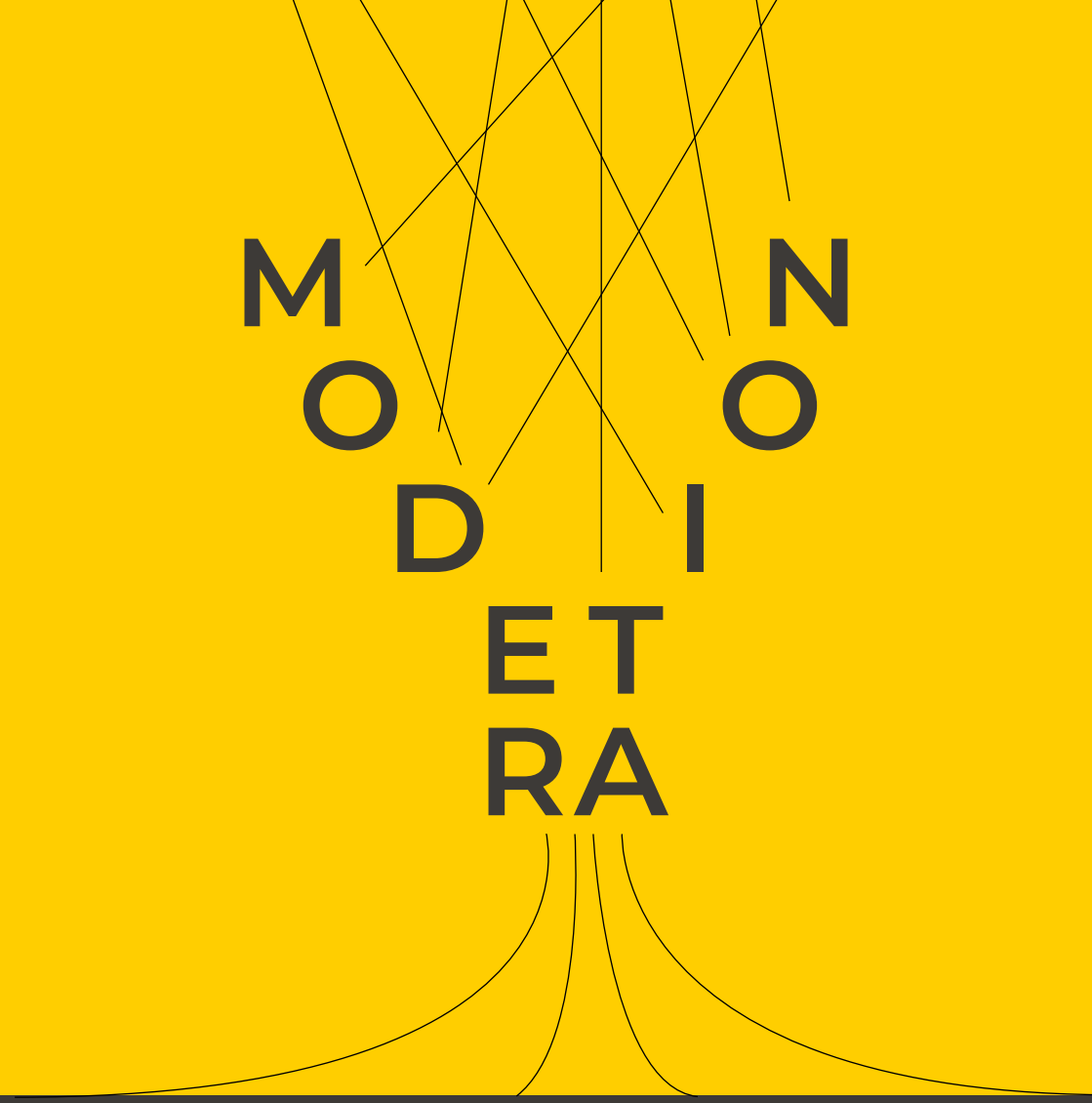


POLICIES, INSTITUTIONS AND DEMOCRACY
JUNE 2020

MODERATING OUR (DIS)CONTENT: RENEWING THE REGULATORY APPROACH

RENAISSANCE numerique

TABLE OF CONTENTS

KEY TAKEAWAYS

- Public policy tends to consider only a handful of platforms in current efforts to regulate toxic online content. Policy discussions at present also fail to grasp the idiosyncrasies and interconnectedness of content moderation across different platforms. The result is that policy response only addresses a piece of the problem.

- Content moderation is not just about removal, but about finding the right balance, positioning, and process, along with policy makers, civil society, and end users themselves. Moderation must be examined in a broader sense, beyond simply the number of contents removed.

- Today, co-regulation remains a bilateral process between the major platform operators and governments. If regulation introduces responsibilities and obligations across the board which are calibrated for the world's largest internet companies, these measures will have disproportionate negative impacts on other actors, which could in turn further reduce the diversity of platforms. Regulatory frameworks must be careful not to further reduce the diversity of platforms available to support a wide range of online expression.

- A new regulatory approach is called for, one that accounts for diverse moderation approaches and protects fundamental rights. We need agile indicators that let us measure the responsiveness of platforms to the real moderation challenges they face, challenges which evolve.

- Nuance, agility, and broad stakeholder and end user participation are necessary to this new regulatory approach. A central question for many is how to look beyond the concept of user threshold, or the number of in-country users on a platform. This concept is inapt, as this figure alone does not illustrate the moderation challenges faced by the platform. Renaissance Numérique advocates for a more process-oriented assessment of platform moderation performance.

