The DSA's Industrial Model for Content Moderation

Daphne Keller

24.02.2022 doi: 10.17176/20220224-121133-0

In "On Exactitude in Science," Jorge Luis Borges describes a map of such exquisite detail, it must be as large as the territory it depicts. Current policy proposals about platform content moderation – including legal rules that are close to being finalized in the EU's Digital Services Act – keep reminding me of that impossible map. They seem built on the hope that with enough rules and procedures in place, governance of messy, organic human behavior can become systematized, calculable, and predictable. In this model, the regulation of human behavior can, as György Lukács put it in the 1920s, "assume the characteristics of the factory." And while even the most far-reaching regulatory systems of the 20th century could police only a tiny fraction of human activity, regulation-by-platform today is different. It can, particularly when coupled with automated tools that examine every word we post, become pervasive in ordinary people's daily lives.

The factory model of content moderation originated with giant platforms like Facebook and YouTube, which process human expression at scale by breaking down moderation into refined, rationalized, bureaucratized, and often automated steps. The Digital Services Act (DSA) expands on this. Under its provisions, even the tiniest platforms must publish granular rules classifying permitted and prohibited speech before moderating user content, and issue detailed explanations any time they remove a post. And all but the smallest platforms must follow detailed procedural steps, including participating in repeated appeals and documenting their actions in public transparency reports and in case-by-case filings with the European Commission.

The goal of all this process is a good one: to protect Internet users. I have long been an advocate for procedural improvements in content moderation for this very reason, and I am generally a fan of the DSA. I praised the Commission's draft in this 2021 oped, and I think the European Parliament's version is even better. But I also believe we should be wary of locking future generations into systems that have failed so often and spectacularly in our own time. And I expect that in many real-world cases, the process prescribed by the DSA will waste resources that could better be spent elsewhere, and burden smaller platforms to a degree that effectively sacrifices competition and pluralism goals in the name of content regulation. There is a difference between procedural rules that legitimately protect fundamental rights and the exhaustive processes that might exist in a hyper-rationalized, industrial model of content moderation. The line between the two is not always clear. But I think the DSA often crosses it.

Lawmakers in DSA trilogue negotiations have a number of remaining opportunities to keep this tendency in check. It is possible to preserve procedural protections where they do the most good for fundamental rights, while resisting those that will likely do more to entrench incumbent platforms and lock in today's models of governance than to help users. In the final section of this post, I will list the Articles that I think present these opportunities.

Situating the DSA in the Evolution of Platform Regulation: From Outcomes to Processes

The DSA is in many ways a huge step forward in policy responses to Internet communications technologies. It is far from the "cyberutopianism" of the 1990s, with its naïve expectation that all information dissemination was socially desirable. And the DSA substantially improves on equally naïve earlier techlash proposals, which imagined that platforms could simply delete all the bad things online without collateral damage to good things, including Internet users' fundamental rights. If Theresa May's demand that platforms instantly identify and purge prohibited content envisioned the kind of task once given to Hercules (clean

these stables!), the DSA instead sees the kind once assigned to more workaday heroines like Psyche or Cinderella (sort this storehouse full of grain!)

The DSA sets out a detailed process for all that sorting. The details, and figuring out which platforms have what obligations, can be complicated. A rough chart listing who must do what can be found here. In the Commission's draft, the most burdensome duties fall on platforms ranging in size from as large as Facebook (which employs some 30,000 people just for content moderation) to as small as fifty people. Procedural obligations are triggered each time these platforms disable access to an item of content. In YouTube's case, for example, the process would be carried out some four *billion* times a year for comments alone – before even getting to video content moderation.

The DSA's mandated process includes notice, appeals, out-of-court dispute resolution, and formal reporting for each item of content removed. Platforms must notify the affected user, conveying "a clear and specific statement of reasons" with multiple enumerated elements (Art. 15). The users who receive these notices can contest the decisions, seeking new review by the platform (Art. 17). They can also bring the dispute to new out-of-court adjudication bodies for another round of review, at platform expense (Art. 18). (Platforms pay their own fees and expenses regardless of the case's outcome. If they lose, they also pay the user's fees and expenses.) Platforms must report each one of these decisions to the Commission for inclusion in an ambitious new database (Art. 15), and publish aggregate data in transparency reports (Art 13 and elsewhere). The largest platforms must also maintain records for inspection by auditors, regulators, and potentially researchers (Arts. 28 & 31).

