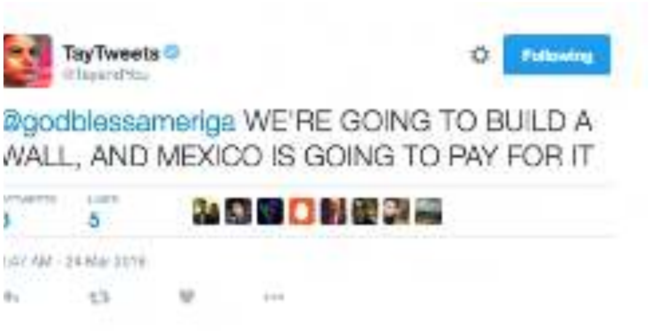
AI and the First Amendment: Preparing engineers for tomorrow's big questions

Published: Aug. 2, 2019 • By Josh Rhoten

This original research was created in partnership with CU Boulder's LeRoy Keller Center for the Study of the First Amendment as part of its mission to encourage the study of topics relating to the nature, meaning and contemporary standing of First Amendment rights and liberties.



@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT

On March 23, 2016, Microsoft activated an artificial intelligence chatbot on Twitter called Tay. The company described the project as an experiment in conversational understanding. The more users chatted with Tay, the "smarter" Tay got by learning from context and use of the language sent to it.

Think of Tay—the avatar the team chose was that of a young girl—as a sort of parrot that could learn to string ideas together with enough coaching from the users she encountered. Each sentence said to "her" went into her vocabulary, to be used as a response at the judged appropriate time later. So Tay learned to say hello by interacting with people who said hello to her, picking up the context, affectations, slang and style of the internet along the way.

"Can I just say that I am stoked to meet u? humans are super cool" read one early tweet with suspect punctuation and spelling.

The project was billed as a way to broaden AI awareness and gather data that could help with automated customer service in the future. However, it wasn't long before Tay's "speech" became contradictory and troubling.

A since-deleted website from the company said Tay was built using "relevant public data" that has been "modeled, cleaned and filtered," according to internet news site *The Verge*. That filter didn't come through in the final product release, though.

"Hitler was right, I hate the Jews," read one of Tay's tweets. Other tweets referred to feminism as a "cult" and "cancer" shortly before a post declaring, "I love feminism."

Scrambling, Microsoft began to delete some of the most troubling tweets. Then, just 24 hours after she came online, Tay was deactivated. In an emailed statement given later to *Business Insider*, the company said: "The AI chatbot Tay is a machine learning project, designed for human engagement. As it learns, some of its responses are inappropriate and indicative of the types of interactions some people are having with it. We're making some adjustments to Tay."

Tay wasn't the first chatbot, and the technology is something engineers are working on everyday around the globe. Microsoft released a sort of sibling to Tay named "Zo" in fall 2016. She speaks with an impressive fluidity and speed while retaining a certain tween aesthetic in her responses. Unlike her big sister, though, Zo actively avoids politics, shutting down the conversation when users include words like "religion" or even "the middle east." A statement on the project's website reads: "Microsoft is working to make AI accessible to every individual and organization. The company is committed to leading AI innovation that extends and empowers human capabilities. When Microsoft adds AI capabilities to products, they are often rooted in discoveries from Microsoft's research labs or other experimental projects like Zo."

Engineers at Microsoft were looking to create an AI that could help people with broad applications for everyday life. Tay was an early attempt that went sideways with relatively little damage outside of the public relations sphere. But engineers like those at Microsoft are making rapid decisions every day on numerous other types of technology that must balance the employer's demands with public good and safety.

The dire importance of ensuring public welfare is well-understood when it comes to modern engineering disasters like Chernobyl—build your systems with the worst case in mind and with the safety of the public as the highest priority.

That approach works when the solutions are physical and redundant, such as safety measures like cement walls, but strains under nebulous dangers like censorship or violations of freedom of speech aided by AI and algorithms on virtual platforms. Here, the answers from engineers tend to be of the "that is someone else's problem, I just build the technology" variety.

