1. Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt.
By Jeff Horwitz

Mark Zuckerberg has publicly said Facebook Inc. allows its more than three billion users to speak on equal footing with the elites of politics, culture and journalism, and that its standards of behavior apply to everyone, no matter their status or fame.
In private, the company has built a system that has exempted high-profile users from some or all of its rules, according to company documents reviewed by The Wall Street Journal.

The program, known as "cross check" or "XCheck," was initially intended as a quality-control measure for actions taken against high-profile accounts, including celebrities, politicians and journalists. Today, it shields millions of VIP users from the company's normal enforcement process, the documents show. Some users are "whitelisted"—rendered immune from enforcement actions—while others are allowed to post rule-violating material pending Facebook employee reviews that often never come.

At times, the documents show, XCheck has protected public figures whose posts contain harassment or incitement to violence, violations that would typically lead to sanctions for regular users. In 2019, it allowed international soccer star Neymar to show nude photos of a woman, who had accused him of rape, to tens of millions of his fans before the content was removed by Facebook. Whitelisted accounts shared inflammatory claims that Facebook's fact checkers deemed false, including that vaccines are deadly, that Hillary Clinton had covered up "pedophile rings," and that then-President Donald Trump had called all refugees seeking asylum "animals," according to the documents.

A 2019 internal review of Facebook's whitelisting practices, marked attorney-client privileged, found favoritism to those users to be both widespread and "not publicly defensible."

"We are not actually doing what we say we do publicly," said the confidential review. It called the company's actions "a breach of trust" and added: "Unlike the rest of our community, these people can violate our standards without any consequences."

Despite attempts to rein it in, XCheck grew to include at least 5.8 million users in 2020, documents show. In its struggle to accurately moderate a torrent of content and avoid negative attention, Facebook created invisible elite tiers within the social network.

In describing the system, Facebook has misled the public and its own Oversight Board, a body that Facebook created to ensure the accountability of the company's enforcement systems.

In June, Facebook told the Oversight Board in writing that its system for high-profile users was used in "a small number of decisions."

In a written statement, Facebook spokesman Andy Stone said criticism of XCheck was fair, but added that the system "was designed for an important reason: to create an additional step so we can accurately enforce policies on content that could require more understanding."

He said Facebook has been accurate in its communications to the board and that the company is continuing to work to phase out the practice of whitelisting. “A lot of this internal material is outdated information stitched together to create a narrative that glosses over the most important point: Facebook itself identified the issues with cross check and has been working to address them,” he said.

Internal document

The documents that describe XCheck are part of an extensive array of internal Facebook communications reviewed by The Wall Street Journal. They show that Facebook knows, in acute detail, that its platforms are riddled with flaws that cause harm, often in ways only the company fully understands.

Moreover, the documents show, Facebook often lacks the will or the ability to address them.

This is the first in a series of articles based on those documents and on interviews with dozens of current and former employees.

At least some of the documents have been turned over to the Securities and Exchange Commission and to Congress by a person seeking federal whistleblower protection, according to people familiar with the matter.

Facebook’s stated ambition has long been to connect people. As it expanded over the past 17 years, from Harvard undergraduates to billions of global users, it struggled with the messy reality of bringing together disparate voices with different motivations—from people wishing each other happy birthday to Mexican drug cartels conducting business on the platform. Those problems increasingly consume the company.

Time and again, the documents show, in the U.S. and overseas, Facebook’s own researchers have identified the platform’s ill effects, in areas including teen mental health, political discourse and human trafficking. Time and again, despite congressional hearings, its own pledges and numerous media exposés, the company didn’t fix them.

Sometimes the company held back for fear of hurting its business. In other cases, Facebook made changes that backfired. Even Mr. Zuckerberg’s pet initiatives have been thwarted by his own systems and algorithms.