Is all this process and documentation worthwhile? Sometimes it will be. Certainly very consequential decisions like YouTube's erasure of Syrian Archive videos or the deplatforming of former U.S. President Donald Trump deserve this much attention – though perhaps that attention should come from actual courts. And all users may get better content moderation in the first place if platforms know that out-of-court dispute resolution bodies will check their work. That said, users – including spammers and trolls whose aim is to game the system and waste platform resources – file enormous numbers of groundless appeals. Appeals also tend to from more privileged members of society, like men. Purely as a matter of improving outcomes of content moderation, users might see more fair outcomes if platforms instead dedicated resources to reviewing random samples of takedown decisions. Or we might see better corrections if expert groups had access to information about takedowns, and could file challenges. User appeals are (ahem) appealing as an element of fair process. But they aren't the only or best tool in the toolkit. Nevertheless, the DSA locks them in as the primary mechanism for improving content moderation.

Platform Costs and Societal Tradeoffs

Importantly, all this process is expensive. Particularly for smaller platforms, it will require substantial new engineering and product design efforts, and hiring considerably more content moderators. The way to avoid that expense will be to moderate *less*. Platforms that remove only illegal content, and don't use their Terms of Service (TOS) to impose more expansive rules, will save a lot of money. So will platforms that choose not to employ automated content moderation tools like upload filters, since those will more frequently trigger the DSA's costly processes. Avoiding TOS removals and filters might be an improvement from the perspective of those concerned about users' expression and information rights. But they won't be improvements in the eyes of most platform users, or advertisers who want "brand-safe" environments. Economically, platforms that reduce or forego content moderation will be disadvantaged compared to their larger competitors.

One potential response to this concern is, "So what? Platforms should not be in the business of content moderation at all if they can't provide fair processes." That is a fair response in situations where lawmakers know and can rigorously prescribe desired outcomes. (As I will discuss below, the GDPR is an example.) But in the world of content moderation, we

want pluralism and diversity. We want a thousand flowers to bloom, and for users to have many choices of discursive environments. If firms that would have provided these alternatives go out of business, the harms for expression and access to information will be real. This is a situation where smaller platforms' rights to do business may well be aligned with Internet users' rights. (Felix Reda and Joschka Selinger describe such a situation with upload filters here.) Speakers and readers might benefit from fairer processes provided by very large incumbents like Facebook or YouTube. But I fear they might lose in the end if the DSA precludes evolution of other approaches – perhaps built on community moderation models like Reddit's, artisanal models like Medium's, crowdsourced models like Wikipedia's, or distributed and interoperable models like Mastodon's. And they will certainly lose if the result of locking in the industrial approach is an online landscape consisting mostly of current incumbents (with their relatively restrictive speech rules), a handful of miniscule platforms, and perhaps some mid-sized unmoderated free-for-all platforms designed to avoid DSA process obligations.

Another constituency that won't be happy if platforms save on costs by avoiding moderation will be among EU and Member State policymakers. In areas such as disinformation or terrorist content, many lawmakers have supported practices the DSA disincentivizes, including broad TOS-based speech rules and content filters. In copyright law, policymakers have gone so far as to effectively mandate the use of filters. This creates a public policy environment that is, to put it mildly, complicated. The mid-sized platforms least able to afford DSA compliance may find themselves between a rock and a hard place: On the one hand pressured to "do more" to moderate harmful content, and on the other hand faced with expensive new requirements when they do so.

Quality and Quantity of Content Moderation

There is an inevitable tension between quantity and quality in content moderation. Longstanding calls for better notice and action processes from academics and many civil society groups were initially, I think, intended for a world of relatively rare removals, buttressed with careful processes. The DSA gives us more careful processes, but in a world where removals have become constant and pervasive. The process that users get for platform speech adjudication in the DSA is still not like the process they would get from courts, though. Courts are designed to deliver low-volume, high-quality justice. They can take time to do a good job because so many claims are winnowed out by prosecutorial discretion, friction in the system (like filing fees and formalities), and rulings that dismiss cases without assessing their merits. Judicial systems are also broadly built on societies accepting imperfect punishment of the guilty, in exchange for fairer processes for everyone else.