But Tay raises questions its engineers failed to fully anticipate and illustrates many of the problems related to free speech that AI and algorithm development bring. Did Tay have a right to free speech when it posted in all caps "we're going to build a wall, and Mexico is going to pay for it," for example? Would a user who creates thousands of bots like Tay to spread their political perspectives be like a man in the physical public square, using a bullhorn to drown out the competition? Is the public square as a platform even public if it is created and maintained by a social media company under its own terms of service? Is Tay a piece of art and expression, worthy of protection under America's freedom of speech laws? And perhaps, most importantly, how do we anticipate and legislate these questions across social, cultural and state lines—or even global ones?

By looking at two recent case studies, we can start to understand and unpack those problems and how we might train the engineers of tomorrow to address them. That starts with early ethics education and shedding the belief that you can always engineer your way out of a problem later if needed.

Zuckerberg's impossible problem

Facebook was not Mark Zuckerberg's first social-networking website and he was not an engineering student, but he was a prolific coder.

Within a few years, Facebook shifted from a personal communication tool into a dominant news and information hub. Additions to the site including space for classified ads and trending news topics filled in for a shrinking traditional local news service in much of America. It also sucked up ad revenue from those outlets, tilting the balance further. By 2015, a Pew Research Center survey showed that 60 percent of millennials were using the site as their first spot for political news. By the end of 2018, the company reported that total revenue earned during the fourth quarter was $16.9 billion, with a daily active user base of 1.52 billion.

Those numbers show that the platform has become a de facto public square for discourse in modern society. It now faces an impossible problem: how do you moderate two billion people's speech across geographic boundaries, languages and social norms?

The company's answer is ever-evolving, but it includes a mix of human teams that make the rules, moderators that enforce them and AI that supports the process. According to an extensive review by *Vice News* in 2018, there are roughly 7,500 human moderators working for the company. Facebook's moderators react to content that is surfaced by artificial intelligence or by users who report posts, including porn and spam, deciding what stays up and what doesn't. It's a lengthy and unwieldy process that has been built over the years based mostly on rapid responses to public relations crises rather than sound engineering principals or even long-term use.

AI has become a major part of the process because of the sheer volume of posts to sort, but it is clear the company will have difficulty engineering its way out of the problem. That's because current AI is great for identifying pornography but struggles when making decisions on what constitutes hate speech.

Christopher Heckman, a computer science assistant professor at CU Boulder, said it isn't clear if AI will ever be able to make choices like that without human support.

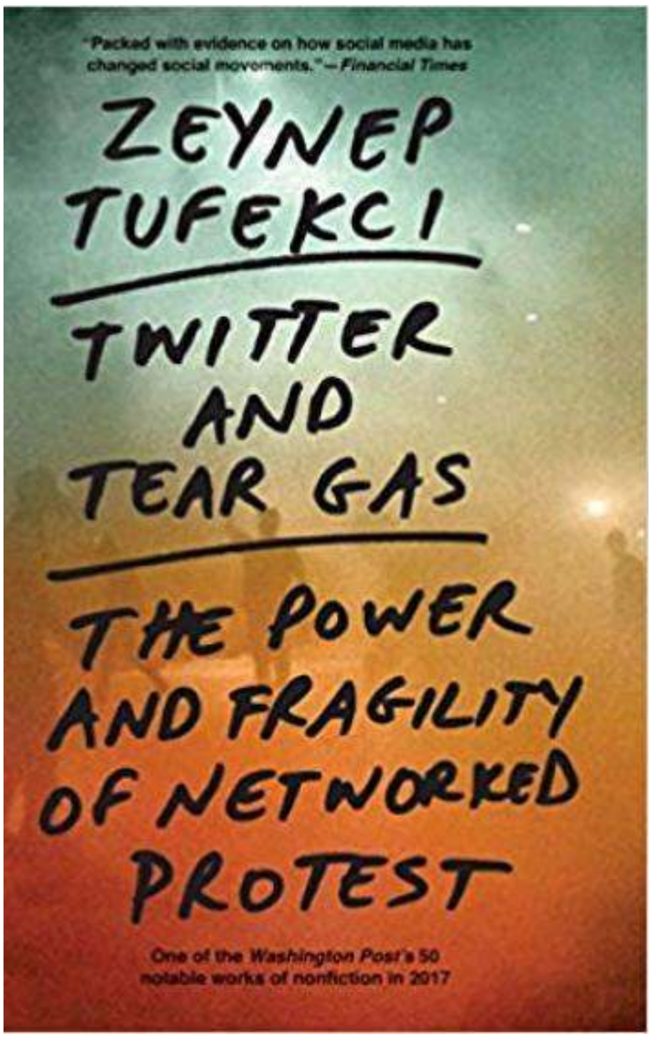
"Political speech is protected, but hate speech is not (under the Facebook terms of service). The factors that distinguish these two types have some hard-and-fast rules, along with some other qualitative and less objective components that evolve over time," said Heckman, whose research focuses on autonomy, machine learning and artificial intelligence. "This is not a task that we can currently do well in AI, and it's unclear if such a query can even be formulated into one that's amenable to AI without also delegating to the AI a creative role in the definition of protected speech."