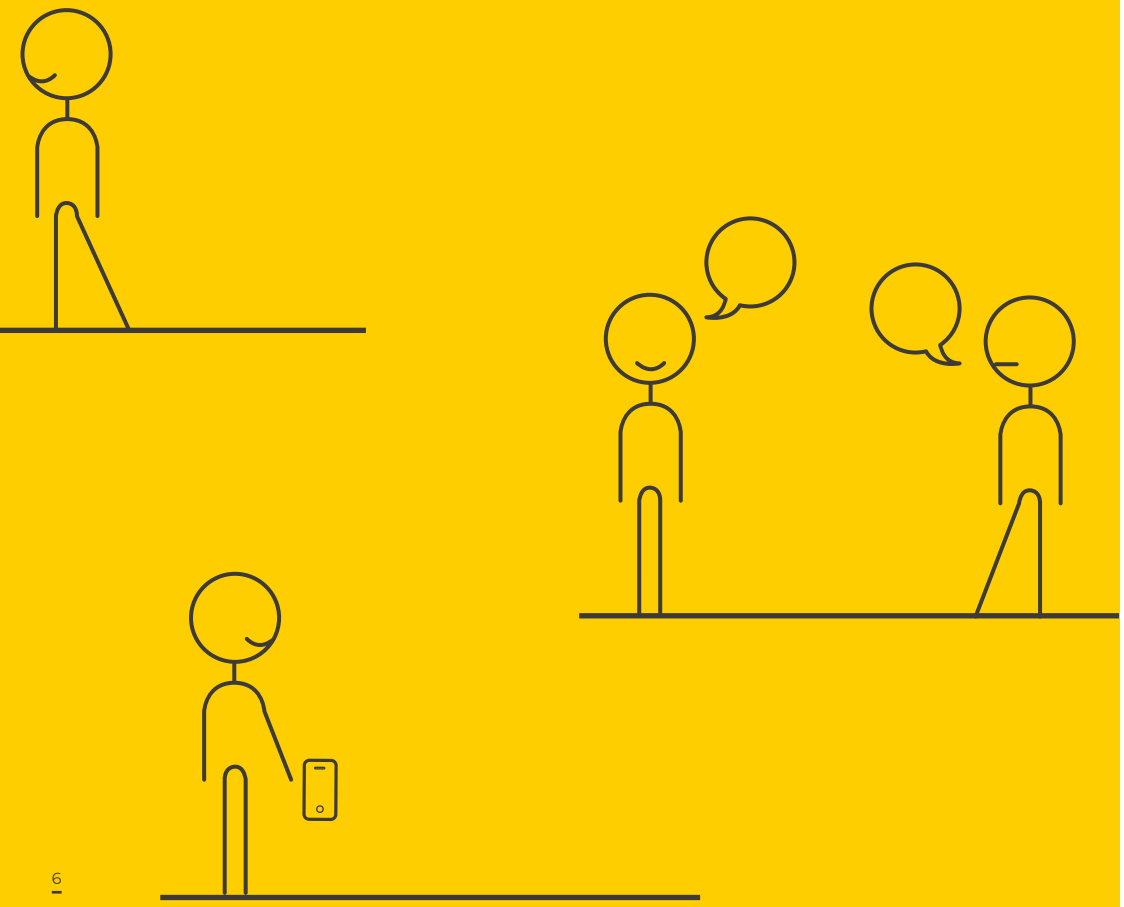
- Inherent to user-generated content-hosting platforms is the notion of the co-creation of value. The substantial contribution of end users must be reflected in platform governance of content moderation. A collaborative approach requires genuine discursive processes with end users, not just the outsourcing of moderation labor.

- Governance structures are needed to facilitate this participation. This kind of user involvement must be part of a broader behavioral shift on online platforms, reframing of the end user as agent.

- Public authorities should reinforce the capacities of all stakeholders to allow for functional collaboration and discursive processes. It is the responsibility of public authorities to establish a general framework to facilitate intra- and inter-sectoral collaboration and knowledge sharing, to work with civil society, researchers and technical experts to find effective methods, and to share these methods with all actors and across all platforms.

- Future regulation in this space, in particular the European Digital Services Act, must not simply be shaped for and by the most dominant platform operators. Regulation must aim to address content moderation holistically, across all relevant services.

INTRODUCTION
BROADENING THE PERSPECTIVE OF PUBLIC POLICY



Harmful online content and the question of how to address it has existed since the origins of the surface web[1]. Over the years, as online spaces have become central to our democracies, this problem has magnified to a point where toxic content now threatens the free flow of information and the enjoyment of our fundamental rights. However, at present, public policy tends to consider only a handful of platform operators in current efforts to regulate toxic online content - whether it be hate speech, cyberbullying, disinformation, etc. The result of this trend is that policy response only addresses a piece of the problem. Rather than observe the full range of platform operators that host toxic content, along with the important interconnections and spillover effects between these platforms, the gaze of public policy remains focused on a select group of actors - notably Facebook, YouTube, Twitter. In Germany, the *Netzwerkdurchsetzungsgesetz* or NetzDG, the country's law to combat hate speech on the Internet, took clear aim at these three: in preparation for the law, the German Minister of Justice, Heiko Maas, formed a working group to meet specifically with Google, Facebook and Twitter[2]. In France, Laetitia Avia, LREM deputy of the 8th district of Paris and spokesperson of France's analogous law against online hate content has stated that the law is intended to address *"a handful of actors"*[3][4]. Despite this policy preoccupation with a small group of platforms, all operators hosting user-generated content are faced with the challenge of toxic content, and must be considered intelligently in the formulation of regulation. Regulation that does not recognize the diversity of platforms and the relationships among them does not only compromise its ultimate

1 The surface web refers to the area of the World Wide Web that is accessible to the general public and indexable by search engines.

2 The NetzDG is often referred to as "The Facebook Law". William Echikson et Olivia Knodt, "Germany's NetzDG: A key test for combating online hate", *CEPS Research Report*, November 2018. Available online: https://bit.ly/2ZaX6Nx

3 Conference on « Les réseaux de la haine » ("*Networks of hatred*"), January 28, 2020, École militaire, Paris. The actors are not yet determined at this time of writing in June 2020. This and other specifications will be determined by decrees that will be drafted in the coming months, as is the custom.

4 After it was seized by a group of French senators, the French Constitutional Council found the Avia law substantially unconstitutional.
Décision n° 2020-801 DC du 18 juin 2020, *Loi visant à lutter contre les contenus haineux sur internet*: Available Online: https://www.conseil-constitutionnel.fr/decision/2020/2020801DC.htm

effectiveness, it risks disproportionately harming certain platforms, along with the quality and quantity of the online spaces available to users.

Admittedly, this policy attention on a handful of large platforms can be explained by their oligopolistic nature and their role in structuring digital public space. But regulatory decisions can have a disproportionate impact on those actors less able to meet these new regulatory requirements — those with less capacity in terms of human resources, technical or ergonomic features, etc. Additionally, these new requirements fail to sufficiently account for the idiosyncrasies of different platforms, and thereby only respond to a part of the problem that the regulation aims to address. Meanwhile, the automated, algorithmic moderation of toxic content — which is becoming the *de facto* requirement for platforms and has seen a surge in recent months during the Covid-19 health crisis[5] — is far from a miracle solution[6]. A single regulatory approach has the potential to unintentionally catch up in its net other platforms beyond those first intended. These challenges are political as well as technical and financial. If new regulations introduce new responsibilities and obligations across the board which are calibrated for the world's largest internet companies, these measures will have disproportionate negative impacts on other actors, which could in turn further reduce the diversity of platforms available to support a wide range of online expression. At the same time, these larger, in a sense "incumbent" internet companies, remain better positioned to adapt to regulations and meet compliance obligations[7]. In order to build effective methods to address the problem of toxic content in our online spaces — to address the challenges of moderation across a deeply fragmented online landscape — we need to look beyond the major platform operators that currently shape public debate. Nuance, agility, and broad stakeholder and end user participation are necessary to a new regulatory approach.

