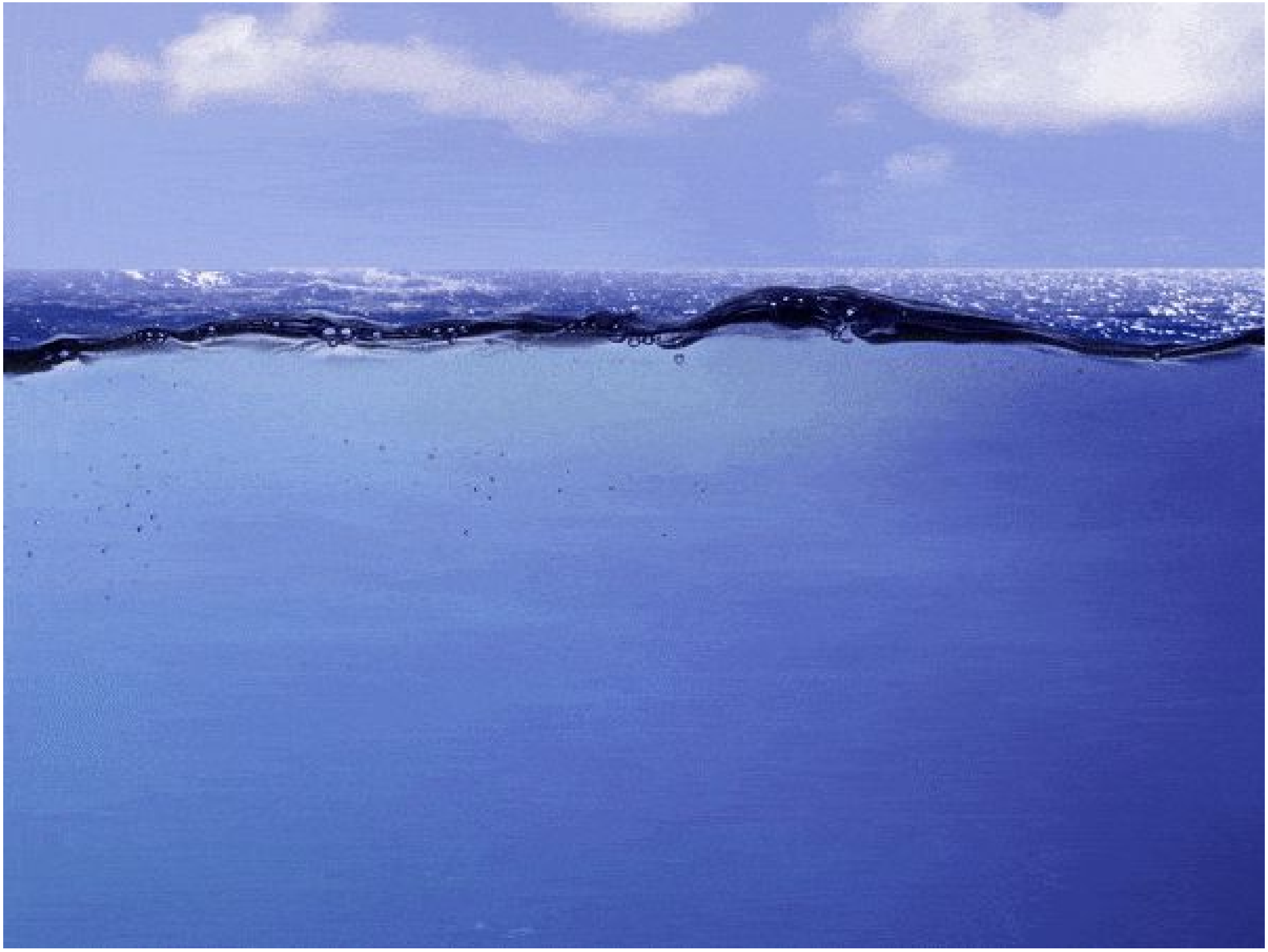
A REPORTER AT LARGE OCTOBER 19, 2020 ISSUE

WHY FACEBOOK CAN’T FIX ITSELF

The platform is overrun with hate speech and disinformation. Does it actually want to solve the problem?

By Andrew Marantz

October 12, 2020



Animation by Javier Jaén; photograph from Getty

□ Listen to this story

When Facebook was founded, in 2004, the company had few codified rules about what was allowed on the platform and what was not. Charlotte Willner joined three years later, as one of the company's first employees to moderate content on the site. At the time, she said, the written guidelines were about a page long; around the office, they were often summarized as, "If something makes you feel bad in your gut, take it down." Her husband, Dave, was hired the following year, becoming one of twelve full-time content moderators. He later became the company's head of content policy. The guidelines, he told me, "were just a bunch of examples, with no one articulating the reasoning behind them. 'We delete nudity.' 'People aren't allowed to say nice things about Hitler.' It was a list, not a framework." So he wrote a framework. He called the document the Abuse Standards. A few years later, it was given a more innocuous-sounding title: the Implementation Standards.

These days, the Implementation Standards comprise an ever-changing wiki, roughly twelve thousand words long, with twenty-four headings—"Hate Speech," "Bullying," "Harassment," and so on—each of which contains dozens of subcategories, technical definitions, and links to supplementary materials. These are located on an internal software system that only content moderators and select employees can access. The document available to Facebook's users, the Community Standards, is a condensed, sanitized version of the guidelines. The rule about graphic content, for example, begins, "We remove content that glorifies violence." The internal version, by contrast, enumerates several dozen types of graphic images—"charred or burning human beings"; "the detachment of non-generating body parts"; "toddlers smoking"—that content moderators are instructed to mark as "disturbing," but not to remove.