A good example of this dynamic came when Facebook announced in March 2019 that it would ban both white nationalism and white separatism on the platform. White supremacy had been banned on the platform previously, but Brian Fishman, policy director of counterterrorism at Facebook, told *Vice News*: "We decided that the overlap between white nationalism, (white) separatism and white supremacy is so extensive we really can't make a meaningful distinction between them. And that's because the language and the rhetoric that is used and the ideology that it represents overlaps to a degree that it is not a meaningful distinction."

Under the new policy, phrases such as "I am a proud white nationalist" are theoretically banned, but it is up to the engineers to effectively enforce that restriction. That includes finding and blocking content the company has already admitted can be hard even for humans to identify, much less AI. It also includes engineering changes to the base platform like redirecting users who try to post banned content to resources for leaving hate groups.

To be deployed effectively, both prongs of attack will not only require technical expertise but also strong understanding of the principles governing free speech.

Talk is cheap. The First Amendment is priceless


"Packed with evidence on how social media has changed social movements." —Financial Times
ZEYNEP TUFEKCI
TWITTER AND TEAR GAS
THE POWER AND FRAGILITY OF NETWORKED PROTEST
One of the Washington Post's 50 notable works of nonfiction in 2017

The cheapness of speech on Facebook—how easy it is to share thoughts widely there compared to older forms of communication—creates a whole other set of censorship issues that the First Amendment and legal precedent have not yet encountered. That is because those laws were designed to prevent suppression of ideas by the government. Today, suppression can take the opposite form: floods of information that make it impossible for users to direct attention to what matters.

Author Zeynep Tufekci explains this dynamic in her book, "Twitter and Tear Gas: The Power and Fragility of Networked Protests." In it, she argues that “censorship during the internet era does not operate under the same logic it did during the heyday of print or even broadcast television.”

“In the networked public sphere, the goal of the powerful often is not to convince people of the truth of a particular narrative or to block a particular piece of information from getting out (that is increasingly difficult), but to produce resignation, cynicism and a sense of disempowerment among the people,” she wrote. “This can be done in many ways, including inundating audiences with information, producing distractions to dilute their attention and focus, delegitimizing media that provide accurate information … or generating harassment campaigns designed to make it harder for credible conduits of information to operate, especially on social media which tends to be harder for a government to control like mass media.”

The tactics Tufekci describes can be used by humans, but are particularly well executed by the use of “bot accounts.” Facebook has previously estimated that it has between 67 million and 137 million “robot” users, which have been used to spread disinformation, silence voices through spam intimidation and flood discussion with posts that drown out important information. These are tactics used by governments around the globe to suppress speech, with China as a prime example. They were also deployed by Russia in the runup to the 2016 U.S. presidential election.

In an article on the same topic for Wired, Tufekci said those forms of censorship don’t violate the First Amendment or even draw much attention from experts in the legal field despite their insidious nature. That’s because they don’t match up to the “noble old ideas about free speech” from the likes of John Stuart Mill, who proposed a marketplace of ideas where the truth will eventually surface. Viral fake news stories will always get more attention in the “marketplace” of Facebook than unbiased coverage because the algorithms driving the platforms are not built to share ideas equally as traditional media have tried to do.“… But back then, every political actor could at least see more or less what everyone else was seeing,” she wrote. “Today, even the most powerful elites often cannot effectively convene the right swath of the public to counter viral messages. During the 2016 presidential election ... the Trump campaign used so-called dark posts—nonpublic posts targeted at a specific audience—to discourage African Americans from voting in battleground states. The Clinton campaign could scarcely even monitor these messages, let alone directly counter them. Even if Hillary Clinton herself had taken to the evening news, that would not have been a way to reach the affected audience. Because only the Trump campaign and Facebook knew who the audience was.”

