

Inside Facebook's Secret Rulebook for Global Political Speech

Under fire for stirring up distrust and violence, the social network has vowed to police its users. But leaked documents raise serious questions about its approach.

By Max Fisher

Dec. 27, 2018

MENLO PARK, Calif. — In a glass conference room at its California headquarters, Facebook is taking on the bonfires of hate and misinformation it has helped fuel across the world, one post at a time.

The social network has drawn criticism for undermining democracy and for provoking bloodshed in societies small and large.

But for Facebook, it's also a business problem.

The company, which makes about \$5 billion in profit per quarter, has to show that it is serious about removing dangerous content. It must also continue to attract more users from more countries and try to keep them on the site longer.

How can Facebook monitor billions of posts per day in over 100 languages, all without disturbing the endless expansion that is core to its business? The company's solution: a network of workers using a maze of PowerPoint slides spelling out what's forbidden.

Every other Tuesday morning, several dozen Facebook employees gather over breakfast to come up with the rules, hashing out what the site's two billion users should be allowed to say. The guidelines that emerge from these meetings are sent out to 7,500-plus moderators around the world. (After publication of this article, Facebook said it had increased that number to around 15,000.)

The closely held rules are extensive, and they make the company a far more powerful arbiter of global speech than has been publicly recognized or acknowledged by the company itself, The New York Times has found.

The Times was provided with more than 1,400 pages from the rulebooks by an employee who said he feared that the company was exercising too much power, with too little oversight — and making too many mistakes.

An examination of the files revealed numerous gaps, biases and outright errors. As Facebook employees grope for the right answers, they have allowed extremist language to flourish in some countries while censoring mainstream speech in others.

Moderators were once told, for example, to remove fund-raising appeals for volcano victims in Indonesia because a co-sponsor of the drive was on Facebook's internal list of banned groups. In Myanmar, a paperwork error allowed a prominent extremist group, accused of fomenting genocide, to stay on the platform for months. In India, moderators were mistakenly told to flag for possible removal comments critical of religion.



The ruins of a home set upon by a Buddhist mob in a deadly attack in Sri Lanka last March. Facebook has been accused of accelerating violence in the country. Adam Dean for The New York Times

The Facebook employees who meet to set the guidelines, mostly young engineers and lawyers, try to distill highly complex issues into simple yes-or-no rules. Then the company outsources much of the actual post-by-post moderation to companies that enlist largely unskilled workers, many hired out of call centers.

Those moderators, at times relying on Google Translate, have mere seconds to recall countless rules and apply them to the hundreds of posts that dash across their screens each day. When is a reference to “jihad,” for example, forbidden? When is a “crying laughter” emoji a warning sign?

Moderators express frustration at rules they say don’t always make sense and sometimes require them to leave up posts they fear could lead to violence. “You feel like you killed someone by not acting,” one said, speaking on the condition of anonymity because he had signed a nondisclosure agreement.

Facebook executives say they are working diligently to rid the platform of dangerous posts.

“It’s not our place to correct people’s speech, but we do want to enforce our community standards on our platform,” said Sara Su, a senior engineer on the News Feed. “When you’re in our community, we want to make sure that we’re balancing freedom of expression and safety.”

Monika Bickert, Facebook’s head of global policy management, said that the primary goal was to prevent harm, and that to a great extent, the company had been successful. But perfection, she said, is not possible.

“We have billions of posts every day, we’re identifying more and more potential violations using our technical systems,” Ms. Bickert said. “At that scale, even if you’re 99 percent accurate, you’re going to have a lot of mistakes.”

The Rules

What does “possible terrorism terminology” mean?

A neutral term which is not praise, support or representation of terrorism

AND either:

1. Is used along with a photo / video:

- I. Of a fighter (unofficial military style clothing which cannot be tied back to an official army + weapons)
- II. Of a suicide bomber or attacker
- III. Of a terrorist leader
- IV. Of a logo / symbol / flag / animation

OR

2. Title of a page/group which represents terrorism
E.g: Page title = “Amaq Agency” (ISIS’ media wing) “Free Syrian Army” “Indian Mujahideen” “Loving PKK”

When is it support for terrorism? Is “martyr” a forbidden word? Moderators are given guides to help them decide.