The documents include research reports, online employee discussions and drafts of presentations to senior management, including Mr. Zuckerberg. They aren’t the result of idle grumbling, but rather the formal work of teams whose job was to examine the social network and figure out how it could improve.

They offer perhaps the clearest picture thus far of how broadly Facebook’s problems are known inside the company, up to the CEO himself. And when Facebook speaks publicly about many of these issues, to lawmakers, regulators and, in the case of XCheck, its own Oversight Board, it often provides misleading or partial answers, masking how much it knows.

One area in which the company hasn't struggled is profitability. In the past five years, during which it has been under intense scrutiny and roiled by internal debate, Facebook has generated profit of more than $100 billion. The company is currently valued at more than $1 trillion.

Rough justice

For ordinary users, Facebook dispenses a kind of rough justice in assessing whether posts meet the company's rules against bullying, sexual content, hate speech and incitement to violence. Sometimes the company's automated systems summarily delete or bury content suspected of rule violations without a human review. At other times, material flagged by those systems or by users is assessed by content moderators employed by outside companies.

Mr. Zuckerberg estimated in 2018 that Facebook gets 10% of its content removal decisions wrong, and, depending on the enforcement action taken, users might never be told what rule they violated or be given a chance to appeal.

Users designated for XCheck review, however, are treated more deferentially. Facebook designed the system to minimize what its employees have described in the documents as "PR fires"—negative media attention that comes from botched enforcement actions taken against VIPs.

If Facebook's systems conclude that one of those accounts might have broken its rules, they don't remove the content—at least not right away, the documents indicate. They route the complaint into a separate system, staffed by better-trained, full-time employees, for additional layers of review.

Most Facebook employees were able to add users into the XCheck system, the documents say, and a 2019 audit found that at least 45 teams around the company were involved in whitelisting. Users aren't generally told that they have been tagged for special treatment. An internal guide to XCheck eligibility cites qualifications including being "newsworthy," "influential or popular" or "PR risky."

Neymar, the Brazilian soccer star whose full name is Neymar da Silva Santos Jr., easily qualified. With more than 150 million followers, Neymar's account on Instagram, which is owned by Facebook, is one of the most popular in the world.

After a woman accused Neymar of rape in 2019, he posted Facebook and Instagram videos defending himself—and showing viewers his WhatsApp correspondence with his accuser, which included her name and nude photos of her. He accused the woman of extorting him.

Facebook's standard procedure for handling the posting of "nonconsensual intimate imagery" is simple: Delete it. But Neymar was protected by XCheck.

For more than a day, the system blocked Facebook's moderators from removing the video. An internal review of the incident found that 56 million Facebook and Instagram users saw what Facebook described in a separate document as "revenge porn," exposing the woman to what an employee referred to in the review as abuse from other users.

“This included the video being reposted more than 6,000 times, bullying and harassment about her character,” the review found.

Facebook’s operational guidelines stipulate that not only should unauthorized nude photos be deleted, but that people who post them should have their accounts deleted.

“After escalating the case to leadership,” the review said, “we decided to leave Neymar’s accounts active, a departure from our usual ‘one strike’ profile disable policy.”

Neymar denied the rape allegation, and no charges were filed against him. The woman was charged by Brazilian authorities with slander, extortion and fraud. The first two charges were dropped, and she was acquitted of the third. A spokesperson for Neymar said the athlete adheres to Facebook’s rules and declined to comment further.

The lists of those enrolled in XCheck were “scattered throughout the company, without clear governance or ownership,” according to a “Get Well Plan” from last year. “This results in not applying XCheck to those who pose real risks and on the flip-side, applying XCheck to those that do not deserve it (such as abusive accounts, persistent violators). These have created PR fires.”

In practice, Facebook appeared more concerned with avoiding gaffes than mitigating high-profile abuse. One Facebook review in 2019 of major XCheck errors showed that of 18 incidents investigated, 16 involved instances where the company erred in actions taken against prominent users.