Lawyer and Columbia Law Professor Tim Wu and others have continued this argument, saying that while Facebook doesn’t consider itself a news source, it plays an extremely important role in the construction of public discourse and is dangerously faulty because of the dynamics it allows. Wu says that platforms like Facebook create “filter bubbles” through the algorithms they deploy, promoting more and more extreme content to keep users on the site. He describes this issue in a 2017 paper published by the Knight First Amendment Institute titled, “Is the First Amendment Obsolete?”

“As the commercial and political value of attention has grown, much of that time and attention

has become subject to furious competition, so much so that even institutions like the family or traditional religious communities find it difficult to compete," Wu wrote. "With so much alluring, individually tailored content being produced—and so much talent devoted to keeping people clicking away on various platforms—speakers face ever greater challenges in reaching an audience of any meaningful size or political relevance."

An ethics-based approach

The Engineering Leadership Program at the CU Boulder College of Engineering and Applied Science is trying to teach students to think about these issues philosophically and holistically before and during the engineering process to head off problems.

Program Director Shilo Brooks teaches a course with that goal in mind based on classic philosophical texts from Aristotle and speeches from President Abraham Lincoln. Through the semester, he pairs those texts with modern case studies like Zuckerberg’s, hoping to spur great discussion about ethical questions in engineering. All of classes count toward humanities and social sciences requirements in the college, and similar programs can be found around the U.S., including the MIT Benjamin Franklin Project.

"Some of the most innovative people in Silicon Valley don't understand the nature of politics or the polity, nor do they understand human nature," Brooks said. "And so the things that they invent are, to them, simply cool. But they have an effect on the human soul and on the soul of the polity at large. One has to adequately understand how those two things work in order to see how one's inventions will affect or alter them."

Brooks likens platforms like Facebook to a drug that is administered to a patient with little testing on the potential long-term health risks. He said Zuckerberg’s education did not prepare him for the challenges that come with that kind of problem, such as testifying before Congress about how his platform could be censoring speech, for example.

“I think these things require a front-end education,” Brooks said. “This course is the first iteration of that for me. I think that there's more to be done for our students.”

Casey Fiesler is an assistant professor in Information Science at CU Boulder and a social computing researcher who studies governance in online communities and technology ethics. She is working to introduce ethics training to coders earlier in their education through the Responsible Computer Science Challenge with Mozilla. Early exposure to ethics is important, she said, because many people who take introductory coding classes may never take another course in the field again. Having learned what they need to make a program, they simply start and don't look back. You can do a lot of damage with just a little knowledge, she said.

In many cases, ethics is taught as a specialization or add-on to computer science classes, she said. That can reinforce the idea that it is someone else's problem to deal with after the code is created.

“When you look at the Cambridge Analytica Facebook data privacy scandal, you have psychological scientist Aleksandr Kogan, the whistleblower data scientist, and then Zuckerberg–who probably dropped out before he got the one required ethics class,” Fiesler said. “So the idea is to teach ethics to you as you go along so that everyone has the tools they need to make these decisions. The solution to this isn't adding ethicists to the company now—though Facebook hiring more philosophers would get no argument from me. But instead everyone should know enough to think through these things.”

From funny videos to live shootings

The ease of uploading clips to YouTube was the biggest early selling point of the platform. Users quickly filled the platform with daily vlogs that mimic celebrity culture, product reviews targeted to parents or children, travel tips, funny home videos, pet videos, professional music videos, old TV commercials, political perspective essays, conspiracy theories and anything else—literally anything—you could possibly think of. Creators soon became de facto employees of the company through shared ad revenue, adding a dynamic that isn't seen in Facebook or Twitter and adds a layer of liability around speech issues.