The Facebook guidelines do not look like a handbook for regulating global politics. They consist of dozens of unorganized PowerPoint presentations and Excel spreadsheets with bureaucratic titles like “Western Balkans Hate Orgs and Figures” and “Credible Violence: Implementation standards.”

Because Facebook drifted into this approach somewhat by accident, there is no single master file or overarching guide, just a patchwork of rules set out by different parts of the company. Facebook confirmed the authenticity of the documents, though it said some had been updated since The Times acquired them.

The company’s goal is ambitious: to reduce context-heavy questions that even legal experts might struggle with — when is an idea hateful, when is a rumor dangerous — to one-size-fits-all rules. By telling moderators to follow the rules blindly, Facebook hopes to guard against bias and to enforce consistency.

Hate Speech			
Attack Summary			
Protected Characteristics	Tier 1	Tier 2	Tier 3
	PCs + QPCs + Sensitive Subsets + Non-Sensitive Subsets + Age Subsets of PC + QPC Subsets	PCs + Sensitive Subsets + Age Subsets of PC	PCs + Sensitive Subsets + Age Subsets of PC
Attacks	- Call for violence, disease or death - Designated dehumanizing comparison - Dehumanizing claims - Mocking Hate Crimes - Visual hate	- Making claims about physical, mental or moral deficiencies. - Expressing contempt - Cursing	- Call for exclusion of segregation

A slide from Facebook’s rulebook on what constitutes hate speech asks moderators to quickly make a series of complex, legalistic judgments per post.

Taken individually, each rule might make sense. But in their byzantine totality, they can be a bit baffling.

Emojis

Indicators of..	Emojis
Condemnation	🙄, 😬, ❤️, 🍌, 🤔, 😏, 👎, 🧐, 🗿, 🚮, 🦹‍♂️, 🦹‍♀️
Praise, Support, Promote	😊, 😊, 👉, ❤️, 💯, ✈️, 🥳, 🙌, 👏, 📢, 🇩🇪, 🙋, 🥰, 😁
Bullying, Mocking	😝, 😜, 🟡
Sexualised text	💀, 🍆, 🍑
Attack, Harm, Call to Action	👊, 🔫, ⚒️, 🛠️, 🖱️
SSI	⚔️, 🔪
Sexual orientation	🌈, 🏳️, 🧺, 🏴, 🏵️, 🏶, 🏷️, 🏸️, 🏹️, 🏼, 🏽, 🏾, ♀, ♂
Exclusion	🚫, 🚲, 🚭, ❌, 🚨, 🚻, 🚣
Dehumanising comparison	🐼, 🐕, 🐘, 🐃, 🐅, 🐆, 🐇, 🐈, 🐉, 🐊, 🐋, 🐌, 🐍, 🐎, 🐏, 🐑, 🐓, 🐔, 🐕, 🐖, 🐗, 🐘, 🐙, 🐛, 🐞, 🐟, 🐠, 🐡, 🐢, 🐣, 🐤, 🐥, 🐦, 🐧, 🐨, 🐩, 🐪, 🐫, 🐬, 🐭, 🐮, 🐯, 🐰, 🐹, 🐺, 🐻, 🐼

The guidelines for identifying hate speech, a problem that has bedeviled Facebook, run to 200 jargon-filled, head-spinning pages. Moderators must sort a post into one of three “tiers” of severity. They must bear in mind lists like the six “designated dehumanizing comparisons,” among them comparing Jews to rats.

Though the Facebook employees who make the rules are largely free to set policy however they wish, and often do so in the room, they also consult with outside groups.