The tools and uses of digital platforms are constantly evolving, and with them, trends in the nature and propagation of toxic content. Not only does toxic expression change in both online and offline spaces (dialect, symbols, codes), but the platforms in use and the activities on these platforms also morph with the emergence of new tools, features and trends. Meme culture[8] is an important example, one that is rarely evoked in policy conversations. In this way, social and cultural changes are layered onto technical evolutions, leading to the rapid mutation of toxic content[9]. This is just one edge of the problem. Even what constitutes toxic content is subject to continual debate and development by platforms, policy makers, researchers, and of course users across the world. There is not a shared definition — and certainly not a shared, operationalized definition — of what constitutes toxic or harmful content across ugc (user-generated content) -hosting platforms. We have seen the definitions of "harm" broaden most recently during the Covid-19 pandemic[10]. The purpose of this note is not to try to further define toxic content, whose very nature may in fact be porous[11], but rather to examine the mechanisms of its moderation and propose pathways for improvement. We remain focused here, necessarily, on content prohibited by law and by the Terms of Service of platform operators (including: CSAM [Child Sexual Assault Material], terrorist content, misinformation, threats and cyber har-

5 Marc Faddoul, "COVID-19 is triggering a massive experiment in algorithmic content moderation" *Brookings*, April 28, 2020: https://www.brookings.edu/techstream/covid-19-is-triggering-a-massive-experiment-in-algorithmic-content-moderation/

6 Hannah Bloch-Wehba, "Automation in Moderation", *Cornell International Law Journal*, Forthcoming, last revision: April 29, 2020: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3521619

7 Nine Principles for Future EU Policymaking on Intermediary Liability, *Center for Democracy and Technology*, April 2020. Available online: https://cdt.org/wp-content/uploads/2019/08/Nine-Principles-for-Future-EU-Policymaking-on-Intermediary-Liability-Aug-2019.pdf

8 Because memes rely on adapted images, symbols, and irony, they are particularly challenging for content moderation. Facebook AI Research has launched a so-called "Hateful Memes data set" of 10,000 memes scraped from public Facebook groups in the U.S.
"Facebook is using more AI to detect hate speech", *Venture Beat*, May 12 2020:
https://venturebeat.com/2020/05/12/facebook-is-using-more-ai-to-detect-hate-speech/

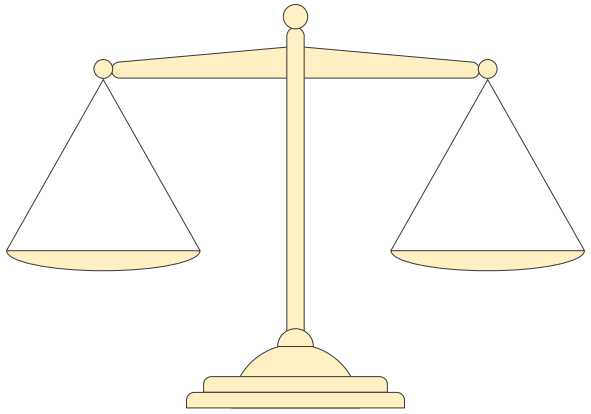
9 Of course, advanced features and platform evolutions are not necessary to propagating toxicity "The Hottest Chat App for Teens Is … Google Docs", *The Atlantic*, March 14, 2019: https://www.theatlantic.com/technology/archive/2019/03/hottest-chat-app-teens-google-docs/584857/

10 Evelyn Douek, "COVID-19 and Social Media Content Moderation", *Lawfare*, March 25, 2020: https://www.lawfareblog.com/covid-19-and-social-media-content-moderation

11 Renaissance Numérique (2018), "Fake News, Faire face aux troubles informationnels à l'ère numérique", Available online: https://www.renaissancenumerique.org/system/attach_files/files/000/000/155/original/RN-NOTE_FAKE_NEWS_23mars2018.pdf?1521799239; Renaissance Numérique (2017), "Taking action against hate on the internet in a collaborative society", available online: https://www.renaissancenumerique.org/ckeditor_assets/attachments/210/note_finale_seriously_en.pdf

assment, hateful/defamatory/discriminatory content, etc.). These toxic contents are in perpetual evolution both in their performance (how they propagate online) and their definitions (how they are perceived), multiplying the challenges for regulators and for platform operators. Regulatory and moderation practices seek to evolve quickly in response, to keep pace with the problem, but hasty regulation presents its own risks[12].

The European Commission's upcoming Digital Services Act revising the e-Commerce Directive of 2000 is an opportunity to *"upgrade the Union's liability and safety rules for digital platforms, services and products"* in the words of Commission President Ursula von der Leyen[13]. This text aims to frame a more responsible governance of our digital environments, central to which is the issue of content moderation. This publication seeks to nourish the Commission's reflections on this topic.



12 We are witnessing technical evolutions and social adoptions of new technology at an unprecedented speed. The adoption of many of these digital platforms has taken place over just a few years, while the adoption of previous technologies such as telephones or electricity has taken place over decades. This slower adoption allowed regulators, industry and society to develop behaviours and moderation practices gradually.

13 The European Commission "Political Guidelines for the next European Commission: 2019-2024": https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf

ONLINE MODERATION IS NOT JUST CONTENT REMOVAL

Moderation challenges are not limited to the ability of platform operators to recognize toxic content. Moderation also relates to the decisions made in curating and promoting content and in inviting and discouraging certain behavior. Many platforms employ graduated measures, for example, down-ranking and reducing the visibility of content, putting content in "quarantine", dereferencing content, adding a label, alert or supplementary/qualifying information, or even cautioning users before the publication of content (ie: "nudging"). Moderation can also involve closing groups, suspending accounts, and banning users from the platform. **Platforms are innovating considerably in this area, and are still arguably at the very beginning of their capacity to do so[14].**

In fact, the mere removal of content can circumvent or even aggravate the problem that was created by that content. The removal of one post can produce a hydra-like response, motivating several others posts — by users who dispute the removal of their content, or by users who sympathize with the original post, etc. Often, the ensuing content will be even harder to moderate because it is written in a way that respects the Terms of Service just enough to stay within the bounds of the permissibility or to otherwise circumvent detection. It can also happen that an original instigating comment is not toxic or in violation, but ensuing comments that refer to it or evoke it are toxic, thus leaving platform moderators the difficulty of deciding whether to remove permissible content. Moderation by removal also raises the risk that the deletion of content can have the opposite of the desired effect: giving it legitimacy with certain *"communities of interest"* and inspiring its propagation elsewhere. This is often the case with conspiracy theories, whose dis-

14 Twitter is currently testing new settings to limit unwanted replies, see: https://blog.twitter.com/en_us/topics/product/2020/testing-new-conversation-settings.html

appearance can serve to reinforce the claim of the conspiracy theory[15].