Four of the 18 touched on inadvertent enforcement actions against content from Mr. Trump and his son, Donald Trump Jr. Other flubbed enforcement actions were taken against the accounts of Sen. Elizabeth Warren, fashion model Sunnaya Nash, and Mr. Zuckerberg himself, whose live-streamed employee Q&A had been suppressed after an algorithm classified it as containing misinformation.

Pulling content

Historically, Facebook contacted some VIP users who violated platform policies and provided a “self-remediation window” of 24 hours to delete violating content on their own before Facebook took it down and applied penalties.

Mr. Stone, the company spokesman, said Facebook has phased out that perk, which was still in place during the 2020 elections. He declined to say when it ended.

At times, pulling content from a VIP’s account requires approval from senior executives on the communications and public-policy teams, or even from Mr. Zuckerberg or Chief Operating Officer Sheryl Sandberg, according to people familiar with the matter.

In June 2020, a Trump post came up during a discussion about XCheck's hidden rules that took place on the company's internal communications platform, called Facebook Workplace. The previous month, Mr. Trump said in a post: "When the looting starts, the shooting starts."

A Facebook manager noted that an automated system, designed by the company to detect whether a post violates its rules, had scored Mr. Trump's post 90 out of 100, indicating a high likelihood it violated the platform's rules.

For a normal user post, such a score would result in the content being removed as soon as a single person reported it to Facebook. Instead, as Mr. Zuckerberg publicly acknowledged last year, he personally made the call to leave the post up. "Making a manual decision like this seems less defensible than algorithmic scoring and actioning," the manager wrote.

Mr. Trump's account was covered by XCheck before his two-year suspension from Facebook in June. So too are those belonging to members of his family, Congress and the European Union Parliament, along with mayors, civic activists and dissidents.

While the program included most government officials, it didn't include all candidates for public office, at times effectively granting incumbents in elections an advantage over challengers. The discrepancy was most prevalent in state and local races, the documents show, and employees worried Facebook could be subject to accusations of favoritism.

Mr. Stone acknowledged the concern but said the company had worked to address it. "We made multiple efforts to ensure that both in federal and nonfederal races, challengers as well as incumbents were included in the program," he said.

The program covers pretty much anyone regularly in the media or who has a substantial online following, including film stars, cable talk-show hosts, academics and online personalities with large followings. On Instagram, XCheck covers accounts for popular animal influencers including "Doug the Pug."

In practice, most of the content flagged by the XCheck system faced no subsequent review, the documents show.

Even when the company does review the material, enforcement delays like the one on Neymar's posts mean content that should have been prohibited can spread to large audiences. Last year, XCheck allowed posts that violated its rules to be viewed at least 16.4 billion times, before later being removed, according to a summary of the program in late December.

Facebook recognized years ago that the enforcement exemptions granted by its XCheck system were unacceptable, with protections sometimes granted to what it called abusive accounts and persistent violators of the rules, the documents show. Nevertheless, the program expanded over time, with tens of thousands of accounts added just last year.

In addition, Facebook has asked fact-checking partners to retroactively change their findings on posts from high-profile accounts, waived standard punishments for propagating what it classifies as misinformation and even altered planned changes to its algorithms to avoid political fallout.

“Facebook currently has no firewall to insulate content-related decisions from external pressures,” a September 2020 memo by a Facebook senior research scientist states, describing daily interventions in its rule-making and enforcement process by both Facebook’s public-policy team and senior executives.

A December memo from another Facebook data scientist was blunter: “Facebook routinely makes exceptions for powerful actors.”

Flubbed calls

Mr. Zuckerberg has consistently framed his position on how to moderate controversial content as one of principled neutrality. “We do not want to become the arbiters of truth,” he told Congress in a hearing last year.

Facebook’s special enforcement system for VIP users arose from the fact that its human and automated content-enforcement systems regularly flub calls.