An Unseen Branch of Government

Key Parties

Jamiat Ulama-e-Islam - Fazal ur Rehman Group (JUI-F)

- Religious political party, "Sunni Deoband", based in Khyber Pakhtunkhwa & Balochistan
- In Coalition (14 seats in 2013-2018), JUI split into JUI-F & JUI-D due to dissent over policies, religious inclination, played role in peace discussions between military & militants
- Current Leader: Maulana Fazal ur Rehman
- JUI-F Leader: Sami ul Haq, more inclined towards imposing Sharia, allegedly support militant activity
- Expected Content: Hate speech (Sharia), Religious agenda, "Maulana Disease"




Fazal ur Rehman



Sami ul Haq

Maulana Diesel ref. Hate speech Shias. Both are religious party.

JUI-S which is lead by Sami Ul Haq has been known as pro Taliban, has been really rigid and

As detailed as the guidelines can be, they are also approximations — best guesses at how to fight extremism or disinformation. And they are leading Facebook to intrude into sensitive political matters the world over, sometimes clumsily.

Increasingly, the decisions on what posts should be barred amount to regulating political speech — and not just on the fringes. In many countries, extremism and the mainstream are blurring.

In the United States, Facebook banned the Proud Boys, a far-right pro-Trump group. The company also blocked an inflammatory ad, about a caravan of Central American migrants, that was produced by President Trump's political team.

In June, according to internal emails reviewed by The Times, moderators were told to allow users to praise the Taliban — normally a forbidden practice — if they mentioned its decision to enter into a cease-fire. In another email, moderators were told to hunt down and remove rumors wrongly accusing an Israeli soldier of killing a Palestinian medic.

"Facebook's role has become so hegemonic, so monopolistic, that it has become a force unto itself," said Jasmin Mujanovic, an expert on the Balkans. "No one entity, especially not a for-profit venture like Facebook, should have that kind of power to influence public debate and policy."

In Pakistan, shortly before elections were held in July, Facebook issued its moderators a 40-page document outlining "political parties, expected trends and guidelines."

Pakistan, one of the world's largest and most fragile democracies, enforces an Election Day blackout on campaigning. Facebook is a center of news and discussion during voting.

The document most likely shaped those conversations — even if Pakistanis themselves had no way of knowing it. Moderators were urged, in one instance, to apply extra scrutiny to Jamiat Ulema-e-Islam, a hard-line religious party. But another religious party, Jamaat-e-Islami, was described as "benign."

Though Facebook says its focus is protecting users, the documents suggest that other concerns come into play. Pakistan guidelines warn moderators against creating a "PR fire" by taking any action that could "have a negative impact on Facebook's reputation or even put the company at legal risk."

In India, Chinmayi Arun, a legal scholar, identified troubling mistakes in Facebook's guidelines.

One slide tells moderators that any post degrading an entire religion violates Indian law and should be flagged for removal. It is a significant curb on speech — and apparently incorrect. Indian law prohibits blasphemy only in certain conditions, Ms. Arun said, such as when the speaker intends to inflame violence.



Facebook's rules for India and Pakistan both include this diagram explaining that the company removes some content to avoid risk of legal challenge or being blocked by governments.

Another slide says that Indian law prohibits calls for an independent Kashmir, which some legal scholars dispute. The slide instructs moderators to "look out for" the phrase "Free Kashmir" — though the slogan, common among activists, is completely legal.

Facebook says it is simply urging moderators to apply extra scrutiny to posts that use the phrase. Still, even this could chill activism in Kashmir. And it is not clear that the distinction will be obvious to moderators, who are warned that ignoring violations could get Facebook blocked in India.

'Things Explode Really Fast'

In the absence of governments or international bodies that can set standards, Facebook is experimenting on its own.

The company never set out to play this role, but in an effort to control problems of its own creation, it has quietly become, with a speed that makes even employees uncomfortable, what is arguably one of the world's most powerful political regulators.

"A lot of this would be a lot easier if there were authoritative third parties that had the answer," said Brian Fishman, a counterterrorism expert who works with Facebook.

"Sometimes these things explode really fast," Mr. Fishman said, "and we have to figure out what our reaction's going to be, and we don't have time for the U.N."

But the results can be uneven.

Consider the guidelines for the Balkans, where rising nationalism is threatening to reignite old violence. The file on that region, not updated since 2016, includes odd errors. Ratko Mladic, a Bosnian war criminal still celebrated by extremists, is described as a fugitive. In fact, he was arrested in 2011.