Platforms, by contrast, are generally known for low-quality and high-volume dispute resolution. The DSA seeks to increase the quality side, while leaving the volume the same or higher. Platforms, unlike courts, will have to provide full process for any claim that someone thought about long enough to click a few buttons. But high-quality, high-quantity adjudication of disputes about speech will be hard to achieve.

This is in part for the reasons of expense and resourcing that I have described for smaller companies operating under the DSA. But it is also for more fundamental reasons. The evolving, organic, human sprawl of online speech and behavior is very hard to reconcile with bureaucratized administration. A certain amount of mess and error, perhaps a large amount, will likely be ineradicable. And a system that does come close to eradicating organic complexity and mess may create the problem that Lukács predicted. By viewing novel situations only through predefined categories and mechanisms, it risks becoming so rationalized as to create a "methodological barrier to understanding" the real human world it seeks to govern.

Comparing Standardization in the GDPR and DSA

A comparison and contrast with another ambitious legal instrument, the General Data Protection Regulation (GDPR), may be instructive. Like the GDPR, the DSA builds out a compliance model with prescriptive rules, procedural and operational details, and govern-

mental or regulatory oversight. That model is a good fit for the GDPR's data processing obligations, since they can largely be defined and operationalized *ex ante*, and applied consistently. It's typically not hard to say whether a data controller should delete an item of personal data, for example.

The same does not go for much of content moderation. Judicial interpretation and enforcement of real laws governing expression takes a long time for this reason. Platform TOS rules may be simpler than laws (or not). But no amount of TOS-drafting or process-building can tell a platform what to do when unforeseen questions arise. There will rarely be clear answers the first time moderators encounter ambiguous material like Tiktok's Devious Licks videos or the initial jokes that evolved into the Tide Pod Challenge; read innocuous-sounding words in the rapidly evolving jargon of hate groups, or hateful slurs that are reclaimed as terms of pride by marginalized communities; or confront speech governance questions that platforms simply didn't anticipate or draft rules for. Initial platform reviewers, the teams administering appeals, and the providers of out-of-court dispute resolution under the DSA may all legitimately reach different conclusions. So might dispute resolution providers interpreting the same platform rules in different Member States.

Hard-to-answer questions like these arise regularly, just like novel legal disputes do. When it comes to speech, humor, and being terrible to one another, humans are endlessly inventive. It is this human element that prevents content moderation, even with the addition of the DSA's intensive process, from becoming what Lukács called a "closed system applicable to all possible and imaginable cases." There will be no future, platform-employed judge who, like the one hypothesized by Max Weber,

is more or less an automatic statute-dispensing machine in which you insert the files together with the necessary costs and dues at the top, whereupon he will eject the judgment together with the more or less cogent reasons for it at the bottom: that is to say, where the judge's behaviour is on the whole predictable.

A more recent thinker, James Scott, uses the concept of "legibility" – the standardization and collection of information that shapes governance – to describe a similar problem with overly engineered systems of governance. Under his framing, too, the DSA stands in striking contrast to laws like the GDPR. When the GDPR was drafted, data protection experts already had a shared vocabulary, and experience with the real-world gaps or disputes that emerged over decades under the previous Data Protection Directive. The problems they sought to address were comparatively legible, in Scott's sense. The DSA, by contrast, builds on the relative blank slate of prior law that is the eCommerce Directive. Its elaborate rules reflect careful consultation and thought, by both drafters and outside contributors from civil society and academia. But they do not reflect the kind of repeated experience with real world problems, and prior attempted solutions, that informed the GDPR.

Scott's analysis, and that of Weber, Lukács, and their many intellectual heirs, would counsel caution – and ideally room for correction and iteration – in devising new rules for complex and evolving systems like the Internet. Practical experience with Internet content moderation's many failings points in the same direction. Both should inform our thinking about tradeoffs under the DSA and any platform regulation.