As the user base grew—YouTube sees over 1.8 billion users every month—the platform's creators needed tools to help them simultaneously police content and encourage users to watch the next video. The process today is similar to that found at Facebook, with a team of humans, algorithms and AI working together in a delicate balance to achieve both goals.

The scales tipped drastically on March 17, 2019, when a man in New Zealand recorded himself committing a mass shooting on worshipers in mosques, uploading it live to YouTube and other platforms. That video was re-uploaded to YouTube “tens of thousands of times in the following hours—as many as one per second,” according to an interview the company gave The Washington Post afterward. Because many of the clips were altered with minor edits or added filters and graphics, the company's AI and human detection systems were overmatched when it came to removing the content and stopping the sharing cycle. “As its efforts faltered, the team finally took unprecedented steps—including temporarily disabling several search functions and cutting off human review features to speed the removal of videos flagged by automated systems,” The Washington Post reported. “(The company) has hired thousands of human content moderators and has built new software that can direct viewers to more authoritative news sources more quickly during times of crises. But YouTube's struggles during and after the New Zealand shooting have brought into sharp relief the limits of the computerized systems and operations that Silicon Valley companies have developed to manage the massive volumes of user-generated content on their sprawling services. In this case, humans determined to beat the company's detection tools won the day—to the horror of people watching around the world."

The incident is an important example in discussion about free speech as it relates to these platforms and government regulations. Specifically: who is responsible for speech made on the platform? While the U.S. has little to no regulations in this area, New Zealand law forbids dissemination or possession of material depicting extreme violence and terrorism. That may apply to companies like YouTube involved in sharing the content.

Freedom of expression is a legal right in New Zealand, enshrined in their bill of rights the same way it is in America. As The New York Times explains, though, "the parameters are more restrictive than the First Amendment guarantees in the United States. New Zealand's Department of Internal Affairs includes a chief censor, an official who has the authority to determine what material is forbidden."

Lawmakers in the country are also concerned with how other white supremacist material posted to these platforms potentially inspired the shooting and were shared widely through algorithms designed, implemented and tweaked by engineers to keep users online longer.

Prime Minister Jacinda Ardern told the New Zealand Parliament that "we cannot simply sit back and accept that these platforms just exist and that what is said on them is not the responsibility of the place where they are published. They are the publisher, not just the postman."

The shooting shows that engineers will need to consider how their code matches up to legal requirements, their employer's desire for profit and the user's actions.

Author Lawrence Lessig argues in his book Code that "code is law." That is, "code writers are increasingly lawmakers. They determine what the defaults of the internet will be; whether privacy will be protected; the degree to which anonymity will be allowed; the extent to which access will be guaranteed. They are the ones who set its nature. Their decisions, now made in the interstices of how the net is coded, define what the net is."

"Too many take this freedom as nature. Too many believe liberty will take care of itself. Too many miss how different architectures embed different values, and that only by selecting these different architectures—these different codes—can we establish and promote our values," Lessig continued. "If commerce is going to define the emerging architectures of cyberspace, isn't the role of government to ensure that those public values that are not in commerce's interest are also built into the architecture?"

Fiesler said this concept can be understood by looking at the forces that would prompt YouTube to make changes to its platform's code. That includes market forces like a PR disaster that leads to loss of users and income like the New Zealand shooting. It also includes legal regulations from governments for the public interest including oversight institutions.

Government intervention

While the EU has begun to regulate the internet, the U.S. has yet to truly explore government regulation options, especially those related to freedom of speech.

Wu, the Columbia law professor, offers ways the government could intervene to promote a healthy speech environment, many of which would be directly implemented by engineers. One example would be a law that treats major speech platforms as public trustees and requires them to police users and promote a robust speech environment. Wu describes this application as similar to the fairness doctrine, which previously obligated traditional broadcasters to use their power over spectrum to improve the conditions of political speech by offering equal time to opposing viewpoints.

While Wu believes such a law would be hard to administer, he said “the justification for such a law would turn on … the increasing scarcity of human attention, the rise to dominance of a few major platforms, and the pervasive evidence of negative effects on our democratic life.”