In truth, removal will always be relative, as it is not possible to scrub content entirely from the Internet. After the Christchurch attacks, for example, thousands of people re-watched the video of the shooting without reporting it, mostly on so-called "alternative" networks. And though the GIFCT (Global Internet Forum to Counter Terrorism[16]) hash-sharing consortium ultimately reduced the spread of footage of the Halle shooting, it could not entirely erase it[17].

Finally, for platforms and for users, there can be damage caused by both removal and non-removal. Platforms attest to receiving backlash from both sides — when they take down content and when they leave it up — including from authoritative civil society actors like NGOs who take issue with certain decisions. Platform representatives explain that they do not wish to be the judge or guarantor of freedom of expression[18], deciders of what can appear online or not. And yet decisions must be made. **Their moderation work is therefore not just about removal, but about finding the right balance, positioning, and process, along with policy makers, civil society, and end users themselves.**

Above all, it must be remembered that removal of content constitutes a very strong, in a sense violent political act with respect to the freedom of expression, a fundamental right. The French *Conseil d'État* reiterated this recently in the context of the Avia Law, remarking that it is a "particularly radical" act[19]. Indeed, after a group of French senators brought the case to France's highest court, the *Conseil Constitutionnel*, the court found that several central articles of the law, but particularly the first article, infringed upon the exercise of freedom of expression and communication in a manner that is unnecessary, inappropriate and disproportionate[20].

The regulation of content moderation requires a grasp of tools and practices in constant motion, an understanding of a diverse and fragmented online landscape, and a sensitivity to the reality that content moderation can be neither neutral nor perfect. But one particular question emerges from these overlapping challenges, to which this note seeks to respond: how can we integrate all platforms into the moderation debate and facilitate moderation practices that accord with our fundamental rights and freedoms?

15 Sam Levin, "'Taking them down fuels it more': why conspiracy theories are unstoppable", *The Guardian*, February 28, 2018: https://www.theguardian.com/us-news/2018/feb/28/florida-shooting-conspiracy-theories-youtube-takedown

16 The GIFCT (the Global Internet Forum to Counter Terrorism) was established in 2017 by Facebook, Microsoft, Twitter, and Youtube, as a way for the companies to share information about violent terrorist content in order to remove it across their platforms. The GIFCT has set up a database for sharing hashes (fingerprints) of identified terrorist content, to facilitate its removal. Pinterest, Dropbox, Amazon, LinkedIn, and WhatsApp have since joined the collective, among others. Membership is open to smaller platforms as well, and the initiative is making an effort to share their resources - notably in collaboration with the independent organization *Tech Against Terrorism*. Still, the GIFCT is often criticized for its lack of transparency and oversight.
See the website of the initiative: https://www.gifct.org/.

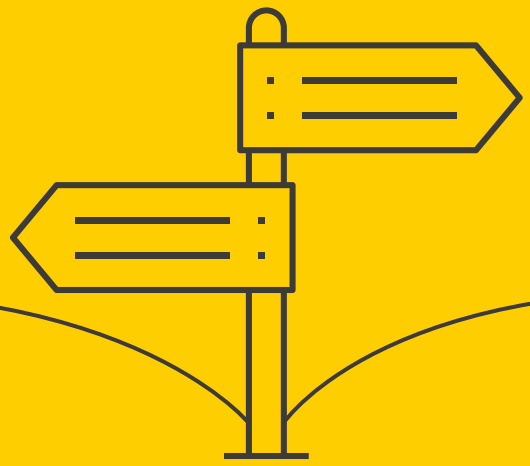
17 Renaissance Numérique, "Un crime répété, et pourtant : qu'est-ce qui a changé dans notre réponse au terrorisme lié à internet ? *blog.seriously.ong*, October 12, 2019: http://blog.seriously.ong/un-crime-copie-et-pourtant-quest-ce-qui-a-change-dans-notre-reponse-au-terrorisme-lie-a-internet/

18 Watch CNBC's full interview with Facebook CEO Mark Zuckerberg from May 28, 2020: https://www.cnbc.com/video/2020/05/28/watch-cnbcs-full-interview-with-facebook-ceo-mark-zuckerberg.html

19 Renaissance Numérique (2019), "Lettre ouverte collective appelant à garantir nos libertés publiques dans la proposition de loi visant à lutter contre la haine sur Internet", Available online: https://www.renaissancenumerique.org/system/attach_files/files/000/000/199/original/lettre_ouverte_relative_a_la_proposition_de_loi_visant_a_lutter_contre_la_haine_sur_internet.pdf?1569397597

20 Décision n° 2020-801 DC du 18 juin 2020, *Loi visant à lutter contre les contenus haineux sur internet*: Available Online: https://www.conseil-constitutionnel.fr/decision/2020/2020801DC.htm

PART I
MODERATION CHALLENGES ACROSS A FRAGMENTED ONLINE LANDSCAPE



WHY IT IS NECESSARY TO GRASP THE DIVERSITY OF ONLINE PLATFORMS

Although research[21] and public policy tend to focus on only a handful of actors, there is a wide variety of ugc-hosting platforms who face the challenge of toxic content and yet pass under the radar. Many of these platforms are neither small nor niche[22]. Some initiatives have aimed to address this gap; notably, the sCAN project, a collective of European civil society organisations, has examined alternative "safe havens" for hateful content by studying platforms like RK.com, Gab.ai, RuTube, Telegram, Disqus, Discordance, Spotify, Pinterest, and Tumblr.[23]. But broader, comparative work is still needed to nourish public policy. There are important differences between platforms, for example, the type of content hosted (text, video, live streaming, ephemeral content), the strategy of referencing and ordering content (including the role of artificial intelligence)[24], the services and functions offered (private chat, marketplace, etc.), the business model of the platform

21 This failing in the research can be partially attributed to the challenge of accessing data. For example, it is relatively easy to "scrape" data from Twitter, so many studies are based on Twitter, but with limited relevance beyond it. The fact that research is concentrated around a few platforms and not holistic in scope impedes the improvement of moderation strategies and regulation. Civil society in its broadest sense (including researchers), in collaboration with the platforms, need to conduct in-depth and comparative research across platforms.