A 2016 document on Western Balkan hate groups, still in use, incorrectly describes Ratko Mladic as a fugitive. Mr. Mladic was arrested in 2011. Though the error is minor, experts say it underscores an inattention to detail in Facebook's guidelines.

The slides are apparently written for English speakers relying on Google Translate, suggesting that Facebook remains short on moderators who speak local languages — and who might understand local contexts crucial for identifying inflammatory speech. And Google Translate can be unreliable: Mr. Mladic is referred to in one slide as “Rodney Young.”

The guidelines, said Mr. Mujanovic, the Balkans expert, appear dangerously out of date. They have little to say about ultranationalist groups stoking political violence in the region.

Nearly every Facebook employee who spoke to The Times cited, as proof of the company's competence, its response after the United Nations accused the platform of exacerbating genocide in Myanmar. The employees pointed to Facebook's ban this spring on any positive mention of Ma Ba Tha, an extremist group that has been using the platform to incite violence against Muslims since 2014.

But puzzled activists in Myanmar say that, months later, posts supporting the group remain widespread.

Understanding the Coup in Myanmar

A military coup. Following a military coup on Feb. 1, 2021, unrest gripped Myanmar. Peaceful pro-democracy demonstrations gave way to insurgent uprisings against the Tatmadaw, the country's military, which ousted the country's civilian leader, Daw Aung San Suu Kyi.

The culprit may be Facebook's own rulebooks. Guidelines for policing hate speech in Myanmar instruct moderators not to remove posts supporting Ma Ba Tha. Facebook corrected the mistake only in response to an inquiry from The Times.



The page on the left is NOT violating as it clearly condemns the hate figure. The page on the right is NOT violating even though the "Ma Ba Tha" logo indicates association with Wirathu. Currently, we only take consideration of Wirathu (hate figure) himself and not the organizations associated with Wirathu (hate figure).

Several months after Facebook said it had banned praise for Ma Ba Tha, a Myanmar supremacist group accused of encouraging ethnic cleansing, the company’s Myanmar guidelines stated that the group was allowed.

Employees also touted their decision to shut down Facebook accounts belonging to senior military officials in Myanmar.

But the company did not initially notify Myanmar’s government, leading the barred officers to conclude that they had been hacked. Some blamed Daw Aung San Suu Kyi, the country’s de facto civilian leader, and the episode deepened distrust between her and the military, lawmakers say.

The Hate List

Facebook’s most politically consequential document may be an Excel spreadsheet that names every group and individual the company has quietly barred as a hate figure.

COUNTRY	NAME	ALTERNATIVE NAME	SYMBOLS / SLOGANS	SYMBOLS / SLOGANS	SYMBOLS / SLOGANS
Indonesia	Front Jihad Islam	FJI			
Indonesia	Laskar Pembela Islam - LPI	Soldiers of the Defenders of Islam			
International	International Gaym Party	IGP			
Israel	Lehava	למנוחת הנבחרות בארץ הקודש			
Italy	Comunità Militante Dei Diodici Raggi, Do.Ra	Do.Ra			
Italy	Lealtà - Azione				
Italy	MAB Manipolo D'avanguardia Bergamo	MAB			

Facebook keeps an internal list of groups and individuals it bars as hate figures, though not all are on the fringe. Facebook users are prohibited from posting content that is deemed to support or praise them.

Moderators are instructed to remove any post praising, supporting or representing any listed figure.

Anton Shekhovtsov, an expert in far-right groups, said he was “confused about the methodology.” The company bans an impressive array of American and British groups, he said, but relatively few in countries where the far right can be more violent, particularly Russia or Ukraine.

Countries where Facebook faces government pressure seem to be better covered than those where it does not. Facebook blocks dozens of far-right groups in Germany, where the authorities scrutinize the social network, but only one in neighboring Austria.

The list includes a growing number of groups with one foot in the political mainstream, like the far-right Golden Dawn, which holds seats in the Greek and European Union parliaments.