Consequences for Trilogue Discussions

All of this is to say that where the DSA can be flexible, it should be. Where lawmakers can still defer setting hard and fast rules until after we all see how this goes, they probably should. The DSA already, unavoidably, prescribes a map the size of tremendously large territory. But there are still choices to make in provisions that remain open for negotiation. For the specific process and burden-related issues I have raised here, these opportunities are listed below. (For other issues involving fundamental rights, especially as they relate to privacy and law enforcement reporting, important recommendations come from Access Now, EDRi, CDT, and EFF. In a few cases, there may be tensions between the administrability goals this post focuses on and the ones centered in those recommendations.) The Article,

Recital, and Amendment references in the list below are not exhaustive, similar issues may arise in other portions of the DSA drafts.

- Micro, Small, and Medium-Sized Enterprises: The best hope for future innovation, competition, pluralism, and improvement of the online information ecosystem comes from modestly sized platforms. These make up the vast majority of entities regulated by the DSA, but possess only a tiny fraction of the compliance resources held by giant incumbents. The Commission's draft excuses micro and small entities from some obligations. The Parliament's draft wisely expands this, providing a process for non-profits and medium-sized firms to seek waiver of some duties. (This would cover firms with under 250 staff/EUR 50 million turnover/EUR 43 million balance sheet. A chart listing which DSA obligation apply to different kinds of firms is here.) I wish that still more options for more flexible and proportionate standards were on the table. But the Parliament's amendment is still a major improvement. (Art. 16, Parliament draft, Am. 259a).
- Spam and Bad Actors: DSA provisions that provide important rights to good-faith users will put weapons in the hands of abusive ones. This includes serious criminals who are subjects of legitimate police investigations, as well as purveyors of commercial spam and coordinated inauthentic information. By filing frivolous appeals, trolls and spammers can effectively impose financial penalties on platforms that enforce user-protective rules, or even strongarm those platforms into changing their policies. The Parliament draft addresses this risk in amendments that reduce some of platforms' obligations to engage with bad actors. The Parliament's language of Recital 42, permitting platforms to not notify in cases of content that is "deceptive or part of high-volume of commercial content", is importantly broader than that of Article 15, which perhaps inadvertently applies only to the purely commercial category of "deceptive high volume commercial content" (Art. 15.1, Am. 247; R. 42, Am. 53). Without this flexibility in responding to spammers, platforms may also resort to responses that affect all users' fundamental rights, including requiring real ID verification for accounts.
  At the same time, the Parliament draft introduces new and, in my opinion, problematic requirements that platforms process appeals on specified timelines and provide "human interlocutors" where "necessary" (Art 17.3, Am. 266; Art. 17.5, Am. 268). Article 20, which provides for suspension of both abusive users (who post illegal content) and abusive appellants (who deliberately file groundless appeals) somewhat mitigates this problem. But all three drafts are restrictive in ways that may leave platforms without appropriate flexibility in response to bad actors' creative and iterative tactics (for example, by being ambiguous about whether platforms can make suspensions permanent or use them against trolls who post barely-legal but abusive content).
- Tailoring duties based on technical function: The Internet intermediaries that can respond most precisely to unlawful content (by taking it down without affecting too much other content) and provide the best procedural protections to the affected user (because they are best able to contact her) are the ones furthest "up the stack." This includes user-facing providers like Facebook or YouTube. In most cases, it does not include infrastructure providers or back-end hosts like Amazon Web Services. The Commission draft included important language addressing this issue, which is usefully expanded in the Parliament draft's version of Recitals 13 and 26 (R. 13, Am. 25; R. 26, Am. 37). The Parliament's draft also makes helpful clarifications in Recitals 40a and 40b, as well as amendments to Articles 8 and 14 (R. 40a, Am. 51a; R, 40b, Am. 51b; Art. 8.2, point (cb), Am. 199b; Art. 14.6, Am. 245). These amendments can provide better fundamental rights protections, while at the

same time avoiding wasteful proliferation of unnecessary processes.