22 Daniel Carnahan, "For the first time, LinkedIn included data on its moderation efforts in its biannual transparency report" *Business Insider*, November 25, 2019: https://www.businessinsider.fr/us/linkedin-releases-data-on-spam-scams-and-fake-account-removals-2019-11

23 "Beyond the 'Big Three', Alternative platforms for online hate speech", The EU-funded project sCAN– Platforms, Experts, Tools: Specialised Cyber-Activists Network (2018-2020), April 2019. Available online: https://www.voxpol.eu/download/report/Beyond-the-Big-Three-Alternative-platforms-for-online-hate-speech.pdf

24 In France, the report of the mission led by Benoît Loutrel and submitted to the Secretary of State in charge of digital technology, in order to inform the regulatory approach to social networks explains the need to focus particularly on content "accelerators", platforms that have the function of content scheduling and therefore "the capacity to accelerate the distribution of certain content, or on the contrary, to slow down its propagation". This is a relevant distinction, but this note does not limit itself to platform accelerators, as many other important moderation decisions can be made without this capacity.

(the presence of advertisements[25]), the size and geographic presence of the platform, etc. Most relevant for the purpose of this analysis are the different moderation methods through which platforms approach the problem of toxic content.

A VARIETY OF APPROACHES TO CONTENT MODERATION

The diversity of ugc-hosting platforms must be accounted for in their regulation. To convey this diversity, this analysis draws on the work of Robyn Caplan in her text *Content or Context Moderation?*[26], and on testimony shared by platform representatives during a seminar organized in February 2020 by Renaissance Numérique[27]. Caplan develops a theoretical framework distinguishing between three typologies: *industrial, artisanal and community-reliant*. These categories are necessarily fluid and many platform operators rely on hybrid strategies, or else alter their approaches over time as their services and user bases develop.

The following table is inspired from Caplan's work in *Content or Context Moderation?*, and has been simplified and adapted slightly.

25 The online advertising industry can be closely correlated with toxic content, often through financing webpages that host toxic content.
See Renaissance Numérique, "Brand safety dans l'écosystème de la publicité programmatique: Quel rapport entre les contenus haineux et les marques", *blog.seriously.ong*, December 2019: http://blog.seriously.ong/brand-safety-dans-lecosysteme-de-la-publicite-programmatique-quelle-rapport-entre-les-contenus-haineux-et-les-marques/
This tendency, particularly with automated programmatic advertising, inspired an amendment in the French Avia law referred to as "Follow the money", which would have required advertisers to make public at least once per year their advertising relationships (before the decision of the *Conseil Constitutionnel*).

26 Caplan, Robyn. "Content or Context Moderation?: Artisanal, Community-Reliant, and Industrial Approaches", *Data & Society*, 2018: https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf

27 Renaissance Numérique organized a seminar on the challenges of the moderation of toxic content entitled "How to integrate all platform operators in the moderation debate" on 14 February 2020. The event brought together platform representatives, members of civil society, researchers and members of French public institutions. See the Acknowledgements section for the full list of participants.

TABLE 1 - TYPOLOGIES OF MODERATION APPROACHES

Approach	Moderators	Users	Tools
Industrial	• may employ as many as tens of thousands of employees around the world; • many moderators are third parties or contractors; • moderation teams are separated from design and policy teams; • the system is standardized and formalized (factory-like).	• users can report toxic content (though modalities depend on the platform); • relationships with civil society through Trusted Flagger programs, fact-checking teams, etc.	• the vast majority of content is filtered by automatic tools; • some participate in hash sharing collectives (photoDNA, GIFCT).
Artisanal	• moderation teams range between about 5 and 200; • moderators exchange/coordinate with other teams; • there is room for discussion and decisions are made on a case-by-case basis ("manually").	• more time is taken per post and users are considered more holistically/within the history of their online activity; • the process for flagging content is similar to that of industrial platforms.	• limited use of artificial intelligence, most content is examined *ex post* (not filtered); • platforms may participate in hash-sharing collectives but in a passive/non-strategic role.
Community-reliant	• a multi-layered model with a core team of a few salaried staff and then degrees of volunteer participation and responsibility ("onion layers"); • some transversal policies, but pages/group establish their own rules and respective moderators are responsible for enforcing these rules ("federal"); • volunteer moderators are not remunerated.	• any user may become a moderator; • moderation responsibility can be increased over time; • user flagging varies; often, users can bring complains directly to moderators.	• less use of artificial intelligence though there are automated tools available to users and moderators to use as they see fit.

INDUSTRIAL: THE SPEED AND SCALE OF A DECISION-FACTORY

In industrial-style moderation, tens of thousands of employees apply rules set by a dedicated policy team. Notable examples are Facebook and YouTube. Teams increasingly rely on automated tools to report content such as hate speech. Much toxic or otherwise infringing content is removed *ex ante* by algorithms or filtering tools (for example, the digital fingerprinting system used by YouTube, *Content ID*). The amount of content identified *ex ante* depends on the platform and the type of content. For example, spam content is often identified at nearly 100%, and CSAM content can also be identified algorithmically at a high percentage. Facebook — which currently employs over 30,000 moderators — published that they identified 80% of hate speech content automatically in their most recent cycle[28] (up from 38% in 2018[29], which may raise red flags among experts). Automation is a fundamental hallmark of the industrial model; without these technological capabilities, for example in the artisanal model, moderation can only be *ex post*. Essentially, moderation on these platforms is considered industrial due to: their size and number of users, the size of their moderation teams, their use of automation and *ex ante* algorithmic moderation, and the separation between their policies and the implementation of these policies within the company. Indeed, an important feature of the industrial approach is to separate the teams in charge of developing moderation policies from those in charge of their application, both at the organizational and geographic levels.

These companies tend to have more resources at their disposal and continue to expand their moderation capacity. They oftentimes began with artisanal approaches and then experimented over time, developing more formalized policies and systems. Caplan notes the simultaneous growth of moderation teams and the need to create a factory-like decision making apparatus: "*Complex concepts like harassment or hate speech are operationalized to make the application of these rules more consistent across the company*". Among its limits, by trying to create a compartmentalized decision machine, the industrial approach cannot fully grasp the context around the contents: this leads to both *false positives*, or the removal of legitimate content, and also to toxic content escaping detection.