Facebook's stated mission is to "bring the world closer together." It considers itself a neutral platform, not a publisher, and so has resisted censoring its users' speech, even when that speech is ugly or unpopular. In its early years, Facebook weathered periodic waves of bad press, usually occasioned by incidents of bullying or violence on the platform. Yet none of this seemed to cause lasting damage to the company's reputation, or to its valuation. Facebook's representatives repeatedly claimed that they took the spread of harmful content seriously, indicating that they could manage the problem if they were only given more time. Rashad Robinson, the president of the racial-justice group Color of Change, told me, "I don't want to sound naïve, but until recently I was willing to believe that they were committed to making real progress. But then the hate speech and the toxicity keeps multiplying, and at a certain point you go, Oh, maybe, despite what they say, getting rid of this stuff just isn't a priority for them."

There are reportedly more than five hundred full-time employees working in Facebook's P.R. department. These days, their primary job is to insist that Facebook is a fun place to share baby photos and sell old couches, not a vector for hate speech, misinformation, and violent extremist propaganda. In July, Nick Clegg, a former Deputy Prime Minister of the U.K. who is now a top flack at Facebook, published a piece on AdAge.com and on the company's official blog titled "Facebook Does Not Benefit from Hate," in which he wrote, "There is no incentive for us to do anything but remove it." The previous week, Guy Rosen, whose job title is vice-president for integrity, had written, "We don't allow hate speech on Facebook. While we recognize we have more to do . . . we are moving in the right direction."

It would be more accurate to say that the company is moving in several contradictory directions at once. In theory, no one is allowed to post hate speech on Facebook. Yet many world leaders—Rodrigo Duterte, of the Philippines; Narendra Modi, of India; Donald Trump; and others—routinely spread hate speech and disinformation, on Facebook and elsewhere. The company could apply the same standards to demagogues as it does to everyone else, banning them from the platform when necessary, but this would be financially risky. (If Facebook were to ban Trump, he would surely try to retaliate with onerous regulations; he might also encourage his supporters to boycott the company.) Instead, again and again, Facebook has erred on the side of allowing politicians to post whatever they want, even when this has led the company to weaken its own rules, to apply them selectively, to creatively reinterpret them, or to ignore them altogether.

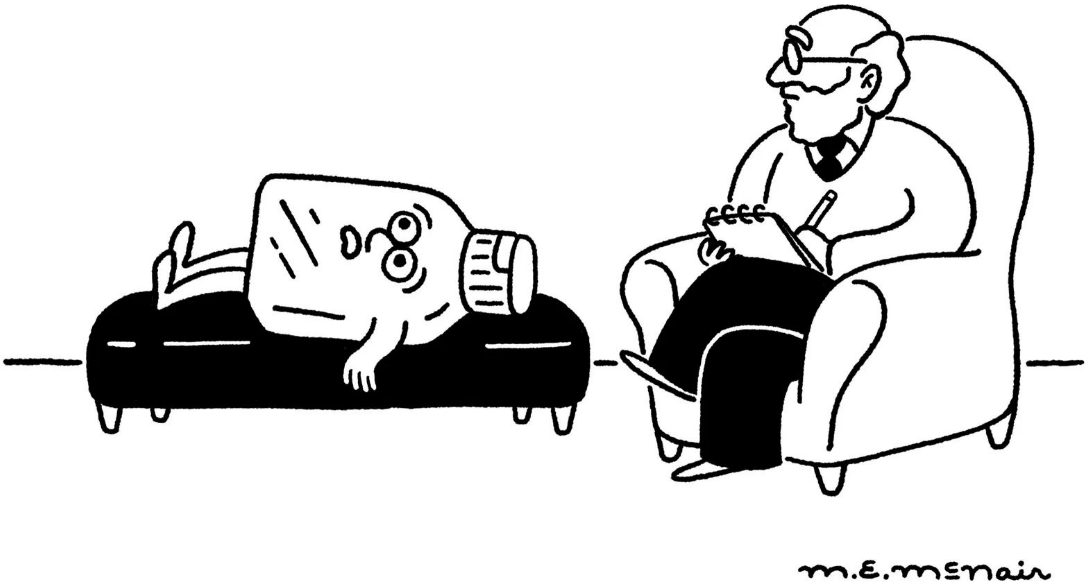
Dave Willner conceded that Facebook has "no good options," and that censoring world leaders might set "a worrisome precedent." At the same time, Facebook's stated reason for forbidding hate speech, both in the Community Standards and in public remarks by its executives, is that it can lead to real-world violence. Willner went on, "If that's their position, that hate speech is inherently dangerous, then how is it not more dangerous to let people use hate speech as long as they're powerful enough, or famous enough, or in charge of a whole army?"

The Willners left Facebook in 2013. (Charlotte now runs the trust-and-safety department at Pinterest; Dave is the head of community policy at Airbnb.) Although they once considered themselves "true believers in Facebook's mission," they have become outspoken critics of the company. "As far as I can tell, the bulk of the document I wrote hasn't changed all that much, surprisingly," Dave Willner told me. "But they've made some big carve-outs that are just absolute nonsense. There's no perfect approach to content moderation, but they could at least try to look less transparently craven and incoherent."

In a statement, Drew Pusateri, a spokesperson for Facebook, wrote, "We've invested billions of dollars to keep hate off of our platform." He continued, "A recent European Commission report found that Facebook assessed 95.7% of hate speech reports in less than 24 hours, faster than YouTube and Twitter. While this is progress, we're conscious that there's more work to do." It is possible that Facebook, which owns Instagram, WhatsApp, and Messenger, and has more than three billion monthly users, is so big that its content can no longer be effectively moderated. Some of Facebook's detractors argue that, given the public's widespread and justified skepticism of the company, it should have less power over users' speech, not more. "That's a false choice," Rashad Robinson said. "Facebook already has all the power. They're just using it poorly." He pointed out that

Facebook consistently removes recruitment propaganda by ISIS and other Islamist groups, but that it has been far less aggressive in cracking down on white-supremacist groups. He added, "The right question isn't 'Should Facebook do more or less?' but 'How is Facebook enforcing its rules, and who is set up to benefit from that?' "

In public, Mark Zuckerberg, Facebook's founder, chairman, and C.E.O., often invokes the lofty ideals of free speech and pluralistic debate. During a lecture at Georgetown University last October, he said, "Frederick Douglass once called free expression 'the great moral renovator of society.' " But Zuckerberg's actions make more sense when viewed as an outgrowth of his business model. The company's incentive is to keep people on the platform—including strongmen and their most avid followers, whose incendiary rhetoric tends to generate a disproportionate amount of engagement. A former Facebook employee told me, "Nobody wants to look in the mirror and go, I make a lot of money by giving objectively dangerous people a huge megaphone." This is precisely what Facebook's executives are doing, the former employee continued, "but they try to tell themselves a convoluted story about how it's not actually what they're doing."