For a tech company to draw these lines is “extremely problematic,” said Jonas Kaiser, a Harvard University expert on online extremism. “It puts social networks in the position to make judgment calls that are traditionally the job of the courts.”

The bans are a kind of shortcut, said Sana Jaffrey, who studies Indonesian politics at the University of Chicago. Asking moderators to look for a banned name or logo is easier than asking them to make judgment calls about when political views are dangerous.

But that means that in much of Asia and the Middle East, Facebook bans hard-line religious groups that represent significant segments of society. Blanket prohibitions, Ms. Jaffrey said, amount to Facebook shutting down one side in national debates.

And its decisions often skew in favor of governments, which can fine or regulate Facebook.

In Sri Lanka, Facebook removed posts commemorating members of the Tamil minority who died in the country's civil war. Facebook bans any positive mention of Tamil rebels, though users can praise government forces who were also guilty of atrocities.

Kate Cronin-Furman, a Sri Lanka expert at University College London, said this prevented Tamils from memorializing the war, allowing the government to impose its version of events — entrenching Tamils' second-class status.

The View From Menlo Park

Facebook's policies might emerge from well-appointed conference rooms, but they are executed largely by moderators in drab outsourcing offices in distant locations like Morocco and the Philippines.

Facebook says moderators are given ample time to review posts and don't have quotas. Moderators say they face pressure to review about a thousand pieces of content per day. They have eight to 10 seconds for each post, longer for videos.

The moderators describe feeling in over their heads. For some, pay is tied to speed and accuracy. Many last only a few exhausting months. Front-line moderators have few mechanisms for alerting Facebook to new threats or holes in the rules — and little incentive to try, one said.

One moderator described an officewide rule to approve any post if no one on hand can read the appropriate language. This may have contributed to violence in Sri Lanka and Myanmar, where posts encouraging ethnic cleansing were routinely allowed to stay up.

Facebook says that any such practice would violate its rules, which include contingencies for reviewing posts in unfamiliar languages. Justin Osofsky, a Facebook vice president who oversees these contracts, said any corner-cutting probably came from midlevel managers at outside companies acting on their own.

This hints at a deeper problem. Facebook has little visibility into the giant outsourcing companies, which largely police themselves, and has at times struggled to control them. And because Facebook relies on the companies to support its expansion, its leverage over them is limited.

One hurdle to reining in inflammatory speech on Facebook may be Facebook itself. The platform relies on an algorithm that tends to promote the most provocative content, sometimes of the sort the company says it wants to suppress.

Facebook could blunt that algorithm or slow the company's expansion into new markets, where it has proved most disruptive. But the social network instills in employees an almost unquestioned faith in their product as a force for good.

When Ms. Su, the News Feed engineer, was asked if she believed research finding that more Facebook usage correlates with more violence, she replied, “I don't think so.”

“As we have greater reach, as we have more people engaging, that raises the stakes,” she said. “But I also think that there's greater opportunity for people to be exposed to new ideas.”

Still, even some executives hesitate when asked whether the company has found the right formula.

Richard Allan, a London-based vice president who is also a sitting member of the House of Lords, said a better model might be “some partnership arrangement” with “government involved in setting the standards,” even if not all governments can be trusted with this power.

Mr. Fishman, the Facebook terrorism expert, said the company should consider deferring more decisions to moderators, who may better understand the nuances of local culture and politics.

But at company headquarters, the most fundamental questions of all remain unanswered: What sorts of content lead directly to violence? When does the platform exacerbate social tensions?

Rosa Birch, who leads an internal crisis team, said she and her colleagues had been posing these questions for years. They are making progress, she said, but will probably never have definitive answers.

But without a full understanding of the platform's impact, most policies are just ad hoc responses to problems as they emerge. Employees make a tweak, wait to see what happens, then tweak again — as if repairing an airplane midflight.

In the meantime, the company continues to expand its reach to more users in more countries.

“One of the reasons why it's hard to talk about,” Mr. Fishman said, “is because there is a lack of societal agreement on where this sort of authority should lie.”

But, he said, “it's harder to figure out what a better alternative is.”