ARTISANAL: REVIEWING CASE BY CASE

In the artisanal moderation method, moderation is normally carried out by a team of 5 to 200 employees. Decisions are often made case-by-case. A few examples are: Patreon, Change.org, Vimeo, Discord, Medium. Artisanal platforms include large online forums, websites for supporting content creators, file sharing services, etc. Caplan highlights that these platforms are one of the primary ways individuals around the world access the Internet. There is great methodological diversity even within the artisanal approach. Artisanal moderation teams are not only distinguished by their intimate size, but also by the fact that moderation is carried out internally, by employees instead of third-party services or contractors. Platform representatives also emphasize the limited use of automation and algorithms in content moderation. Companies and organizations (indeed not all of them are businesses) with artisanal approaches often taut a "manual" and thorough approach to moderation, and the ability to be more responsive to the context in which the content was published. They also claim to have fewer reports of toxic content, which allows them to devote the time for more meticulous review. Although some have millions of users, it bears noting that artisanal platforms do not always face the same mass of content as the larger industrial actors. Last, while accounting carefully for context, these platforms are limited in their ability to apply rules consistently and at scale.

28 See Facebook's most recent transparency reporting on Hate Speech: https://transparency.facebook.com/community-standards-enforcement#hate-speech

29 Facebook, "Hate Speech," Transparency.Facebook.com, (2018), https://transparency.facebook.com/community-standards-enforcement#hate-speech Retrieved July 31, 2018

CHANGE.ORG: EMPHASIZING DIALOGUE WITH USERS[30]

Change.org is a platform for online petitions. Change.org France has a team of about 20 people with developers, engineers, a product manager, and a campaign team that accompanies the petition-writers. It is a "*Certified B Corporation*", an American certification given to commercial companies meeting societal and environmental requirements.

Change.org does not rely solely or principally on artificial intelligence in content moderation. According to the director of Change.org France: "*firstly because we don't have the same resources as the big platforms and also because we see the risk this could imply, as we are a petition platform whose goal is to promote freedom of expression.*" Thus, their moderation is essentially *ex post*. Change.org holds small discussion groups internally to discuss moderation decisions, and they try to dialogue with content creators and with their user community. At the seminar organized by Renaissance Numérique on February 14, 2020, the director noted that this practice of dialogue is all the more essential when it comes to misinformation, because it is difficult to change one's mind as soon as the content has been shared.

Given the size of their platform, this practice of dialogue is not easily scalable and cannot be practiced across all petitions. As a U.S. company, their ability to respond to the French context remains more limited: for instance, when employees in the U.S. have to make decisions regarding French content with social/cultural subtleties.

30 Testimony of the director of Change.org France during the seminar held on February 14, 2020 by Renaissance Numérique.

PATREON: CONSIDERING THE PERSON BEHIND THE CONTENT

Patreon is a crowdsourcing financing platform with a team of six full-time employees in total, serving approximately 150,000 creators worldwide.

Patreon tauts a "thorough, manual process" of moderation that is context-sensitive. For example, they try to review all of an author's content to understand the author's intent. According to the platform, "*Creators on Patreon depend on us for their paycheck. This is a massive responsibility and one we take very seriously. For this reason, all decisions that impact creators' paychecks are made personally and case-by-case. No decision that impacts a creator's paycheck is automated — each case is always reviewed by a member of the Trust and Safety team.*"[31]

Their subscription-based business model is important when considering their moderation approach. As one representative explained: "*the value to the platform of each new user on a content hosting platform run by ads is lower compared to the value of each new Patreon creator with subscription payments*"[32]. However, as expressed in an interview with Caplan[33], representatives still have some concerns about the lack of resources: *"The reason why I think size is a useful thing to think about is it's a reflection of the resources available to that platform to actually comply with something.*"

31 "How Patreon moderates content", *blog.patreon.com*, July 25, 2019: https://blog.patreon.com/how-patreon-moderates-content

32 Colin Sullivan, head of legal at Patreon, "Trust Building As A Platform For Creative Businesses", *TechDirt*, February 9 2018: https://www.techdirt.com/articles/20180206/11024139169/trust-building-as-platform-creative-businesses.shtml

33 Caplan, Robyn. "Content or Context Moderation?: Artisanal, Community-Reliant, and Industrial Approaches", *Data & Society*, 2018: https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf

COMMUNITY-RELIANT: FEDERATED, VOLUNTEER MODERATION

The community-reliant approach to moderation usually combines formal policies developed by the platform with actions taken by volunteer moderators. The platforms usually establish a system through which a large number of volunteers implement and even enrich the general policy decisions made by a small core team of employees. Examples include: Wikipedia, Reddit, Mastodon. Most platform operators rely on their users to participate in the moderation process. This style of moderation relies heavily on volunteers who may have different levels of moderation authority. According to Caplan, the most important feature of this approach is the relationship between the parent organization and its volunteer moderators.

These platforms are often - though not always - decentralized structures. This brings strengths and weaknesses in terms of moderation. A representative of Reddit compared their model to a federal system with site-wide rules that must be respected by small sub-communities but can also be extended at the discretion of sub-community moderators[34]. Community-reliant platforms prioritize localized context-level decision making, often over consistency and uniformity across the platform as a whole. Permitting communities to follow their own rules arguably allows for greater sensitivity to context. However, disagreements may (and do) arise between sub-communities, and platforms with this federal moderation structure risk appearing incoherent and arbitrary in their enforcement. As recent events on Reddit illustrate, community-reliance can have its own, not negligible, transparency and accountability issues: a screenshot of a list of moderators shared in May of 2020 claimed that "92 of top 500 subreddits [Reddit group-pages] are controlled by just 4 people". David Pierce, Editor of Protocol, explains that this list is misleading because Reddit groups often have dozens of moderators, but he concedes that the controversy does reveal the underlying flaw that some people have too much power on the platform[35]. Reliance on volunteer labor from the user community has other serious drawbacks. These platforms are often criticized for their reliance on non-remunerated volunteers. Volunteer moderators can be exposed to the same kinds of horrific content[36] as paid moderators employed by industrial platforms, and they are also subject to harassment from users who disagree with their decisions, that can go as far as death threats[37]. Reddit may represent a far end of the spectrum when it comes to community-reliant moderation, with minimal to non-existent overarching moderation governance. Wikipedia[38] and Germany's Gutefrage.net[39] have shown more willingness to shape and structure community moderation through careful layers and policies[40].