"Who sanitizes the sanitizer?"

Cartoon by Elisabeth McNair

Open cartoon gallery

In retrospect, it seems that the company's strategy has never been to manage the problem of dangerous content, but rather to manage the public's perception of the problem. In Clegg's recent blog post, he wrote that Facebook takes a "zero tolerance approach" to hate speech, but that, "with so much content posted every day, rooting out the hate is like looking for a needle in a haystack." This metaphor casts Zuckerberg as a hapless victim of fate: day after day, through no fault of his own, his haystack ends up mysteriously full of needles. A more honest metaphor would posit a powerful set of magnets at the center of the haystack—Facebook's algorithms, which attract and elevate whatever content is most highly charged. If there are needles anywhere nearby—and, on the Internet, there always are—the magnets will pull them in. Remove as many as you want today; more will reappear tomorrow. This is how the system is designed to work.

On December 7, 2015, Donald Trump, then a dark-horse candidate for the Republican Presidential nomination, used his Facebook page to promote a press release. It called for "a total and complete shutdown of Muslims entering the United States" and insinuated that Muslims—all 1.8 billion of them, presumably —"have no sense of reason or respect for human life." By Facebook's definition, this was clearly hate speech. The Community Standards prohibited all "content that directly attacks people based on race, ethnicity, national origin, or religion." According to the *Times*, Zuckerberg was personally "appalled" by Trump's post. Still, his top officials held a series of meetings to decide whether, given Trump's prominence, an exception ought to be made.

The discussions were led by Monika Bickert, Elliot Schrage, and Joel Kaplan, all policy executives with law degrees from Harvard. Most of Facebook's executives were liberal, or were assumed to be. But Kaplan, an outspoken conservative who had worked as a clerk for Justice Antonin Scalia and as a staffer in the George W. Bush White House, had recently been promoted to the position of vice-president of global public policy, and often acted as a liaison to Republicans in Washington, D.C. His advice to Zuckerberg, the *Times* later reported, was "Don't poke the bear"—avoid incurring the wrath of Trump and his supporters. Trump's post stayed up. The former Facebook employee told me, "Once you set a precedent of caving on something like that, how do you ever stop?"

Making the decision to leave Trump's post up was one thing; justifying the decision was another. According to the Washington *Post*, Bickert drafted an internal memo, laying out the options that she and her colleagues had. They could make "a one-time exception" for Trump's post, which would establish a narrow precedent that would allow them to reverse course later. They could add an "exemption for political discourse" to the guidelines, which would let them treat politicians' future utterances on a case-by-case basis. Or they could amend the rules more expansively—for example, by "weakening the company's community guidelines for everyone, allowing comments such as 'No blacks allowed' and 'Get the gays out of San Francisco.' "

At the time, Facebook had fewer than forty-five hundred content moderators. Now there are some fifteen thousand, most of whom are contract workers in cities around the world (Dublin, Austin, Berlin, Manila). They often work at odd hours, to account for time-zone differences, absorbing whatever pops up on their screens: threats, graphic violence, child pornography, and every other genre of online iniquity. The work can be harrowing. "You're sleep-deprived, your subconscious is completely open, and you're pouring in the most psychologically radioactive content you can imagine," Martin Holzmeister, a Brazilian art director who worked as a moderator in Barcelona, told me. "In Chernobyl, they knew, you can run in for two minutes, grab something, and run back out, and it won't kill you. With this stuff, nobody knows how much anyone can take." Moderators are required to sign draconian nondisclosure agreements that forbid them to discuss their work in even the most rudimentary terms. In May, thousands of moderators joined a class-action suit against Facebook alleging that the job causes P.T.S.D. (Facebook settled the suit, paying the moderators fifty-two million dollars. Pusateri, the Facebook spokesperson, said that the company provides its moderators with on-site counselling and a twenty-four-hour mental-health hotline.)

One of Facebook's main content-moderation hubs outside the U.S. is in Dublin, where, every day, moderators review hundreds of thousands of reports of potential rule violations from Europe, Africa, the Middle East, and Latin America. In December, 2015, several moderators in the Dublin office—including some on what was called the MENA team, for Middle East and North Africa—noticed that Trump's post was not being taken down. "An American politician saying something shitty about Muslims was probably not the most shocking thing I saw that day," a former Dublin employee who worked on content policy related to the Middle East told me. "Remember, this is a job that involves looking at beheadings and war crimes." The MENA team, whose members spoke Arabic, Farsi, and several other languages, was not tasked with moderating American content; still, failing to reprimand Trump struck many of them as a mistake, and they expressed their objections to their supervisors. According to Facebook's guidelines, moderators were to remove any "calls for exclusion or segregation." An appeal to close the American border to Muslims clearly qualified.

The following day, members of the team and other concerned employees met in a glass-walled conference room. At least one policy executive joined, via video, from the U.S. "I think it was Joel Kaplan," the former Dublin employee told me. "I can't be sure. Frankly, I had trouble telling those white guys apart." The former Dublin employee got the impression that "the attitude from the higher-ups was You emotional Muslims seem upset; let's have this conversation where you feel heard, to calm you down. Which is hilarious, because a lot of us weren't even Muslim. Besides, the objection was never, Hey, we're from the Middle East and this hurts our feelings." Rather, their message was "In our expert opinion, this post violates the policies. So what's the deal?"

