A Systemic Approach to Implementing the DSA's Human-in-the-Loop Requirement

Rachel Griffin, Erik Stallman

A Systemic Approach to Implementing the DSA's Human-in-the-Loop Requirement

VB verfassungsblog.de/a-systemic-approach-to-implementing-the-dsas-human-in-the-loop-requirement/

Rachel Griffin

Erik Stallman

This article belongs to the debate » From the DMCA to the DSA—A Transatlantic Dialogue on Online Platform Regulation and Copyright
22 February 2024

Policymakers and the public are increasingly concerned about a lack of transparency and accountability in content moderation. Opaque and incontestable content moderation decisions have potential impacts on freedom of expression and media freedom, and well-known issues of discrimination and bias. In the EU, improving fairness and accountability in content moderation is one important policy objective of the 2022 Digital Services Act (DSA).

Our contribution focuses on a core component of this legislative framework: Article 20 DSA, which sets out rules for online platforms' internal complaint-handling systems. Article 20 requires platforms to allow users to challenge moderation decisions, and have their complaints reviewed "under the supervision of appropriately qualified staff". Although scholars and commentators have raised important questions about the utility of trying to regulate complex, large-scale content moderation systems via 'due process' for individuals, this approach is now entrenched in European law. Accordingly, our focus here is on how Article 20 can and should be interpreted going forward. Specifically, does Article 20 require a human content moderator to review every content moderation decision on request? And should it?

Drawing on the broader literature on "human in the loop" requirements in artificial intelligence (AI) governance, we argue that formalistically requiring a human to look over every complaint is both normatively problematic and practically counterproductive. We set out an alternative approach, in which human review is oriented towards improving automated moderation systems at a systemic level, rather than correcting individual decisions. We argue that this is both permitted by the DSA text, and normatively preferable as a way of achieving the DSA's ultimate policy goals of preventing arbitrariness and discrimination in moderation.

What level of human review does Article 20 require?

Article 20 requires online platforms to establish 'easy to access, user-friendly' systems which allow users to complain about any moderation decision. This includes all kinds of actions (or inaction) on flagged content – from terminating an entire account to hiding a single comment – as well as decisions not to remove content, and decisions to reduce visibility or impose other interventions short of removal. This implies a vast number of decisions potentially subject to review. Article 20(4) requires platforms to consider complaints "in a timely, non-discriminatory, diligent and non-arbitrary manner" and reverse decisions where the complaint shows that they are not justified by the law or by platforms' content policies.

The vast majority of moderation decisions potentially subject to Article 20 complaints are fully automated – the only feasible way of monitoring content across platforms with millions or billions of users. A crucial question is therefore whether Article 20 requires complaints to be reviewed by human moderators. The answer not only implies potentially enormous investments of labour time and resources, but also has important implications for the overall effectiveness of the DSA.

Superficially, requiring human moderators to review complaints could seem like the most natural interpretation of Article 20. However, a close reading suggests otherwise. The key provision is Article 20(6), which requires 'decisions [to be] taken under the supervision of appropriately qualified staff, and not solely on the basis of automated means' (our emphasis). This seems to leave space for humans to play a more high-level supervisory role, rather than examining every individual complaint. Further guidance is provided by Recital 58:

> "providers of online platforms should be required to provide for **internal complaint-handling systems, which** meet certain conditions that aim to ensure that **the systems are** easily accessible and lead to swift, non-discriminatory, non-arbitrary and fair outcomes, **and are subject to human review** where automated means are used." (our emphasis)

Crucially, "are subject to human review" here refers to "systems", not to "complaints". Thus, it is the complaint-handling system as a whole which must be subject to human review and supervision – not necessarily every individual moderation decision. In the following sections, we will argue that this interpretation is not just legally permissible, but strongly preferable as a way of improving the quality, reliability and fairness of content moderation.

What is the point of human review?

The optimal design of human review processes in content moderation ultimately depends on what purposes they are meant to serve. Yet the DSA provides surprisingly little guidance on this. Recital 58 states that, 'Recipients of the service should be able to easily and effectively contest [moderation] decisions […] Therefore, providers of online platforms should be required to provide for internal complaint-handling systems.' The ultimate purpose of allowing recipients to contest moderation decisions is left unstated.