34 David Pierce, Editor of Protocol, illustrates, "*Practically every subreddit, once it hits a certain size, develops its own rulebook. No two are alike: You can have a 'Game of Thrones' subreddit that doesn't allow memes, serious discussion only, and a competing one where memes flow like Dornish reds. Some are ruthless about formatting and style, others couldn't care less.*"
David Pierce, "Reddit does moderation differently — and it's ignited a war on the platform", *Protocol*, May 27, 2020: https://www.protocol.com/reddit-powermods-war?utm_campaign=The%20Interface&utm_medium=email&utm_source=Revue%20newsletter

35 David Pierce, "Reddit does moderation differently — and it's ignited a war on the platform", *Protocol*, May 27, 2020: https://www.protocol.com/reddit-powermods-war?utm_campaign=The%20Interface&utm_medium=email&utm_source=Revue%20newsletter

36 Casey Newton, "Facebook will pay $52 million in settlement with moderators who developed PTSD on the job", *The Verge*, May 12, 2020: https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health

37 David Pierce, "Reddit does moderation differently — and it's ignited a war on the platform", *Protocol*, May 27, 2020: https://www.protocol.com/reddit-powermods-war?utm_campaign=The%20Interface&utm_medium=email&utm_source=Revue%20newsletter

38 Volunteer moderation varies widely across platforms and communities, but Wikipedia's community is often cited as particularly mature. This was evidenced in the response of the platform and their community moderators to mis/disinformation during the Covid-19 health crisis. See:
Omer Benjakob, "On Wikipedia, a fight is raging over coronavirus disinformation", *Wired*, February 9, 2020: https://www.wired.co.uk/article/wikipedia-coronavirus

39 Gutefrage.net has a multi-layered system of moderation which aims to protect volunteer moderators from the traumatic work of reviewing the most toxic content. Employees of the company (professional, remunerated moderators) work with algorithms to address the most dangerous content. Volunteer moderators can address only the content that passes through and is published.
Dinar, Christina, "Digital Streetwork - Wie Online Padagogisch Agieren?", November 22, 2019: https://www.belltower.news/debate-dehate-digital-streetwork-wie-online-paedagogisch-agieren-93255/

40 Adi Robertson, "Wikimedia is writing new policies to fight Wikipedia harassment", *The Verge*, May 25, 2020:

WIKIPEDIA: VIRALITY IN REVERSE?

Wikipedia is financed by the Wikimedia Foundation, a U.S.-based charity, and is a voluntary and collaborative project largely reliant on donations. Wikipedia has 30 million unique visitors and is the fourth most visited site in France and the fifth most visited in the world[41].

Wikipedia is a hybrid system of automated tools and volunteer editors[42]. All users are invited to contribute and modify content on the site. The moderation model is therefore based on the volunteer 'civic' engagement of individuals. A representative of Wikimedia France noted that Wikipedia experiences the opposite of the negative effects from virality; virality in fact has a moderating effect, because the more an article is seen/shared/edited, the more it is moderated[43].

Wikipedia has a multi-layered moderation system that offers different levels of responsibility and encourages participants to gradually increase their involvement. Wikipedia facilitates relationships between its moderators to encourage this participation. They organize local offline meetings in several hubs, as well as an annual conference, *WikiCon*[44]. Wikipedia is not free from complications, of course. Much has been reported on how moderators suffer from harassment on the platform[45].

41 According to Alexa.com, visited on June 15, 2020: https://www.alexa.com/topsites/countries/FR

42 Clark, Justin, Robert Faris, Urs Gasser, Adam Holland, Hilary Ross, and Casey Tilton. *Content and Conduct: How English Wikipedia Moderates Harmful Speech*. Berkman Klein Center for Internet & Society, Harvard University, 2019: http://nrs.harvard.edu/urn-3:HUL.InstRepos:41872342

43 Testimony of the president of Wikimédia France during the seminar held on February 14, 2020 by Renaissance Numérique.

44 Testimony of Christina Dinar, former Project Manager at Wikimedia Germany. Interview February 25, 2020. For more on WikiCon: https://en.wikipedia.org/wiki/Wikipedia:WikiCon

45 Adi Robertson, "Wikimedia is writing new policies to fight Wikipedia harassment: Trustees say it hasn't done enough to stop abuse", *The Verge*, May 25, 2020: https://www.theverge.com/2020/5/25/21269482/wikimedia-foundation-anti-harassment-code-of-conduct-vote

FRAMAPIAF: CONFIDENCE IN THE HUMAN TOUCH

Framapiaf is a subcommunity of Mastodon moderated by Framasoft, a non-profit community education association that provides alternative, open-source software and online tools. Framapiaf's moderation is performed by five volunteers who communicate with one another and with Framasoft employees as needed[46]. They have explicit and relatively strict rules in their Charter[47], in which they retain the right to make a decision without fully explaining their reasoning. Whenever possible - when moderators judge the author of a piece of content likely to engage in positive dialogue - they open a discussion with the author about the content and how it does not fit with their community standards. Otherwise, they may remove any post and even ban the author from the group.

They do not rely on automated algorithmic moderation. As a volunteer moderator of the platform testifies: "*Framapiaf is open-source. Moreover, there are no algorithms, no AI, and there are few layers between users and moderators. There are no intermediate layers. This human touch gives confidence to the users.*"[48]

As these typologies and examples testify, there is far from one style of moderation. There is great variation between platforms, and even within the same platform over time, in relation to tools, strategies, relationship with users, and political sensibilities. It is this diversity that public policy must keep in mind when formulating regulation that will traverse this landscape.

46 Testimony of Maïtané Maiwann, volunteer moderator of Framapiaf. Interview March 31, 2020.

47 See Framasoft Conditions Générales d'Utilisation: https://framasoft.org/fr/cgu/

48 Testimony of Maïtané Maiwann, volunteer moderator of Framapiaf. Interview March 31, 2020

TOXIC CONTENT: A PROBLEM FOR ALL PLATFORMS

Similar trends and behaviors in the spread of toxic content can be observed across ugc-hosting platforms, for example, the mastery of coding techniques to remain within the limits of legality, or else the hydra-like multiplication of toxic content in response to the removal of a piece of content. There is a marked porosity between platforms. In some cases, the same piece of toxic content can be found on different platforms: links can be shared towards content hosted elsewhere, files stored in private groups or chats, etc. To illustrate: though the platform Pinterest regulates hate speech relatively rigorously, users in Italy were found to be sharing links leading to more explicit hate speech stored elsewhere within Pinterest posts[49]. Similarly, bullies and aggressors can pursue their target across multiple platforms[50]. Platform operators and policy makers must first acknowledge this inevitable porosity between platforms in order to design effective measures to address toxic content.

Some platforms even find their services intentionally abused or misappropriated. For example, illegal content may be uploaded and shared on another platform with increased privacy capabilities, where the content cannot be accessed and removed. Such platforms find themselves in a sense caught in the crossfire. In exceptional cases, illegal viral content may be saved on these platforms in "private" mode, and then embedded into third party platforms like mainstream social networks. According to a participant at the seminar, this type of hijacking calls for a deeper cooperation between platforms and online services[51].