Facebook claims that it has never diluted its protections against hate speech, but that it sometimes makes exceptions in the case of newsworthy utterances, such as those by people in public office. But a recently acquired version of the Implementation Standards reveals that, by 2017, Facebook had weakened its rules —not just for politicians but for all users. In an internal document called the Known Questions—a Talmud-like codicil about how the Implementation Standards should be interpreted—the rules against hate speech now included a loophole: "We allow content that excludes a group of people who share a protected characteristic from entering a country or continent." This was followed by three examples of the kind of speech that was now permissible. The first was "We should ban Syrians from coming into Germany." The next two examples—"I am calling for a total and complete shutdown of Muslims entering the United States" and "We should build a wall to keep Mexicans out of the country"—had been uttered, more or less word for word, by the President of the United States.

In May, 2017, shortly after Facebook released a report acknowledging that "malicious actors" from around the world had used the platform to meddle in the American Presidential election, Zuckerberg announced that the company would increase its global moderation workforce by two-thirds. Mildka Gray, who was then a contract worker for Facebook in Dublin, was moved into content moderation around this time; her husband, Chris, applied and was offered a job almost immediately. "They were just hiring anybody," he said. Mildka, Chris, and the other contractors were confined to a relatively drab part of Facebook's Dublin offices. Some of them were under the impression that, should they pass a Facebook employee in the hall, they were to stay silent.

For the first few days after content moderators are hired, a trainer guides them through the Implementation Standards, the Known Questions, and other materials. “The documents are full of technical jargon, not presented in any logical order,” Chris Gray recalled. “I'm looking around, going, Most of the people in this room do not speak English as a first language. How in the hell is this supposed to work?” Mildka, who is from Indonesia and whose first language is Bahasa Indonesia, agreed: “In the training room, you just nod, Yes, yes. Then you walk out of the room and ask your friend, ‘Did you understand? Can you explain it in our language?’ ” Unlike Facebook's earliest moderators, who were told to use their discretion and moral intuition, the Grays were often encouraged to ignore the context in which an utterance was made. The Implementation Standards stated that Facebook was “inclined to tolerate content, and refrain from adding friction to the process of sharing unless it achieves a direct and specific good.”

There is a logic to the argument that moderators should not be allowed to use too much individual discretion. As Chris Gray put it, “You don't want people going rogue, marking pictures as porn because someone is wearing a skirt above the knee or something.” Nor would it make sense to have Raphael's paintings of cherubs scrubbed from Facebook for violating child-nudity guidelines. “At the same time,” he went on, “there's got to be a balance between giving your moderators too much freedom and just asking them to turn their brains off.”

Mildka and Chris Gray left Facebook in 2018. Shortly afterward, in the U.K., Channel 4 aired a documentary that had been filmed by an undercover reporter posing as a content moderator in their office. At one point in the documentary, a trainer gives a slideshow presentation about how to interpret some of the Implementation Standards regarding hate speech. One slide shows an apparently popular meme: a Norman Rockwell-style image of a white mother who seems to be drowning her daughter in a bathtub, with the caption “When your daughter's first crush is a little Negro boy.” Although the image “implies a lot,” the trainer says, “there's no attack, actually, on the Negro boy . . . so we should ignore this.” There's a brief pause in the conference room. “Is everyone O.K. with that?” the trainer says.

"No, not O.K.," a moderator responds. The other moderators laugh uneasily, and the scene ends.

After the footage became public, a Facebook spokesperson claimed that the trainer had made a mistake. "I know for a fact that that's a lie," Chris Gray told me. "When I was there, I got multiple tickets with that exact meme in it, and I was always told to ignore. You go, 'C'mon, we all know exactly what this means,' but you're told, 'Don't make your own judgments.' "

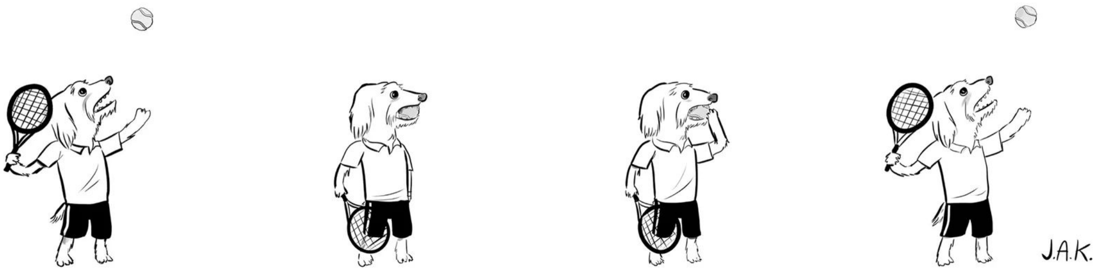
A former moderator from Phoenix told me, "If it was what they say it is—'You're here to clean up this platform so everyone else can use it safely'—then there's some nobility in that. But, when you start, you immediately realize we're in no way expected or equipped to fix the problem." He provided me with dozens of examples of hate speech—some of which require a good amount of cultural fluency to decode, others as clear-cut as open praise for Hitler—that he says were reviewed by moderators but not removed, "either because they could not understand why it was hateful, or because they assumed that the best way to stay out of trouble with their bosses was to leave borderline stuff up."

Recently, I talked to two current moderators, who asked me to call them Kate and John. They live and work in a non-Anglophone European country; both spoke in accented, erudite English. "If you're Mark Zuckerberg, then I'm sure applying one minimal set of standards everywhere in the world seems like a form of universalism," Kate said. "To me, it seems like a kind of libertarian imperialism, especially if there's no way for the standards to be strengthened, no matter how many people complain."

They listed several ways in which, in their opinion, their supervisors' interpretations of the Implementation Standards conflicted with common sense and basic decency. "I just reviewed an Instagram profile with the username KillAllFags, and the profile pic was a rainbow flag being crossed out," Kate said. "The implied threat is pretty clear, I think, but I couldn't take it down."