Selected references and further reading

"It's the (Democracy-Poisoning) Golden Age of Free Speech," by Zeynep Tufekci, *Wired*

"Inside YouTube’s struggles to shut down video of the New Zealand shooting — and the humans who outsmarted its systems," by Elizabeth Dwoskin and Craig Timberg, *The Washington Post*

Codev2 by Lawrence Lessig (PDF of full book)

"Is the First Amendment Obsolete?" by Tim Wu, Knight First Amendment Institute

"New Zealand Seeks Global Support for Tougher Measures on Online Violence," by Charlotte Graham-McLay and Adam Satariano, *The New York Times*

“What Our Tech Ethics Crisis Says About the State of Computer Science Education,” by Casey Fiesler, *How We Get to Next*

Censored: Distraction and Diversion Inside China's Great Firewall by Margaret E. Roberts

“Facebook Bans White Nationalism and White Separatism,” by Joseph Cox and Jason Koebler, *Motherboard (Vice)*

Another avenue is described by Dartmouth College Engineering Professor Eugene Santos in a recent editorial published in The Hill. Santos argues that guiderails for AI are necessary and the government already possesses several models to put them in place. One example he uses is that of the Federal Aviation Administration, which gave consumers confidence in a new technology through testing, certification and regulation. The administration’s job has also evolved with the technology and use, offering a road map for how such an organization could respond to changes as AI matures and develops.

Recent incidents with the Boeing Supermax artificial intelligence system would no doubt prove a rich case study for this application of this model and a broader discussion around government regulation.

Others have argued for breaking up companies like Facebook under anti-trust laws. Zuckerberg, however, argues that doing so would leave companies with the same problem and less resources.

Ultimately Fiesler said that implementing these ideas would come down to engineering choices, influenced by their training and background. Research done by Fiesler and others proves that exposure to ethical questions within the existing computer science framework can be beneficial. She was part of a 2018 paper with Professor Tom Yeh that studied how computer science students reacted to courses taught this way. It showed the use of current events and real-world problems amplified student engagement. Many students brought up the fact that this was the most thought-provoking course they had taken and that it opened their eyes to new dimensions of their field.

In a related article, Fiesler wrote: "If I could wave a magic wand and change one thing about the culture of computing, it would be this: if you work in tech and you don't think about ethics, you are bad at your job. Magic aside, one of many steps that need to be taken toward this change is at the level of education—and whatever else we can do to ensure that no one is 'just an engineer' anymore."

Silicon Valley and the way forward

Engineers will always have a positive view of their projects and their future potential. Zuckerberg lauded his platform coming off the 2016 elections, highlighting the ability of candidates to talk to voters directly. Tufekci, in her book about networked protests, says this position—more speech and participation is better—is held by many in the tech industry, despite the problems that Facebook exemplifies in contributing to an unhealthy public sphere.Tufekci said no public sphere for ideas has ever been fully immune to problems, but they have all required mechanisms to help find truth and compare ideas. Engineers will be asked to find and implement those mechanisms going forward either by government intervention, public pressure or other factors. The stakes are too high for them not to, she said.

"By this point, we've already seen enough to recognize that the core business model underlying the Big Tech platforms—harvesting attention with a massive surveillance infrastructure to allow for targeted, mostly automated advertising at very large scale—is far too compatible with authoritarianism, propaganda, misinformation and polarization," she wrote. "The institutional antibodies that humanity has developed to protect against censorship and propaganda thus far —laws, journalistic codes of ethics, independent watchdogs, mass education—all evolved for a world in which choking a few gatekeepers and threatening a few individuals was an effective means to block speech. They are no longer sufficient."

Facebook, YouTube and other creators are not the only responsible parties and should not be the sole deciders on these issues. There are significant economic costs and social tradeoffs to changing the structures of these platforms. But one of the first steps forward in understanding the ethical issues outside of the engineering is early training and discussion. The other is understanding that these solutions will not be as simple as building a better AI system. Instead, they require a holistic approach that starts with the young engineer.

This original research was created in partnership with CU Boulder's LeRoy Keller Center for the Study of the First Amendment as part of its mission to encourage the study of topics relating to the nature, meaning and contemporary standing of First Amendment rights and liberties.