Turning to the broader literature on human oversight in AI governance, Rebecca Crootof, Margot Kaminski and W. Nicholson Price identify six possible reasons to impose 'human in the loop' requirements:

> "Humans may play (1) corrective roles to improve system performance, including error, situational, and bias correction; (2) justificatory roles to increase the system's legitimacy by providing reasoning for decisions; (3) dignitary roles to protect the dignity of the humans affected by the decision; (4) accountability roles to allocate liability or censure; (5) interface roles to link the systems to human users; and (6) "warm body" roles to preserve human jobs."

Considering their relevance to content moderation, we first want to emphasise that (6) is here a very bad reason. Moderators' working conditions are notoriously appalling. Major platforms outsource most such labour to Global South countries with lower wages and fewer worker protections, but even for workers in Global North markets – often migrants with few other employment options – it is characterised by fast-paced and stressful work, poor pay, and intense managerial surveillance. While these conditions could conceivably be improved, there is nothing in the DSA (a supposedly "comprehensive" regulation of online content governance) that tries to achieve this – an important point we will return to later. Reviewing harmful or offensive content is also, to some extent, an inherently repetitive, unpleasant, and psychologically taxing job.

It follows from this that reason (3) is also of questionable relevance. We do not believe it serves human dignity to allow every social media user to demand that some poorly paid and treated worker on the other side of the world quickly glances at their content. Reason (4) is also less relevant to content moderation, as the DSA's provisions on intermediary liability and regulatory oversight already regulate platforms' liability for moderation decisions. The most relevant goals for "human in the loop" requirements in relation to Article 20 are therefore (1) improving the performance of moderation systems, including by correcting errors and bias, and (2) and (5), justifying decisions and making them comprehensible to human users.

Human review of every contested decision is neither practical nor desirable

To effectively achieve these goals, we start with two observations about the roles that humans should not play in content moderation. First, it is neither practical nor desirable to have humans review every contested automated moderation decision. Automated moderation exists largely because humans can't operate at the scale required for timely action on content hosted on large platforms. In three months, YouTube removed 9 million videos and 1.16 billion comments. As Evelyn Douek notes, "even the smaller fraction of content moderation decisions that are appealed would still overload anything but an impractically large workforce."

Arguably, moderation workforces already have become impractically large and overloaded. Facebook alone has 15,000 content moderators worldwide. Yet moderators are also highly overworked, required to follow rigidly-defined workflows and meet demanding quotas which do not permit nuanced consideration of individual decisions. Research on 'humans in the loop' in AI shows that it is generally difficult for humans to identify and correct errors, due to "automation bias", where people tend to trust and defer to decision-making software. Increasing moderators' workloads is less likely to improve content moderation decisions than it is to lead to more frequent rubberstamping of automated decisions.

Furthermore, if Article 20 is interpreted as relying on a huge workforce to review and correct an enormous volume of contested automated moderation decisions, it is remarkable that the DSA contains virtually no regulation of these workers' pay, working conditions, qualifications and training (beyond some basic transparency requirements for very large online platforms, set out in Article 42). An inflexible and ill-defined human oversight requirement which effectively requires a permanent layer of low-paid, overworked, and over-stressed content moderators is not only in itself normatively problematic, but also seems like a suboptimal way to improve moderation quality.

Second, even assuming platforms could overcome workforce constraints, it is doubtful that a body of consistent reasoned decisions resolving content moderation complaints is a realistic or even desirable outcome. The scale, complexity, and diversity of content available on large online platforms means that "invoking judicial-style norms of reasoning and precedent is doomed to fail." Removing a platform's discretion as to which decisions are subject to further review still leaves a lot of room to tailor the reasoning and outcome of those reviews to limit their current or future impact, as an intensive study of the Meta Oversight Board has shown. And even a fully independent review body faithfully applying its own reasoned decisions to emerging cases would frequently find itself needing to depart from that precedent or a platform's own guidance. Community guidelines are perpetually revised in response to changing circumstances that those guidelines did not anticipate and for which they are a poor fit.

A better approach to human supervision

All content moderation systems are human/machine hybrids regardless of the degree of automation. Moderation software is designed by human engineers, and AI systems are trained on human decisions and evaluations, while hash-matching systems (like YouTube's ContentID system for copyright enforcement) are designed to search for copies of rightsholder-supplied reference files. On the basis that these hybrid systems are the appropriate target for supervision, rather than individual contested decisions, we identify four key considerations to improve their accuracy and proportionality.