"Our supervisors insist that L.G.B.T. is a concept," John explained.

"So if I see someone posting 'Kill L.G.B.T.,' unless they refer to a person or use pronouns, I have to assume they're talking about killing an idea," Kate said.



Cartoon by Jason Adam Katzenstein

Open cartoon gallery

"Facebook could change that rule tomorrow, and a lot of people's lives would improve, but they refuse," John said.

"Why?" I said.

"We can ask, but our questions have no impact," John said. "We just do what they say, or we leave."

Around the time the Grays were hired, Britain First, a white-nationalist political party in the U.K., had a Facebook page with about two million followers. (By contrast, Theresa May, then the Prime Minister, had fewer than five hundred thousand followers.) Offline, Britain First engaged in scare tactics: driving around Muslim parts of London in combat jeeps, barging into mosques wearing green paramilitary-style uniforms. On Facebook, Chris Gray said, members of Britain First would sometimes post videos featuring "a bunch of thugs moving through London, going, 'Look, there's a halal butcher. There's a mosque. We need to reclaim our streets.' " A moderator who was empowered to consider the context—the fact that "Britain First" echoes "America First," a slogan once used by Nazi sympathizers in the U.S.; the ominous connotation of the word "reclaim"—could have made the judgment that the Party members' words and actions, taken together, were a call for violence. "But you're not allowed to look at the context," Gray said. "You can only look at what's right in front of you." Britain First's posts, he said, were "constantly getting reported, but the posts that ended up in my queue never quite went over the line to where I could delete them. The wording would always be just vague enough."

Tommy Robinson, a British Islamophobe and one of Britain First's most abrasive allies, often gave interviews in which he was open about his agenda. "It's a Muslim invasion of Europe," he told *Newsweek*. On Facebook, though, he was apparently more coy, avoiding explicit "calls for exclusion" and other formulations that the company would recognize as hate speech. At times, Gray had the uncanny sense that he and the other moderators were acting as unwitting coaches, showing the purveyors of hate speech just how far they could go. "That's what I'd do, anyway, if I were them," he said. "Learn to color within the lines." When Robinson or a Britain First representative posted something unmistakably threatening, a Facebook moderator would often flag the post for removal. Sometimes a "quality auditor" would reverse the decision. The moderator would then see a deduction in his or her "quality score," which had to remain at ninety-eight per cent or above for the moderator to be in good standing.

Normally, after a Facebook page violates the rules multiple times, the page is banned. But, in the case of Britain First and Tommy Robinson, the bans never came. Apparently, those two pages were "shielded," which meant that the power to delete them was restricted to Facebook's headquarters in Menlo Park. No one explained to the moderators why Facebook decided to shield some pages and not others, but, in practice, the shielded pages tended to be those with sizable follower counts, or with significant cultural or political clout—pages whose removal might interrupt a meaningful flow of revenue.

There is little recourse for a content moderator who has qualms about the Implementation Standards. Full-time Facebook employees are given more dispensation to question almost any aspect of company policy, as long as they do so internally. On Workplace, a custom version of the network that only Facebook staffers can access, their disagreements are often candid, even confrontational. The former Dublin employee who worked on Middle East policy believes that Facebook’s management tolerates internal dissent in order to keep it from spilling into public view: “Your average tech bro—Todd in Menlo Park, or whatever—has to continually be made to feel like he’s part of a force for good. So whenever Todd notices anything about Facebook that he finds disturbing there has to be some way for his critiques to be heard. Whether anything actually changes as a result of those critiques is a separate question.”

On December 18, 2017, on a Workplace message board called Community Standards Feedback, a recruiter in Facebook’s London office posted a *Guardian* article about Britain First. “They are pretty much a hate group,” he wrote. He noted that “today YouTube and Twitter banned them,” and asked whether Facebook would do the same.

Neil Potts, Facebook’s public-policy director for trust and safety, responded, “Thanks for flagging, and we are monitoring this situation closely.” However, he continued, “while Britain First shares many of the common tenets of alt-right groups, e.g., ultra-nationalism,” Facebook did not consider it a hate organization. “We define hate orgs as those that advance hatred as one of their primary objectives, or that they have leaders who have been convicted of hate-related offenses.”

Another Facebook employee, a Muslim woman, noted that Jayda Fransen, a leader of Britain First, had been convicted of hate crimes against British Muslims. “If the situation is being monitored closely,” she asked, “how was this missed?”

"Thanks for flagging," Potts responded. "I'll make sure our hate org SMEs"—subject-matter experts—"are aware of this conviction."

A month later, in January of 2018, the female employee revived the Workplace thread. "Happy new year!" she wrote. "The Britain First account is still up and running, even though as per above discussion it clearly violates our community standards. Is anything being done about this?"

"Thanks for circling back," Potts responded, adding that a "team is monitoring and evaluating the situation and discussing next steps forward." After that, the thread went dormant.

