

# Dilemma of Deactivation in the 'Networked Public Sphere'

September 21, 2011

One of the important protections that U.S. law offers Internet intermediaries is immunity from liability when they remove content or disable user activity that they consider inappropriate. This protection, which is found in a [law known as Section 230](#) [1], supports the right of social media services to kick off bullies, remove sexually explicit material, and generally write and enforce their own terms of service. But as the cliché goes, with great power comes great responsibility: the freedom to remove content or ban users is not one that service providers should take lightly.

Indeed, as social media services have struggled to serve user desires for safe social networks and to preserve company reputations, [often there arises the difficult and complex decision](#) [2] about when – and how – to remove content and disable user accounts. Sometimes, the decisions are clear. As Court Justice Potter Stewart famously said about obscene material, "I know it when I see it." But other decisions are not so black-and-white. Indeed, platforms have even removed artistic renderings of nudes that were mistaken for pornography.

And what should service providers do when an individual's account is flagged for hosting a violent video even though that video was in fact posted to expose an episode of police brutality? What impact does a mistake on the company's part – perhaps a mistaken allegation that a user's provided name is not her real name – have on that user's opportunities for expression? When a decision is made to remove a piece of content or deactivate an account, what processes should be adopted to mitigate potential harms and the likelihood of false positives?

Last year, CDT began working with Harvard's [Berkman Center on Internet & Society](#) [3] to develop a set of guidelines that could help companies think through these issues. Building off a process initiated by the [Global Network Initiative](#) [4], and consulting with a wide range of social media companies, advocates, and academics, we began crafting a set of good practices for companies that host, and users that create, user-generated content. Our [final report](#) [2], titled "Account Deactivation and Content Removal: Guiding Principles and Practices for Companies and Users," lays out some of the steps that platform providers can take to reduce violations of their terms of service, minimize the harm to users whose content is removed or whose accounts are deactivated, and help users understand their own responsibilities with respect to the content they create.

For example, one portion of the report is devoted to talking about that admittedly unsexy topic: user education. Users may inadvertently violate a platform's rules because these rules are written in dense, cryptic legalese or because they are not translated into a local language. Service providers can help educate users and minimize unintentional rule violations by providing plain language guidance about what is acceptable behavior and how to avoid content takedown. Such guidance should be translated into all languages in which the service is offered. For example, Facebook's Community Standards (a plain language version of their terms of service) have now been translated into close to 30 languages and have been opened up to community translation for the rest of the languages that Facebook supports.

## Warnings and Appeals

Companies should also craft predictable and transparent escalation processes for handling identified rule violations. In cases where disputed content is not dangerous or illegal, rather than removing it, the person posting the content should be given an initial warning. Where an account holder's behavior is problematic but not threatening, service providers should employ partial measures, such as limiting the types of activities an account is allowed to engage in, before actually deactivating the account. For example, when Blogger identifies a blog with adult content that has not been properly labeled as such, the service sometimes inserts an unavoidable "mature content" interstitial between the referring URL and the allegedly mature content blog. The interstitial allows Blogger to warn

visitors that certain content may be inappropriate for some users. In other instances, instead of removing a blog altogether, Blogger will "sandbox" it for a certain period of time, during which only the blog author can access the material. Many content platforms, however, could still afford to be a little less trigger-happy. And any escalation action should be accompanied by a clear explanation of what the user did to violate the rules, a description of the next steps in the process (probation? suspension?), and an opportunity for the user to appeal or ask questions.

Appeal processes, it turns out, are crucial. Platforms that host user-generated content often combine automated and human review of material that has been identified as potentially problematic, but even the most sophisticated automated processes and the most experienced human review teams have been caught making some pretty big mistakes. YouTube, for example, allows a user to appeal community flags on her video (and resulting "strikes" on her account). If YouTube agrees the video did not violate the Community Guidelines, it reinstates the video. If YouTube denies the appeal, the user is not allowed to appeal another flagged video for sixty days, which discourages meritless appeals. By allowing users to appeal content removal or account deactivation decisions, and by instituting processes for handling those appeals, platform providers like YouTube help ensure that a mistake made during the content evaluation process does not permanently silence a user's message or otherwise cause undue harm.

Service providers vary in terms of history, mission, content hosted, size, and user base and there is no one-size-fits-all answer. But social media services have become critical not only for other businesses, but for nearly all aspects of our social, economic, and political lives. And human rights and democracy movements rely on these tools in powerful ways. All companies can strive to be transparent and consistent in their interactions with users and in how they establish, communicate, and implement the "ground rules." By giving greater thought and attention to these issues, these companies can have a significant impact on user rights and user satisfaction.

- 
- [global Internet freedom](#)
- [Free Expression](#)

Copyright © 2013 by Center for Democracy & Technology. CDT can be freely copied and used as long as you make no substantive changes and clearly give us credit. [Details](#).

**Source URL:** <https://cdt.org/blogs/erica-newland/219dilemma-deactivation-networked-public-sphere>

#### Links:

- [1] <http://www.cdt.org/blogs/mark-stanley/219shielding-messengers-section-230-and-free-speech-online>
- [2] [http://www.cdt.org/files/pdfs/Report\\_on\\_Account\\_Deactivation\\_and\\_Content\\_Removal.pdf](http://www.cdt.org/files/pdfs/Report_on_Account_Deactivation_and_Content_Removal.pdf)
- [3] <http://cyber.law.harvard.edu/>
- [4] <http://www.globalnetworkinitiative.org>