First, instead of requiring cursory human review of every individual decision, the best way to evaluate and improve automated moderation is through more systematic oversight: for example, requiring policy experts to review statistically representative samples of decisions. Today's advanced AI tools, which are increasingly being deployed by major platforms for moderation tasks that would previously have required human intervention, rely on learning patterns from enormous datasets. However, recent technological advances are increasingly relying on smaller volumes of high-quality data, carefully curated or even produced to order by highly-qualified workers. A smaller, better-trained and better-paid moderation workforce, which carefully evaluates and provides detailed feedback on a subset of decisions, can oversee and improve moderation systems more effectively than an army of low-paid clickworkers – as well as being preferable from a labour rights perspective. Similarly, where failings are identified in hash-matching tools like ContentID, which scan for and remove copies of millions of files, it would be more productive to identify systematic flaws in the processes for (mis)identifying unlicensed and unlawful reproductions of content in their reference databases, rather than just trying to correct errors piecemeal.

Second, for this kind of systematic review to be effective, human reviewers must be able to understand what triggers automated flagging. Drawing on the extensive research literature on AI explainability, moderation systems should be designed to provide human supervisors "meaningful information about the logic involved" in moderation decisions. Conversely, their feedback should improve the automated system's decision-making in future. For example, if the machine failed to distinguish news reporting about terrorist activity from terrorist recruitment propaganda, the human reviewer could identify characteristics that help reinforce the distinction. This "bilateral explainability" should also factor into Article 20's requirement for supervision by "appropriately qualified" staff. Reviewers should have the qualifications and ability to facilitate machine-readable policy refinements that can minimise future errors.

Third, human supervision should be proportionate to different types of moderation decisions. Given the potential economic, reputational, and emotional consequences when users' entire accounts are removed, such decisions should receive more thorough review than, for example, demonetising content or hiding a comment. Meaningful review of deplatforming decisions should not be reserved for sitting presidents: we would suggest that in general, if someone will completely lose access to a platform, they should be able to appeal to a human customer service representative (potentially with some narrow exceptions, such as spam and duplicate accounts). In these serious cases, human review should not just involve a quick glance at a decision, but should enable meaningful communication with moderators. Furthermore, where machine learning led to an erroneous deplatforming decision, the human supervisor should ensure the machine learns from its mistake. That could mean reviewing and reannotating the relevant pieces of content used to train the machine learning classifiers that contributed to the erroneous decision.

Finally, human supervisors can appreciate what types of content pose particular concerns in a specific social, cultural, or political context: for example, political misinformation in the lead-up to a close election, or vaccine misinformation during a pandemic. Expert staff can dynamically allocate limited human and computing resources to address current and emerging threats. And given that the DSA itself may increase the risk of 'coordinated flagging', including misuse or manipulation of the Article 20 complaint system, platforms should dedicate some of their data science and cybersecurity resources to monitoring and addressing these risks – as they have historically done for threats like coordinated disinformation campaigns.

Conclusions

In the context of content moderation, we have argued against formalistic interpretations of human oversight requirements that simply require a person to confirm algorithmic decisions – whether based on the premise that the "human touch" somehow makes decisions more respectful of people's dignity, or on the optimistic assumption that having humans look at a decision is sufficient to correct algorithmic errors and bias. Instead, human review under Article 20 DSA should be geared towards improving the reliability and explainability of algorithmic moderation systems as a whole, as well as providing meaningful communication and support to users in the most consequential decisions (deplatforming).

These basic principles have wider relevance for tech regulation. For example, "human in the loop" requirements are also established in the EU's GDPR and proposed AI Act, as well as under various US legal frameworks. Ultimately, the optimal design of hybrid decision-making systems needs to be adapted to specific contexts. However, the approach we have set out here – interpreting "human in the loop" requirements purposively, and considering how review processes can be designed to serve the legislation's underlying normative and policy goals, rather than just checking a box – could also provide a helpful starting point for interpreting such requirements across different areas of AI regulation.

SUGGESTED CITATION Griffin, Rachel; Stallman, Erik: *A Systemic Approach to Implementing the DSA's Human-in-the-Loop Requirement, VerfBlog,* 2024/2/22, https://verfassungsblog.de/a-systemic-approach-to-implementing-the-dsas-human-in-the-loop-requirement/, DOI: 10.59704/b2a7a2ee0ff8bd31.

LICENSED UNDER CC BY SA