

ROOM FOR IMPROVEMENT. ANALYSING REDRESS POLICY ON FACEBOOK, INSTAGRAM, YOUTUBE AND TWITTER

This article reviews how users can appeal (in)action taken against content and/or accounts on four major social media platforms. It provides policy advice in the context of the next steps of the EU Digital Services Act (DSA), in particular Article 17 concerning internal complaint handling and redress mechanisms. We find that there is ample room for improvement in current appeal mechanisms, most notably in informing and allowing the person who flagged the content or account to seek redress for platform (in)action.

By Trisha Meyer, Professor in Digital Governance and Participation at the Vrije Universiteit Brussel, and Claire Pershan, Policy Coordinator at EU DisinfoLab

When users sign up to an online platform, they agree to the platform's terms of service, policies and community guidelines. Inaction or partial action by social media platforms on content that violates their policies enables the spread of online harms like disinformation and violence. It can also be seen as a [failure to take responsibility](#) in nurturing online spaces that uphold democracy, the rule of law and human rights. [Discrepancy](#) or [confusion](#) between platforms' content moderation policies and their practices erodes users' rights.

As with many content moderation challenges, this problem is more pronounced in [less widely spoken languages](#) and [places where platforms invest less](#) in content moderation.

Some platforms have processes through which users can report violations of the platform's terms of service. But currently there is no stable legal environment where platforms are obliged to respond to users when they make wrongful content moderation decisions - that is, when they [wrongfully act](#) on content as well as when they [fail to act](#) against violations of their platform policies. This is the challenge we examine by drawing on the public resources of four platforms, in which they explain their rules and guidelines and the enforcement of those conditions (through their help, support and transparency centers).

The EU's Digital Services Act will oblige online platforms to make an internal complaint handling system (redress mechanism) available to users, through which users can lodge complaints against content moderation decisions.

Some platforms have such redress mechanisms already, with varying formulations, scope, and accountability. To add to the confusion, a number of different terms are employed to describe "reporting" and "redress" mechanisms. In this article we will refer to "reporting" as the first action a person (an individual user or an entity) undertakes when they alert a platform to content that potentially violates terms of service or policies. Legal regimes, including the Digital Services Act, generally use the term "notify" or "notice and action" to refer to a person's alert to a platform containing potentially illegal content (DSA Article 14). "Flagging" meanwhile is a term we avoid, since it also connotes "Trusted Flaggers", or entities which have a privileged reporting status vis-a-vis platforms (this will be codified in the DSA's Article 19). Finally, in this article, "appeals" refers to a person seeking "remedy" or "redress" on a decision the person feels was wrongful. Importantly, in our view, appeal should be available to both the person against whom action has been taken, as well as the person who reported the content or account.

Why this matters to us

In September 2021, EU DisinfoLab released [a summary](#) of the responses we had received from social media platforms after alerting them to the influence operations we found on their platforms. As we explain, we received inconsistent responses, and sometimes no response at all, after reporting these violations of terms of service to the companies. For instance, in 2019 we detected and exposed a network spreading disinformation in partnership with an established media outlet. We reported this network to a large platform on which it was operating, but the platform did not respond. Two years later we discovered this network, still active, now spreading COVID-19 and anti-vaccine disinformation. We reported it again, this time the platform took a large number of the accounts down. The single actor behind all these accounts, however, appealed and managed to get the moderation decisions reversed, thereby returning online and continuing spreading disinformation and breaching terms of service. Only years later, after we had released a blogpost pointing out this platform's inaction, were the accounts in question finally moderated.

Researchers reporting disinformation networks are not the only group who would benefit from a consistent, reliable reporting and appeals system. The [reports of victims of hate speech and online violence](#) - as well as other behaviors that violate platform terms of service - frequently go unanswered.

The status quo: a look at four major platforms

We took a deep dive into the policies and community guidelines of Facebook, Instagram, YouTube and Twitter, scouring through their help, support and transparency centers, to better understand (a) how users can appeal actions taken by platforms against their content and/or account as well as (b) how those notifiers can appeal platform (lack of) action. The main primary sources are provided in the endnote to this article¹.

Meta (Facebook and Instagram)

Reporting abuse - <https://www.facebook.com/help/1753719584844061>

How to report things - https://help.instagram.com/2922067214679225?helpref=hc_fnav

How to appeal to the Oversight Board - <https://transparency.fb.com/en-gb/oversight/appealing-to-oversight-board/>

Taking down violating content - <https://transparency.fb.com/en-gb/enforcement/taking-action/taking-down-violating-content/>

Google (YouTube)

Appeal Community Guidelines actions -

https://support.google.com/youtube/answer/185111?hl=en&ref_topic=9387060

Reporting and enforcement - https://support.google.com/youtube/topic/2803138?hl=en&ref_topic=6151248

Report inappropriate videos, channels, and other content on YouTube -

https://support.google.com/youtube/answer/2802027?hl=en&ref_topic=9387085#zippy=

Twitter

How we enforce our rules - <https://help.twitter.com/en/resources/rules>

Report abusive behavior - <https://help.twitter.com/en/safety-and-security/report-abusive-behavior>

Our approach to policy development and enforcement philosophy - <https://help.twitter.com/en/rules-and-policies/enforcement-philosophy>

¹ The following primary sources were consulted in investigating platforms' redress mechanisms. We only list the main sources; a full list can be provided upon request.