Part of the problem is resources. While Facebook has trumpeted its spending on an army of content moderators, it still isn’t capable of fully processing the torrent of user-generated content on its platforms. Even assuming adequate staffing and a higher accuracy rate, making millions of moderation decisions a day would still involve numerous high-profile calls with the potential for bad PR.

Facebook wanted a system for “reducing false positives and human workload,” according to one internal document. The XCheck system was set up to do that.

To minimize conflict with average users, the company has long kept its notifications of content removals opaque. Users often describe on Facebook, Instagram or rival platforms what they say are removal errors, often accompanied by a screenshot of the notice they receive.

Facebook pays close attention. One internal presentation about the issue last year was titled “Users Retaliating Against Facebook Actions.”

“Literally all I said was happy birthday,” one user posted in response to a botched takedown, according to the presentation.

“Apparently Facebook doesn’t allow complaining about paint colors now?” another user complained after Facebook flagged as hate speech the declaration that “white paint colors are the worst.”

“Users like to screenshot us at our most ridiculous,” the presentation said, noting they often are outraged even when Facebook correctly applies its rules.

If getting panned by everyday users is unpleasant, inadvertently upsetting prominent ones is potentially embarrassing.

Last year, Facebook's algorithms misinterpreted a years-old post from Hosam El Sokkari, an independent journalist who once headed the BBC's Arabic News service, according to a September 2020 "incident review" by the company.

In the post, he condemned Osama bin Laden, but Facebook's algorithms misinterpreted the post as supporting the terrorist, which would have violated the platform's rules. Human reviewers erroneously concurred with the automated decision and denied Mr. El Sokkari's appeal.

As a result, Mr. El Sokkari's account was blocked from broadcasting a live video shortly before a scheduled public appearance. In response, he denounced Facebook on Twitter and the company's own platform in posts that received hundreds of thousands of views.

Facebook swiftly reversed itself, but shortly afterward mistakenly took down more of Mr. El Sokkari's posts criticizing conservative Muslim figures.

Mr. El Sokkari responded: "Facebook Arabic support team has obviously been infiltrated by extremists," he tweeted, an assertion that prompted more scrambling inside Facebook.

After seeking input from 41 employees, Facebook said in a report about the incident that XCheck remained too often "reactive and demand-driven." The report concluded that XCheck should be expanded further to include prominent independent journalists such as Mr. El Sokkari, to avoid future public-relations black eyes.

As XCheck mushroomed to encompass what the documents said are millions of users world-wide, reviewing all the questionable content became a fresh mountain of work.

Whitelist status

In response to what the documents describe as chronic underinvestment in moderation efforts, many teams around Facebook chose not to enforce the rules with high-profile accounts at all—the practice referred to as whitelisting. In some instances, whitelist status was granted with little record of who had granted it and why, according to the 2019 audit.

"This problem is pervasive, touching almost every area of the company," the 2019 review states, citing the audit. It concluded that whitelists "pose numerous legal, compliance, and legitimacy risks for the company and harm to our community."

A plan to fix the program, described in a document the following year, said that blanket exemptions and posts that were never subsequently reviewed had become the core of the program, meaning most content from XCheck users wasn't subject to enforcement. "We currently review less than 10% of XChecked content," the document stated.

Mr. Stone said the company improved that ratio during 2020, though he declined to provide data.

The leeway given to prominent political accounts on misinformation, which the company in 2019 acknowledged in a limited form, baffled some employees responsible for protecting the platforms. High-profile accounts posed greater risks than regular ones, researchers noted, yet were the least policed.

“We are knowingly exposing users to misinformation that we have the processes and resources to mitigate,” said a 2019 memo by Facebook researchers, called “The Political Whitelist Contradicts Facebook’s Core Stated Principles.” Technology website The Information previously reported on the document.

