User attention is the new frontier in content regulation – Improving the accountability of soft content moderation by social media platforms

Gastautor 2020-04-24T07:39:00

von Alexander Pirang

Social media platforms have an inglorious history of algorithmically promoting and amplifying misinformation, hate, and other repulsive behavior. These are serious challenges to the health of our information eco-systems, and they continue to persist, despite platforms' pledges to tackle the problem. However, we should not overlook another looming challenge in this space: algorithmic content curation has powerful gatekeeping functions, and platforms are increasingly putting them to use.

Indeed, there has been a marked paradigm shift over recent years in how social media platforms police online content. In addition to traditional "hard" moderation tools such as content removals or account suspensions, platform companies are increasingly relying on recommender systems, i.e. algorithms that guide users' attention towards selected content, and other design choices in order to systematically reduce the visibility and spread of unwanted content.

Systematic demotion of content and freedom of expression concerns

Facebook's policy change on "borderline content" offers an instructive example. In 2018, Mark Zuckerberg announced that his company would start to target borderline content, i.e. content that toes the line of what is acceptable, but that falls just short of violating the Community Standards. According to Zuckerberg, Facebook would achieve this by "penalizing borderline content so it gets less distribution and

engagement." Since then, Facebook announced that it would also begin to demote anti-vaccination and "low-quality content."

Other platforms systemically reduce user exposure to certain content as well. TikTok reportedly curbed the reach of videos showing people with disabilities; Twitter penalizes certain Tweets by turning off engagements such as likes, replies, and retweets.

The underlying logic of this approach is straightforward: let people continue to post things that are annoying, unsavory, or offensive, while shielding most users from having to actually encounter such material. Targeting user exposure instead of content allows platform companies to have the cake and eat it, too; they are able to deflect the private censorship criticism associated with hard content restriction and still massively reduce the prevalence of unwanted content.

Despite its general classification as a "soft" content restriction mode, reducing the visibility of content can easily turn into a massive freedom of expression problem. Platforms purport to be showing users what they do (not) want to see —less borderline content, for instance. That means that platforms prioritize their user communities' (presumed) preferences over the actual free speech interests of affected content providers and recipients, who are, respectively, interested in sharing and receiving that content.

Moreover, what we presumably want to see may be very different from what we *ought* to be exposed to as informed citizens in democratic societies. This raises hard questions as to how public values and individual user choices can be implemented into platforms' content curation, while reconciling them with the freedom of platform companies to customize their services. Importantly, this is where the content moderation debate overlaps with the broader issue of content curation.

Demotion and other changes to the visibility of content also threatens to undo much of the progress that has only recently been made in prompting platform companies to become more accountable regarding hard content moderation. Practices targeting user exposure mainly affect content that falls outside of platforms' policies for prohibited material; it is therefore unclear what rules apply to such measures (try your hand at defining "borderline content").

Given the worrisome implications for freedom of expression, platforms need to be held accountable for what they hide from our attention. One avenue—and the approach commonly favored by European policymakers and civil society organizations—is to increase the transparency of platforms' recommender systems.

Yet, transparency as an accountability mechanism has limitations. Entering into the complex debate on meaningful transparency is beyond the scope of this post. Instead, it focusses on other safeguards that appear to be more low-hanging fruit, namely notifications and appeal mechanisms.

Most large platforms have already implemented complaints procedures, which allow users to appeal certain takedown decisions and/or account-level sanctions.

In contrast, users affected by demotion are left out in the cold. Platforms give no pointers to tell if a piece of content was made less visible in search results or feeds. This uncertainty is jarring; as Robin Mansell pointed out, “[c]itizens cannot choose to view what they are not aware of or to protest about the absence of content which they cannot discover.”

The current European regulatory framework fails to adequately address this problem. If anything, the EU’s Code of Practice on Disinformation incentivizes its signatories, among them Facebook, Google, and Twitter, to “[d]ilute the visibility of disinformation by improving the findability of trustworthy content,” without mentioning even minimal safeguards. The Audiovisual Media Services Directive (AVMSD), which regulates video-sharing platforms, does require Member States to ensure that out-of-court redress mechanisms are available for users subjected to state-sanctioned content moderation (Art. 28b(7) AVMSD); however, measures reducing the spread of content appear to fall outside the scope of the AVMSD.

Proposal for an accountability mechanism

In light of the freedom of expression implications, platforms should notify users of changes to the visibility of content in the same way as when content is taken down. These notifications should state the reason for and the specifics of the measure taken. This safeguard would give affected users content-specific knowledge of how soft moderation affects the reach of certain categories of content.

In a next step, platforms should ideally allow users to appeal decisions impacting the reach of their content. This approach is not entirely unproblematic, since platform companies are in principle free to prioritize certain content over other material (although it should be noted that German courts are increasingly applying the doctrine of horizontal fundamental rights obligations in the context of takedown decisions). Appeal mechanisms would therefore have to be sufficiently narrow in scope to exempt general content curation.

Statutory regulation to this end may to be too heavy-handed, given the moving target and the fact that different platforms use different tools. A more flexible approach at the EU level, perhaps in the form of a Code of Conduct, seems more promising. Such a co-regulatory approach could also ensure a degree of public oversight, for instance by including binding reporting obligations.

In sum, European policymakers, researchers, civil society activists, and platform companies have to move fast to come to an understanding on how to best achieve accountability in a space that appears to be the new frontier in content regulation.

Zitiervorschlag: Alexander Pirang, User attention is the new frontier in content regulation – Improving the accountability of soft content moderation by social media platforms, JuWissBlog Nr. 64/2020 v. 24.04.2020, https://www.juwiss.de/64-2020/.