LEAKED FACEBOOK FILES REVEAL FACEBOOK IS AN OUT-OF-CONTROL SOCIOPATHIC COMPANY

Revealed: Facebook's internal rulebook on sex, terrorism and violence

Leaked policies guiding moderators on what content to allow are likely to fuel debate about social media giant’s ethics

The Facebook Files: sex, violence and hate speech

-

Nick Hopkins

Facebook’s secret rules and guidelines for deciding what its 2 billion users can post on the site are revealed for the first time in a Guardian investigation that will fuel the global debate about the role and ethics of the social media giant.

The Guardian has seen more than 100 internal training manuals, spreadsheets and flowcharts that give unprecedented insight into the blueprints Facebook has used to moderate issues such as violence, hate speech, terrorism, pornography, racism and self-harm.

Facebook will let users livestream self-harm, leaked documents show

Read more

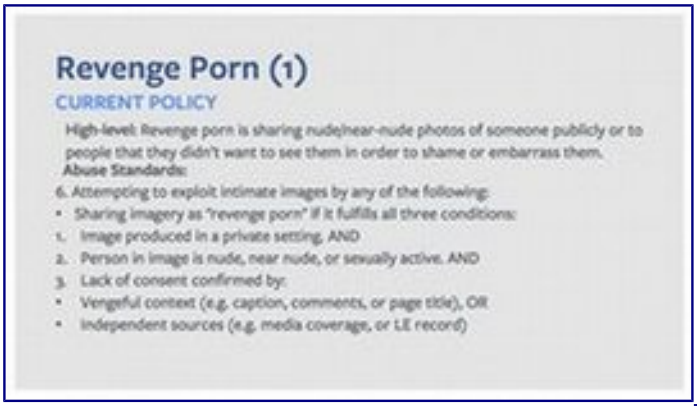
There are even guidelines on match-fixing and cannibalism.

The Facebook Files give the first view of the codes and rules formulated by the site, which is under huge political pressure in Europe and the US.

They illustrate difficulties faced by executives scrabbling to react to new challenges such as “revenge porn” – and the challenges for moderators, who say they are overwhelmed by the volume of work, which means they often have “just 10 seconds” to make a decision.

“Facebook cannot keep control of its content,” said one source. “It has grown too big, too quickly.”

Many moderators are said to have concerns about the inconsistency and peculiar nature of some of the policies. Those on sexual content, for example, are said to be the most complex and confusing.



Facebook Twitter Pinterest

A slide on Facebook's revenge porn policy. Photograph: Guardian

One document says Facebook reviews more than 6.5m reports a week relating to potentially fake accounts – known as FNRP (fake, not real person).

Using thousands of slides and pictures, Facebook sets out guidelines that may worry critics who say the service is now a publisher and must do more to remove hateful, hurtful and violent content.

Yet these blueprints may also alarm free speech advocates concerned about Facebook's de facto role as the world's largest censor. Both sides are likely to demand greater transparency.

The Guardian has seen documents supplied to Facebook moderators within the last year. The files tell them:

Facebook's internal manual on non-sexual child abuse content

Read more

- Remarks such as "Someone shoot Trump" should be deleted, because as a head of state he is in a protected category. But it can be permissible to say: "To snap a bitch's neck, make sure to apply all your pressure to the middle of her throat", or "fuck off and die" because they are not regarded as credible threats.
- Videos of violent deaths, while marked as disturbing, do not always have to be deleted because they can help create awareness of issues such as mental illness.
- Some photos of non-sexual physical abuse and bullying of children do not have to be deleted or "actioned" unless there is a sadistic or celebratory element.
- Photos of animal abuse can be shared, with only extremely upsetting imagery to be marked as "disturbing".
- All "handmade" art showing nudity and sexual activity is allowed but digitally made art showing sexual activity is not.
- Videos of abortions are allowed, as long as there is no nudity.

- Facebook will allow people to livestream attempts to self-harm because it “doesn’t want to censor or punish people in distress”.
- Anyone with more than 100,000 followers on a social media platform is designated as a public figure – which denies them the full protections given to private individuals.

Facebook's manual on credible threats of violence

Read more

Other types of remarks that can be permitted by the documents include: “Little girl needs to keep to herself before daddy breaks her face,” and “I hope someone kills you.” The threats are regarded as either generic or not credible.

In one of the leaked documents, Facebook acknowledges “people use violent language to express frustration online” and feel “safe to do so” on the site.

It says: “They feel that the issue won’t come back to them and they feel indifferent towards the person they are making the threats about because of the lack of empathy created by communication via devices as opposed to face to face.

Facebook Twitter Pinterest

Facebook’s policy on threats of violence. A tick means something can stay on the site; a cross means it should be deleted. Photograph: Guardian

“We should say that violent language is most often not credible until specificity of language gives us a reasonable ground to accept that there is no longer simply an expression of emotion but a transition to a plot or design. From this perspective language such as ‘I’m going to kill you’ or ‘Fuck off and die’ is not credible and is a violent expression of dislike and frustration.”