We focused on platform action taken based on their community guidelines. This includes content and behavior that the platforms deem (likely to be) illegal, but also addresses harmful ('awful yet lawful') and other content and behavior that the platforms consider violating their policies and standards. This vast category ranges from sexual exploitation and hate speech to inauthentic behavior and misinforming content. Importantly, we did not consider appeal and redress when legal requests are made (as opposed to objections based on platform policy). This latter category includes copyright requests, government requests, but also content restrictions based on local laws, such as the German NetzDG law.

This table summarizes our main findings, which we expand upon in the following paragraphs.

Content/account moderation		Meta (Facebook/Instagram)	Google (YouTube)	Twitter
Reporting and review of content and accounts	Machine-driven	x		x
	User-driven	x	X	x
	Third party	x	x	x
Appeal (person whose content/account was actioned upon)	Notification of action	x	x	x
	Possibility to appeal	(x)	x	x
Appeal (person who reported violating content/account)	Notification of action	x	?	x
	Possibility to appeal	(x)	?	?
Escalation options		x		

Table 1: Facebook, Instagram, YouTube and Twitter redress mechanisms (own compilation)

All four platforms offer multiple opportunities for reporting, including by users and "Trusted Flaggers". Transparency reports provide figures on removal of content and accounts on a quarterly basis. On Meta, [actioned content](#) refers to warning labels applied to content and disabling of accounts as well. **However, there is no metric recording the number of reports (based on community guidelines) that went unanswered or did not lead to content being removed.** None could be found for Google and Meta, combining two datasets would allow us to obtain this figure for Twitter.


 In response to COVID-19, we've taken **steps** to protect our extended workforce and reduce in-office staffing. As a result, we are temporarily relying more on technology to help with some of the work normally done by human reviewers, which means we are removing more content that may not be violative of our policies. This impacts some of the metrics in this report and will likely continue to impact metrics moving forward. For the latest updates on how we're addressing the COVID-19 situation, please visit g.co/yt-covid19.

Figure 1: [Google](#) over-blocking disclaimer

However, during the COVID-19 pandemic, platforms rely more heavily on automated machine-driven review. **Google acknowledges that over-blocking likely occurs.** Between October to December 2021, Google [reports](#) that they removed 3,754,215 videos on YouTube. They received 213,346 appeals and as a result reinstated 43,331 pieces of content. 5.7% of content removals were appealed and 1.2% of that content reinstated. Strikingly, this means that 20.3% of all appealed content was reinstated, meaning the initial judgment was wrong for one in five pieces of appealed content.

At the same time, **on Meta's Facebook, due to the pandemic, there is less possibility for appeal by users** whose content or accounts were considered violating community standards.

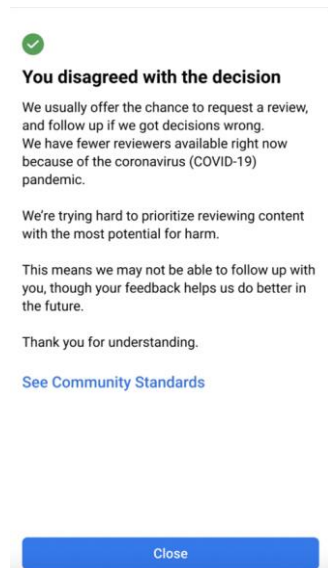


Figure 2: [Facebook](#) limited appeal disclaimer

However Meta did [announce](#) on January 19, 2022, that **on Facebook they are beginning to provide appeals not just for content that they took action on, but also for content that was reported but not acted on.**

What can be appealed

Today, we offer appeals for the vast majority of violation types on Facebook. We don't offer appeals for violations with extreme safety concerns, such as child exploitation imagery.

We are beginning to provide appeals not just for content that we took action on, but also for content that was reported but not acted on. These reporter appeals are not included in the Community Standards Enforcement Report.

Figure 3: [Facebook](#) new appeal for inaction

Moreover, to their credit, **Meta is the only platform that currently provides an escalation option (redress mechanism) available through the [Oversight Board](#)** for "most content you posted on Facebook or Instagram that has been taken down and most content posted by another person that has been left up on Facebook or Instagram." This indicates awareness of the problem and the potential value of redress for platform inaction. However the Oversight Board is currently not capable of providing this kind of redress at scale for all users. (In its [record of cases](#) in its Transparency Center, the Oversight Board lists 29 cases since 1 December 2020.) The Oversight Board is seemingly more suited to addressing a small number of complex questions at a high level and to making policy recommendations.