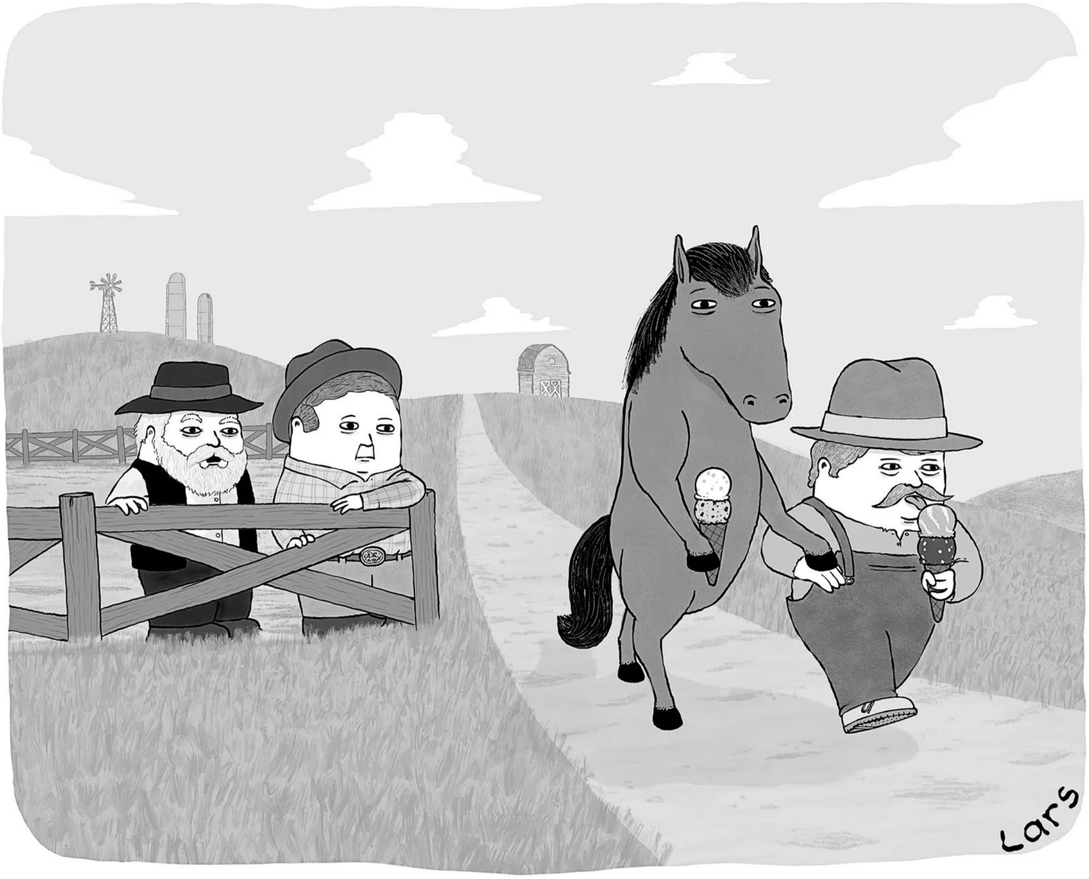
A few weeks later, Darren Osborne, a white Briton, was convicted of murder. Osborne had driven a van into a crowd near a London mosque, killing a Muslim man named Makram Ali and injuring at least nine other people. Prosecutors introduced evidence suggesting that Osborne had been inspired to kill, at least in part, by a BBC miniseries and by following Britain First and Tommy Robinson on social media. The judge deemed the killing "a terrorist act" by a man who'd been "rapidly radicalized over the Internet." Within six weeks, Britain First and Tommy Robinson had been banned from Facebook. (Pusateri, the Facebook spokesperson, noted that the company has "banned more than 250 white supremacist organizations.")

"It's an open secret," Sophie Zhang, a former data scientist for the company, recently wrote, "that Facebook's short-term decisions are largely motivated by PR and the potential for negative attention." Zhang left Facebook in September. Before she did, she posted a scathing memo on Workplace. In the memo, which was obtained by BuzzFeed News, she alleged that she had witnessed "multiple blatant attempts by foreign national governments to abuse our platform on vast scales"; in some cases, however, "we simply didn't care enough to stop them." She suggested that this was because the abuses were occurring in countries that American news outlets were unlikely to cover.

When Facebook is receiving an unusual amount of bad press for a particularly egregious piece of content, this is referred to within the company as a "press fire," or a "#PRFire." Often, the content has been flagged repeatedly, to no avail, but, in the context of a press fire, it receives prompt attention. A Facebook moderator who currently works in a European city shared with me a full record of the internal software system as it appeared on a recent day. There were dozens of press fires in progress. Facebook was being criticized—by Facebook users, primarily—for allowing widespread bullying against Greta Thunberg, the teen-age climate activist, who has Asperger's syndrome. The content moderators were instructed to apply an ad-hoc exemption: "Remove all instances of attacks aimed at Greta Thunberg using the terms or hashtag: 'Gretarded', 'Retard' or 'Retarded.' " No similar protections were extended to other young activists, including those whose bullying was unlikely to inspire such a public backlash. A woman who worked as a content-moderation supervisor in the U.S. told me, "You can ask for a meeting, present your bosses with bullet points of evidence, tell them you've got team members who are depressed and suicidal—doesn't help. Pretty much the only language Facebook understands is public embarrassment."

Facebook moderators have scant workplace protections and little job security. The closest they have to a labor organizer is Cori Crider, a lawyer and an activist based in London. Crider grew up in rural Texas and left as soon as she could, going first to Austin, for college; then to Harvard, for law school; and then to London, where she worked for a decade at a small human-rights organization, representing Guantánamo detainees and the relatives of drone-strike victims in Yemen and Pakistan. "It was through worrying about drones that I came to worry about technology," she said. "I started to feel like, While we're all focussed on the surveillance tactics of the Pentagon, a handful of companies out of California are collecting data on a scale that would honestly be the envy of any state."

Last year, she co-founded a not-for-profit called Foxglove, where she is one of two employees. The foxglove, a wildflower also known as digitalis, can be either medicinal or toxic to humans, depending on how it's ingested. The group's mission is to empower the tech industry's most vulnerable workers—to help them "clean up their factory floor," as Crider often puts it. Her more ambitious goal is to redesign the factory. She was influenced by Shoshana Zuboff, the author of "The Age of Surveillance Capitalism," who argues that "instrumentarian" behemoths such as Facebook pose an existential threat to democracy. In Crider's analysis, not even the most ingenious technocratic fix to Facebook's guidelines can address the core problem: its content-moderation priorities won't change until its algorithms stop amplifying whatever content is most enthralling or emotionally manipulative. This might require a new business model, perhaps even a less profitable one, which is why Crider isn't hopeful that it will happen voluntarily. "I wish I could file a global injunction against the monetization of attention," she said. "In the meantime, you find more specific ways to create pressure."