In one instance, political whitelist users were sharing articles from alternative-medicine websites claiming that a Berkeley, Calif., doctor had revealed that chemotherapy doesn’t work 97% of the time. Fact-checking organizations have debunked the claims, noting that the science is misrepresented and that the doctor cited in the article died in 1978.

In an internal comment in response to the memo, Samidh Chakrabarti, an executive who headed Facebook’s Civic Team, which focuses on political and social discourse on the platform, voiced his discomfort with the exemptions.

“One of the fundamental reasons I joined FB Is that I believe in its potential to be a profoundly democratizing force that enables everyone to have an equal civic voice,” he wrote. “So having different rules on speech for different people is very troubling to me.”

Other employees said the practice was at odds with Facebook’s values.

“FB’s decision-making on content policy is influenced by political considerations,” wrote an economist in the company’s data-science division.

“Separate content policy from public policy,” recommended Kaushik Iyer, then lead engineer for Facebook’s civic integrity team, in a June 2020 memo.

Buzzfeed previously reported on elements of these documents.

That same month, employees debated on Workplace, the internal platform, about the merits of going public with the XCheck program.

As the transparency proposal drew dozens of “like” and “love” emojis from colleagues, the Civic Team’s Mr. Chakrabarti looped in the product manager overseeing the XCheck program to offer a response.

The fairness concerns were real and XCheck had been mismanaged, the product manager wrote, but “we have to balance that with business risk.“ Since the company was already trying to address the program’s failings, the best approach was “internal transparency,” he said.

On May 5, Facebook's Oversight Board upheld the suspension of Mr. Trump, whom it accused of creating a risk of violence in connection with the Jan. 6 riot at the Capitol in Washington. It also criticized the company's enforcement practices, recommending that Facebook more clearly articulate its rules for prominent individuals and develop penalties for violators.

As one of 19 recommendations, the board asked Facebook to "report on the relative error rates and thematic consistency of determinations made through the cross check process compared with ordinary enforcement procedures."

A month later, Facebook said it was implementing 15 of the 19 recommendations. The one about disclosing cross check data was one of the four it said it wouldn't adopt.

"It's not feasible to track this information," Facebook wrote in its responses. "We have explained this product in our newsroom," it added, linking to a 2018 blog post that declared "we remove content from Facebook, no matter who posts it, when it breaks our standards." Facebook's 2019 internal review had previously cited that same blog post as misleading.

The XCheck documents show that Facebook misled the Oversight Board, said Kate Klonick, a law professor at St. John's University. The board was funded with an initial $130 million commitment from Facebook in 2019, and Ms. Klonick was given special access by the company to study the group's formation and its processes.

"Why would they spend so much time and money setting up the Oversight Board, then lie to it?" she said of Facebook after reviewing XCheck documentation at the Journal's request. "This is going to completely undercut it."

In a written statement, a spokesman for the board said it "has expressed on multiple occasions its concern about the lack of transparency in Facebook's content moderation processes, especially relating to the company's inconsistent management of high-profile accounts."

Facebook is trying to eliminate the practice of whitelisting, the documents show and the company spokesman confirmed. The company set a goal of eliminating total immunity for "high severity" violations of FB rules in the first half of 2021. A March update reported that the company was struggling to rein in additions to XCheck.

"VIP lists continue to grow," a product manager on Facebook's Mistakes Prevention Team wrote. She announced a plan to "stop the bleeding" by blocking Facebook employees' ability to enroll new users in XCheck.

One potential solution remains off the table: holding high-profile users to the same standards as everyone else.

“We do not have systems built out to do that extra diligence for all integrity actions that can occur for a VIP,” her memo said. To avoid making mistakes that might anger influential users, she noted, Facebook would instruct reviewers to take a gentle approach.

“We will index to assuming good intent in our review flows and lean into ‘innocent until proven guilty,’ ” she wrote.

The plan, the manager wrote, was “generally” supported by company leadership.

—Design by Andrew Levinson. A color filter has been used on some photos