It adds: “People commonly express disdain or disagreement by threatening or calling for violence in generally facetious and unserious ways.”

Facebook conceded that “not all disagreeable or disturbing content violates our community standards”.

Monika Bickert, Facebook’s head of global policy management, said the service had almost 2 billion users and that it was difficult to reach a consensus on what to allow.

Facebook Twitter Pinterest

A Facebook slide on threats of violence. The ticks mean these statements need not be deleted. Photograph: Guardian

“We have a really diverse global community and people are going to have very different ideas about what is OK to share. No matter where you draw the line there are always going to be some grey areas. For instance, the line between satire and humour and inappropriate content is sometimes very grey. It is very difficult to decide whether some things belong on the site or not,” she said.

mod policy

“We feel responsible to our community to keep them safe and we feel very accountable. It’s absolutely our responsibility to keep on top of it. It’s a company commitment. We will continue to invest in proactively keeping the site safe, but we also want to empower people to report to us any content that breaches our standards.”

She said some offensive comments may violate Facebook policies in some contexts, but not in others.

Facebook’s leaked policies on subjects including violent death, images of non-sexual physical child abuse and animal cruelty show how the site tries to navigate a minefield.

The files say: “Videos of violent deaths are disturbing but can help create awareness. For videos, we think minors need protection and adults need a choice. We mark as ‘disturbing’ videos of the violent deaths of humans.”

Such footage should be “hidden from minors” but not automatically deleted because it can “be valuable in creating awareness for self-harm afflictions and mental illness or war crimes and other important issues”.

A slide on animal cruelty. Photograph: Guardian

Regarding non-sexual child abuse, Facebook says: “We do not action photos of child abuse. We mark as disturbing videos of child abuse. We remove imagery of child abuse if shared with sadism and celebration.”

One slide explains Facebook does not automatically delete evidence of non-sexual child abuse to allow the material to be shared so “the child [can] be identified and rescued, but we add protections to shield the audience”. This might be a warning on the video that the content is disturbing.

Facebook confirmed there are “some situations where we do allow images of non-sexual abuse of a child for the purpose of helping the child”.

Facebook's rules on showing cruelty to animals

Read more

Its policies on animal abuse are also explained, with one slide saying: “We allow photos and videos documenting animal abuse for awareness, but may add viewer protections to some content that is perceived as extremely disturbing by the audience.

“Generally, imagery of animal abuse can be shared on the site. Some extremely disturbing imagery may be marked as disturbing.”

Photos of animal mutilation, including those showing torture, can be marked as disturbing rather than deleted. Moderators can also leave photos of abuse where a human kicks or beats an animal.

Facebook Twitter Pinterest

The documents include guidance on non-sexual child abuse. Composite: alamy

Facebook said: “We allow people to share images of animal abuse to raise awareness and condemn the abuse but remove content that celebrates cruelty against animals.”

The files show Facebook has issued new guidelines on nudity after last year’s outcry when it removed an iconic Vietnam war photo because the girl in the picture was naked.

It now allows for “newsworthy exceptions” under its “terror of war” guidelines but draws the line at images of “child nudity in the context of the Holocaust”.

Facebook told the Guardian it was using software to intercept some graphic content before it got on the site, but that “we want people to be able to discuss global and current events … so the context in which a violent image is shared sometimes matters”.

Some critics in the US and Europe have demanded that the company be regulated in the same way as mainstream broadcasters and publishers.

Facebook Twitter Pinterest

A Facebook slide on its Holocaust policy. Photograph: Guardian

But Bickert said Facebook was “a new kind of company. It’s not a traditional technology company. It’s not a traditional media company. We build technology, and we feel responsible for how it’s used. We don’t write the news that people read on the platform.”

A report by British MPs published on 1 May said “the biggest and richest social media companies are shamefully far from taking sufficient action to tackle illegal or dangerous content, to implement proper community standards or to keep their users safe”.

Sarah T Roberts, an expert on content moderation, said: “It’s one thing when you’re a small online community with a group of people who share principles and values, but when you have a large percentage of the world’s population and say ‘share yourself’, you are going to be in quite a muddle.

“Then when you monetise that practice you are entering a disaster situation.”

Facebook's internal guidance on showing graphic violence

Read more

Facebook has consistently struggled to assess the news or “awareness” value of violent imagery. While the company recently faced harsh criticism for failing to remove videos of Robert Godwin being killed in the US and of a father killing his child in Thailand, the platform has also played an important role in disseminating videos of police killings and other government abuses.

In 2016, Facebook removed a video showing the immediate aftermath of the fatal police shooting of Philando Castile but subsequently reinstated the footage, saying the deletion was a “mistake”.