We were not able to find evidence of similar possibilities of redress for users beyond appeal in the platform policies on YouTube or Twitter. Google [clarifies](#) that "you may appeal each strike only once." Importantly, **on YouTube and Twitter, there currently seems to be no option whatsoever for persons reporting content or accounts to appeal platform (in)action taken.** Indeed, in their support pages, Google does not even explain

whether or how notifiers will be informed about the outcome of their report, and Twitter [follows up](#) only if/when action has been taken.

What happens after I submit a report?

After you submit a report, you will see a confirmation message from us alerting you that we received your report (it may take up to 24 hours before you see a message). We will review the reported account and/or Tweet(s), and/or Direct Message(s). If we determine that the account, and/or Tweet(s), and/or Direct Message(s) are in violation of our policies, we will take action (ranging from a warning to permanently suspending the account). You will receive a follow up from us if we need more information from you, or when we take action on the reported account, and/or Tweet(s), and/or Direct Message(s).

Additionally, the original content of reported Tweets will be replaced with a notice stating that you reported it. You may click through and view the Tweet should you wish.

Note: Additionally, you will receive an in-product notification if an action is taken on an account that you recently reported. This action may or may not be related to your report.

Figure 4: [Twitter](#) follow-up on reports

What change the DSA would bring

Existing practices hinder the average user from seeking redress and discourage them from assisting in content moderation by alerting platforms to violations of their terms of service. They also add obscurity to how platforms enforce their content policies.

The EU's Digital Services Act is an opportunity to improve the existing power asymmetries between users and platforms, giving users the possibility to act. It will require online platforms to have a notice and action reporting mechanism for illegal content (Articles 14-15) and require them to give users access to an internal complaint handling mechanism (essentially a chance for an appeal) on content that is illegal or that violates terms and conditions, through its Article 17.

According to the draft of the agreed text of the Digital Services Act, which was seen by EU DisinfoLab on May 19:

1. Providers of online platforms shall provide recipients of the service, including individuals or entities that have submitted a notice, for a period of at least six months following the decision referred to in this paragraph, the access to an effective internal complaint-handling system, which enables the complaints to be lodged electronically and free of charge, against the decision taken by the provider of the online platform upon the receipt of a notice or against the following decisions taken by the provider of the online platform on the ground that the information provided by the recipients is illegal content or incompatible with its terms and conditions:

- (a) decisions whether or not to remove or disable access to or restrict visibility of the information;*
- (b) decisions whether or not to suspend or terminate the provision of the service, in whole or in part, to the recipients;*
- (c) decisions whether or not to suspend or terminate the recipients' account;*
- (d) decisions whether or not to suspend, terminate or otherwise restrict the ability to monetise content provided by the recipients.*

This version of the text would ensure a balanced complaint-handling system, enabling users to notify platforms when they have not acted on content that infringes the law or their terms and conditions. Platforms must also strengthen their human review. According to Article 17.5:

"Providers of online platforms shall ensure that the decisions, referred to in paragraph 4, are taken under the control of appropriately qualified staff, not solely taken on the basis of automated means".

Crucially, the DSA also allows the possibility of escalating a complaint by ensuring access to the DSA's certified out-of-court dispute settlement bodies (Article 18). Important details about these mechanisms are not yet clear, but this is seemingly the DSA's effort to impose standards for out-of-court dispute settlement, sometimes referred to as Alternative dispute resolution (ADR) or Digital Dispute Resolution (DDR) depending on the mechanism. This ensures that users can take further action against platforms when platforms choose not to enforce the rules that users abide by as defined in the terms of service.

What will this look like on the platforms? Won't this be abused?

Importantly, the DSA could create a ['two-track'](#) system for both reporting and appeals. If platforms maintain their current mechanisms and add further channels to meet the DSA's requirements, users will be faced with multiple options for how to report and how to appeal content, some of which may be available to only certain types of content. For instance, the DSA's 'notice-and-action' reporting mechanism will apply to content that is illegal under EU law, while platforms' own reporting mechanisms may be available for a wider range of content, for instance all content violating terms of service. Platforms will be unlikely to simply replace their existing mechanisms which serve users globally. There is a risk that this two-track system will add confusion to the reporting and appeals process that is already opaque and onerous for users.

For this reason, platforms should meet the DSA's obligations with transparent and user-friendly tools. Platforms must also uphold their side of the bargain regarding Article 20, for "measures and protections against misuse". Platforms may suspend users who frequently submit complaints that are "manifestly unfounded."

While there is no consistent data on abusive reporting or appeals in the area of disinformation, experts and platforms have documented misuse of reporting systems, for instance to silence oppositional expression or as a means of antagonistic online behaviour. There are also cases when users will appear to submit repeated requests simply as [a way to be heard](#), because the platform doesn't recognise the violation at first. It is crucial that the design and implementation of these systems is appropriate for good faith reporting and appeals and mitigates the potential of abuse.