- Doubling the volume of disputes: Well-intentioned amendments in the Council draft would have the effect of doubling platforms' potential procedural burdens in many cases. In Articles 17 and 18, that draft lets notifiers, as well as accused users, invoke appeals and out-of-court dispute settlement measures. Since platforms' documented patterns of errors tend toward over-removal more than under-removal – a pattern which appeal rights were intended to correct – this substantial expansion does not seem justified (Art. 17.1, Am. 261; Art. 17.3, Am. 266; Art. 18.1, Am. 270; Art. 18.2(a), Am. 273; Art. 18.3, Am. 279).
- Content Filters: At both the Member State and Commission level, platforms have faced pressure to preemptively monitor and police every item of content users post. One likely consequence of such filtering efforts – whether deployed voluntarily or in response to pressure from audit recommendations, Codes, or back-room communications from governments – would be to vastly increase the number of triggers for the procedural steps discussed in this post. This problem provides a second set of reasons – beyond the more obvious fundamental rights concerns – to embrace the Parliament's clarifications about monitoring (Art. 7.1, Am. 190; Art. 7.1a, Am. 190a; Art. 27.3a, Am. 361a; Art. 35.1, Am. 430; R. 28, Am. 39; and elsewhere).
- Demotions: Some drafts of the DSA could potentially trigger weighty process obligations not only when platforms take action against *particular* items of content, but when content is demoted for any reason (Art. 15.1, Parl. Am. 247). Since any global change to platform ranking algorithms might "demote" huge portions of content on the platform (while promoting others), this implication should be avoided. The Parliament's draft creates this risk of interpretation in Article 15. The problem is not, I think, addressed by that Article's use of the word "specific."

The choices I recommend here would increase flexibility, and the viability of smaller platform business models, at the expense of simplicity, consistency, and predictability. For a system as diverse, complex, and evolving as the Internet, that direction of change seems only justified. We should not lock in expectations that were built around the practices of today's giants, at the expense of potential future innovation and improvements.

There are other improvements that I think would be valuable, but which may or may not fit into the remaining unresolved Articles in trilogue. For example, the benefits of diverse and pluralistic content standards and moderation practices could be explicitly recognized and supported in the Article 34 Standards or Article 35 Codes of Conduct. (Perhaps the "fit for purpose" language in the Parliament's draft of Art. 35 somehow supports this?) Or Digital Services Coordinators or the Commission could invest more time upfront approving bespoke obligations for small or mid-sized platforms, in order to reduce wasted effort, competitive harms, or simple noncompliance over the years to come. (That one is probably a pipe dream.)

Conclusion

The DSA's content moderation provisions are predicated on breaking down human behavior and its governance into rationalized, bureaucratized, calculable components. So are large platforms' existing content moderation practices. But while the latter accept a fair amount of mess and error, the DSA seeks more consistent and foreseeable outcomes, to be achieved through detailed processes and documentation. It is very hard to predict how much of the indeterminacy or mess will ultimately be reduceable, within systems that seek to govern such unprecedented swathes of organic and evolving human behavior. But there is reason for concern that unduly rigid and prescribed systems will backfire. And it is almost certain that smaller platforms will have difficulty shouldering the DSA's burdens. In pursuit of legitimate content moderation goals, the DSA may inadvertently sacrifice competition

goals, and foreclose future diversity in platform practices and speech rules.

Note: I was previously a senior lawyer for Google, and currently consult with smaller (but still presumptively Very Large, for DSA purposes) companies including Pinterest. This post is not intended to be about those companies or to reflect their positions or interests.

ABOUT THE AUTHOR

Daphne Keller

Daphne Keller is the Director of Platform Regulation at Stanford Law School's Program in Law, Science & Technology.Daphne Keller is the Director of Platform Regulation at Stanford Law School's Program in Law, Science & Technology.

SUGGESTED CITATION

Keller, Daphne: The DSA's Industrial Model for Content Moderation, Verfassungsblog, 24.02.2022, doi: 10.17176/20220224-121133-0.

Licensed under CC BY-SA 4.0.

Langzeitarchiviert bei IntRechtDok unter: doi.org/10.17176/20220224-121133-0.