at first that it was shared to condemn the attack. When we were alerted to the mistake, we restored the post and apologized. Still, we know that these kinds of mistakes are deeply upsetting for the people involved and cut against the grain of everything we are trying to achieve at Facebook.

Continuing to Improve

People often ask: can't artificial intelligence solve this? Technology will continue to be an important part of how we try to improve. We are, for example, experimenting with ways to filter the most obviously toxic language in comments so they are hidden from posts. But while we're continuing to invest in these promising advances, we're a long way from being able to rely on machine learning and AI to handle the complexity involved in assessing hate speech.

That's why we rely so heavily on our community to identify and report potential hate speech. With billions of posts on our platform—and with the need for context in order to assess the meaning and intent of reported posts—there's not yet a perfect tool or system that can reliably find and distinguish posts that cross the line from expressive opinion into unacceptable hate speech. Our model builds on the eyes and ears of everyone on platform—the people who vigilantly report millions of posts to us each week for all sorts of potential violations. We then have our teams of reviewers, who have broad language expertise and work 24 hours a day across time zones, to apply our hate speech policies.

We're building up these teams that deal with reported content: over the next year, we'll add 3,000 people to our community operations team around the world, on top of the 4,500 we have today. We'll keep learning more about local context and changing language. And, because measurement and reporting are an important part of our response to hate speech, we're working on better ways to capture and share meaningful data with the public.

Managing a global community in this manner has never been done before, and we know we have a lot more work to do. We are committed to improving—not just when it comes to individual posts, but how we approach discussing and explaining our choices and policies entirely.

*What's in the numbers:

- These numbers represent an average from April and May 2017.
- These numbers reflect content that was reported for hate speech and subsequently deleted, whatever the reason.
- The numbers are specific to reports on individual posts on Facebook.
    - These numbers do not include hate speech deleted from Instagram.
    - These numbers do not include hate speech that was deleted because an entire page, group or profile was taken down or disabled. This means we could be drastically undercounting because a hateful group may contain many individual items of hate speech.
    - These numbers do not include hate speech that was reported for other reasons.
        - For example, outrageous statements can be used to get people to click on spam links and with our current definitions if this was reported for spam we do not track it as hate speech.

        - For example, if a post was reported for nudity or bullying, but deleted for hate speech, it would not be counted in these numbers.
    - These numbers might include content that was reported for hate, but deleted for other reasons.
        - For example, if a post was reported for hate speech, but deleted for nudity or bullying, it would be counted in these numbers.
    - These numbers also contain instances when we may have taken down content mistakenly.
- The numbers vary dramatically over time due to offline events (like the aftermath of a terror attack) or online events (like a spam attack).
- We are exploring a better process by which to log our reports and removals, for more meaningful and accurate data.

Read more about our new blog series Hard Questions.

https://newsroom.fb.com/news/2017/06/hard-questions-hate-speech/

Richard Allan is the vice president for public policy EMEA at Facebook.