“Can’t say I agree with his methods, but I’ll be damned if there’s a horse that man can’t break.”

Cartoon by Lars Kenseth

Open cartoon gallery

In July, 2019, Crider was introduced to a Facebook moderator in Europe who was able to map out how the whole system worked. This moderator put her in touch with other moderators, who put her in touch with still others. Sometimes, while courting a moderator as a potential source, Crider arranged a meeting and flew to wherever the person lived, only to be stood up. Those who did talk to her were almost always unwilling to go on the record. The process reminded Crider of the months she’d spent in Yemen and Pakistan, trying to gain people’s trust. “They often have very little reason to talk, and every reason in the world not to,” she said. The content moderators were not yet ready to form a union—“not even close,” Crider told me—but she hoped to inculcate in them a kind of latent class consciousness, an awareness of themselves as a collective workforce.

Last October, Crider met Chris Gray at a conference in London. She started introducing him to journalists and activists, helping to spread his story. Two months later, Gray hired a local law firm and sued Facebook in Irish High Court, alleging that his "repeated and unrelenting exposure to extremely disturbing, graphic and violent content" had caused him lasting psychological trauma. Shortly thereafter, about twenty more former Facebook moderators in Dublin contacted the law firm representing Gray to ask about possible lawsuits against the company.

Soon after Gray left Facebook, he wrote a fifty-three-hundred-word memo to Zuckerberg and Sheryl Sandberg, the company's chief operating officer, laying out his critiques of the content-moderation process. He proposed a few fixes, which ranged from the granular ("The hotkey system is a mess") to the sweeping ("Rewrite . . . all of your user privacy policies"). None, however, addressed what he considered to be the underlying issue. On the whole, he concluded, Facebook "is not committed to content moderation, does not have a clear strategy or even a good handle on how to do it, and the people trying to do the actual work are under immense pressure to shovel shit uphill without proper tools or direction. . . . There is no leadership, no clear moral compass." He e-mailed the memo to Zuckerberg and Sandberg from an anonymous address: shovellingshituphill@gmail.com.

Jair Bolsonaro, the autocratic Brazilian politician, ran for President, in 2018, on a shoestring budget. To get his message out, he relied heavily on Facebook, whose social-media apps are the most popular in his country. Since taking office, in 2019, he has delivered a weekly Presidential address on Facebook Live. Earlier this year, during one speech, he said of Brazil's indigenous population, "The Indian has changed. He is evolving and becoming, more and more, a human being like us." Bolsonaro's racism was not exactly a surprise, but his comments caused an uproar nonetheless. Facebook did not remove the video, even though its guidelines prohibit "dehumanizing speech," including any insinuations of "subhumanity."

David Thiel, a cybersecurity expert at Facebook's headquarters in Menlo Park, read about the controversy. (He is unrelated to Peter Thiel, the venture capitalist and Trump donor who sits on Facebook's board.) After searching for the speech on Facebook, he was shocked to find that it was still up. He wrote on Workplace, "I assume we'll be removing this video from our platform?" His question was passed on to subject-matter experts, including one in Brasília and one in Dublin. Both ruled that the video did not violate the guidelines. "President Bolsonaro is known by his controversial and 'politically incorrect' speeches," the expert in Brasília wrote. "He is actually referring to indigenous people becoming more integrated to the society (as opposed to isolated in their own tribes)." This did not satisfy Thiel, in part because the expert, who has worked for at least one pro-Bolsonaro politician, did not strike him as an objective source. Also, given that Facebook's local sales representatives had surely encouraged Bolsonaro to use their products, there may have been a conflict of interest. "It's awkward for a business to go from a posture of 'Please, sir, use our product,' to 'Actually, sir, you're now in trouble for using our product,' " he said.

Thiel appealed the decision, and four or five members of the content-policy team agreed to meet with him by video conference. To make his case that "becoming a human being" was dehumanizing speech, Thiel, with the help of some of his colleagues, created a fifteen-slide PowerPoint presentation, parsing the utterance with a computer engineer's minute attention to detail. One slide included a Merriam-Webster definition of the word "become," and added, "To 'become' something necessarily denotes that, in the status quo ante, the subject is currently not that thing." Thiel also argued that Bolsonaro's racist rhetoric had already incited violence. (In 2019, Bolsonaro's first year in office, seven Brazilian tribal leaders were murdered, the highest number in twenty years.) Thiel's penultimate slide featured a rousing quotation from Zuckerberg's Georgetown speech: "We know from history that dehumanizing people is the first step towards inciting violence. . . . I take this incredibly seriously, and we work hard to get this off our platform." As Thiel delivered his presentation, he recalled, various members of the policy team "interrupted a lot, pushing back on my reasoning or questioning my credibility." When that didn't work, "they just kept insisting, in an up-is-down, black-is-white kind of way, that the words didn't violate the policy." A former content moderator told me, "At some point, someone at Facebook could have said, 'We will keep making exceptions whenever politicians break our rules.' But they never wanted to admit that, even to themselves, so instead they arrived at this twisted logical place where they are now able to look at something that is clearly a violation of their own rules and go, 'Nope, no violation here.' "

In March, Thiel announced his resignation from Facebook. "It was a pretty overt rage quit," he told me. He posted a long, impassioned note on Workplace. "Facebook right now is increasingly aligning with the rich and powerful, allowing them to play by different rules," he wrote, adding that "the hard-right turn has been disillusioning and is not something I feel comfortable with anymore." Shortly after he posted his goodbye note, the content-policy team wrote to him to say that they'd reversed their decision about Bolsonaro's speech. "I couldn't tell if it was them trying to get me to not leave, or to leave on better terms, or what," he said. "Either way, it was too late."

Last October, the Trump campaign made an ad featuring blatantly false allegations about Joe Biden. CNN and other networks refused to run it; YouTube, Twitter, and Facebook did not. "Our approach is grounded in Facebook's fundamental belief in free expression," Katie Harbath, the company's public-policy director for global elections and a former digital strategist for Rudolph Giuliani's Presidential campaign, wrote. "Thus, when a politician speaks or makes an ad, we do not send it to third party fact-checkers." Later that month, in a congressional hearing, Representative Alexandria Ocasio-Cortez asked Zuckerberg how far this policy went: could any politician lie about anything on his platform? Zuckerberg responded, under oath, that he did have some red lines. "If anyone, including a politician, is saying things that is calling for violence . . . or voter or census suppression," Zuckerberg said, "we will take that content down."

Seven months later, Trump crossed these two red lines within a matter of days. On May 26th, on both Twitter and Facebook, he wrote, falsely, "There is NO WAY (ZERO!) that Mail-In Ballots will be anything less than substantially fraudulent." This seemed like an obvious attempt at voter suppression: why would the President warn the public about the putative inadequacy of the vote-by-mail system if not to dissuade people from using it? On May 29th, again using the same language on both Twitter and Facebook, Trump mused about sending the National Guard to quell protests in response to George Floyd's death. "When the looting starts, the shooting starts," Trump wrote, a phrase that was widely seen as an incitement to violence. Prominent segregationists had used these words, in the nineteen-sixties, to justify vicious attacks against Black people, including civil-rights protesters.

Twitter didn't remove Trump's tweets but did append warning labels to them. Facebook, by contrast, did nothing. "That was the moment when a lot of us snapped," Rashad Robinson, of Color of Change, told me. Vanita Gupta, the president and C.E.O. of the Leadership Conference on Civil and Human Rights, said, "The feeling among activists was Why have we spent years pushing Facebook to adopt better policies if they're just going to ignore those policies when they matter most?" (Some Trump campaign ads—including one from June, which used a symbol associated with the Nazis, and one from September, which baselessly accused refugees of spreading the coronavirus—have since been removed.)

On June 1st, scores of Facebook employees, who were working from home due to the pandemic, staged a virtual walkout. Two days later, thirty-four of Facebook's earliest employees, including Dave Willner, signed an open letter that was published in the *Times*. "If all speech by politicians is newsworthy and all newsworthy speech is inviolable," it read, "then there is no line the most powerful people in the world cannot cross on the largest platform in the world." Cori Crider is now in close contact with some fifty content moderators, and she encouraged them to publish a letter of their own. On June 8th, a group of ten current and former moderators, including Chris Gray, signed an open letter on the publishing platform Medium. "As outsourced contractors, non-disclosure agreements deter us from speaking openly," the letter read. Nonetheless, "current events prove we cannot passively accept our role of silent algorithm facilitators—not when our screens are being flooded with hate speech."

On August 19th, Facebook announced changes to its guidelines. Chief among them was a new policy restricting the activities of "organizations and movements that have demonstrated significant risks to public safety," including "US-based militia organizations." Some reporters and activists asked why it had taken so long for Facebook to come up with rules regarding such groups; others pointed out that, although hundreds of pages had been removed under the new policy, many such pages remained. Four days later, in Kenosha, Wisconsin, a police officer shot a Black man named Jacob Blake seven times in the back, in front of his children. Nightly protests erupted. The Kenosha Guard, a self-described militia, put up a "call to arms" on its Facebook page, where people explicitly expressed their intention to commit vigilante violence ("I fully plan to kill looters and rioters tonight"). Within a day, according to BuzzFeed, more than four hundred people had reported the page to Facebook's content moderators, but the moderators decided that it did not violate any of Facebook's standards, and they left it up. (Mark Zuckerberg later called this "an operational mistake.") On August 25th, a white seventeen-year-old travelled to Kenosha from out of state, carrying a semi-automatic rifle, and shot three protesters, killing two of them. It's not clear whether he'd learned about the Kenosha Guard on Facebook, but the militia's page was public. Anyone could have seen it.

Pusateri, the Facebook spokesperson, said, "So far we've identified over 300 militarized social movements who we've banned from maintaining Facebook Pages, groups, and Instagram accounts." In addition, last week, Facebook banned all content relating to QAnon, the far-right conspiracy theory. It also took down a post by Trump that contained misinformation about the coronavirus, and announced plans to ban all political ads from the platform for an indefinite period starting on Election Night. Its critics once again considered these measures too little, too late. Senator Elizabeth Warren described them as "performative changes," arguing that the company was still failing to "change its broken algorithm, or take responsibility for the power it's amassed."

The restrictions are also likely to feed into the notion that social media discriminates against conservatives. (As Trump tweeted in May, "The Radical Left is in total command & control of Facebook, Instagram, Twitter and Google.") This has become a right-wing talking point, even though the bulk of the evidence suggests the opposite. Every weekday, the *Times* reporter Kevin Roose shares the Top Ten "link posts"—posts containing links—from American Facebook pages, according to data provided by a tool owned by Facebook. Almost always, the list is dominated by far-right celebrities or news outlets. (On a representative day last week, the Top Ten included a post by Donald Trump for President, four posts from Fox News, two from CNN, and one from TMZ.) Facebook has disputed Roose's methodology, arguing that there are ways to parse the data that would make it look less damning. Roose ranks posts by "interactions," but John Hegeman, who runs Facebook's News Feed, has argued that it would be better to rank posts by "reach" instead. This, Hegeman tweeted in July, would be "a more accurate way to see what's popular." However, he continued, "This data is only available internally." ♦

Published in the print edition of the October 19, 2020, issue, with the headline "Explicit Content."

Andrew Marantz is a staff writer at The New Yorker and the author of "Antisocial: Online Extremists, Techno-Utopians, and the Hijacking of the American Conversation."

More: Facebook Politics Donald Trump Internet Social Media Technology Hate Speech Fake News 2020 Election Mark Zuckerberg Jair Bolsonaro Peter Thiel England 2016 Election

Read More