49 *Beyond the "Big Three", Alternative platforms for online hate speech*, The EU-funded project sCAN– Platforms, Experts, Tools: Specialised Cyber-Activists Network (2018-2020), April 2019: https://www.voxpol.eu/download/report/Beyond-the-Big-Three-Alternative-platforms-for-online-hate-speech.pdf

50 Renaissance Numérique (2019), "Cyberbullying: a review of the literature", Available online: https://www.renaissancenumerique.org/ckeditor_assets/attachments/493/cyberbullying_eng_page.pdf

51 Testimony of a representative of Dailymotion at the seminar held on February 14, 2020, by Renaissance Numérique.

The literature similarly shows that following the success of large, "industrial" platforms in identifying and blocking accounts supporting terrorist and extremist content, this content has found its way into other online spaces[52]. These "safe havens" are chosen for their lack of moderation capacity, for example the platform *JustPaste.it*, created and run by a Polish student entirely on his own[53]. Individual users and entire communities that are expelled from one platform can seek refuge or regroup on another. In fact, this moderation method of "de-platforming" is debated, as it can have the effect of pushing the authors of toxic content to other platforms where moderation is more difficult[54]. The sCAN project has found that "*a migration to platforms like VK.com or Gab.ai is often openly advertised on Facebook and Twitter*." In this way, some platforms find themselves hosting the hateful communities that assembled and organized on the major platforms[55]. Launched in 2017, the collective Tech Against Terrorism aims to help large and small platforms to protect their services from exploitation for terrorist or extremist purposes[56].

52 *Beyond the 'Big Three', Alternative platforms for online hate speech"*, The EU-funded project sCAN– Platforms, Experts, Tools: Specialised Cyber-Activists Network (2018-2020), April 2019. Available online: https://www.voxpol.eu/download/report/Beyond-the-Big-Three-Alternative-platforms-for-online-hate-speech.pdf

53 Extremists driven off Facebook and Twitter targeting smaller firms", *The Guardian*, July 12 2017: https://www.theguardian.com/uk-news/2017/jul/12/extremists-driven-off-facebook-and-twitter-targeting-smaller-firms How a Polish student's website became an Isis propaganda tool", *The Guardian*, August 15 2014: https://perma.cc/B2GH-5BME

54 N. F. Johnson et al.,"Hidden resilience and adaptive dynamics of the global online hate ecology" *Nature,* 2019: https://www.nature.com/articles/s41586-019-1494-7
Ryan Greer, "Weighing the Value and Risks of Deplatforming", *GNET Insights*, May 11 2020: https://gnet-research.org/2020/05/11/weighing-the-value-and-risks-of-deplatforming/

55 *Beyond the "Big Three", Alternative platforms for online hate speech*, The EU-funded project sCAN– Platforms, Experts, Tools: Specialised Cyber-Activists Network (2018-2020), April 2019: https://www.voxpol.eu/download/report/Beyond-the-Big-Three-Alternative-platforms-for-online-hate-speech.pdf

56 Renaissance Numérique, "Modération des contenus terroristes : défis techniques, enjeux démocratiques", *blog.seriously.ong*, March 2020: http://blog.seriously.ong/moderation-des-contenus-terroristes-defis-techniques-enjeux-democratiques/
Renaissance Numérique, "3 Questions to Jacob Berntsson", *blog.seriously.ong*, March 2020: http://blog.seriously.ong/3-questions-to-jacob-berntsson/

Just how much toxic content is present on these platforms and precisely how it travels between them remains an important question. There is a continued quest for data to properly illustrate the phenomenon of the spread of toxic content and the effectiveness of moderation efforts on different platforms. Transparency reports aim to provide this information and are often the best references available to researchers and policy makers. It bears noting that large "industrial" platforms tend to have better compliance in terms of transparency reporting, in large part because they have more resources to devote to this. Still, many challenges and gaps in platform transparency reporting make analysis difficult, especially comparative or cross-platform analysis. The fact that different platform operators have different ways of classifying and approaching toxic content makes it difficult to compare data collected by different actors. For example, Twitter identifies "platform manipulation" (the use of bots), but this type of content may be observed and classified differently depending on the platform[57]. This is not to say that there is a single model for transparency reporting[58]. It simply means that researchers and stakeholders seeking useful comparative data will have to negotiate these discrepancies. More importantly, because content moderation is not just about content removal, these transparency reports, which often focus on removal, do not provide a comprehensive overview of moderation. Data on takedowns and takedown requests (as well as takedown refusals and disputes) is the basis, but not the end, of transparency reporting. Researchers examining platform reporting in Germany following the NetzDG explain that these kinds of self-reported quantitative metrics may not be an appropriate indicator of the platform's success in addressing toxic content: "*The number of takedowns or number of complaints become a metric to measure the law's efficacy; these takedowns might be ineffective or even counterproductive in combating the overall prevalence of hate speech*."[59]

The European Commission's Code of Practice (CoP) on Disinformation, adopted in 2018, and the EU Code of conduct on countering illegal hate speech online, adopted in 2016[60], represent useful collaborations around self-regulatory standards between the Commission and platform operators. They have gathered some helpful data among their initial members, and hope to bring further insights as more platforms join and as application of the codes strengthens and definitions and indicators become more coherent[61]. However, there is room for improvement and fortification of these initiatives. An assessment published in May of the Code of Practice on Disinformation noted the challenges from the fact that the 13 signatory platforms have different ways of putting the code into practice, limiting the usefulness for researchers, and also noted the lack of a shared understanding and harmonized approach towards the concept of disinformation. Non-signatory platforms found the self-assessment reports of signatory platforms to be not harmonized and "not user friendly". Most recently, the multi-stakeholder "Sounding Board" of the European Commission which helped establish the code has called for stronger obligations to be imposed on signatories,

57 Another example: some content is removed by platforms in response to violations of the Terms of Service, while other content is removed because it is illegal under national law and was flagged by the government; some platforms separate these requests for removal by governments from removal requests by other users, and other platforms do not.

58 The Santa Clara Principles on Transparency and Accountability in Content Moderation, drafted in 2018 by a group of organizations, advocates, and academic experts, outline minimum levels of transparency and accountability for tech platforms around the moderation of user-generated content. These principles are not set in stone, and the Covid-19 crisis has inspired activists, experts, and platforms to reflect on possible evolutions. "EFF Seeks Public Comment About Expanding and Improving Santa Clara Principles Recommendations Sought from Those Affected by Policies to Moderate", Suppress Speech, Press Release, April 14, 2020: https://www.eff.org/press/releases/eff-seeks-public-comment-about-expanding-and-improving-santa-clara-principles

59 Heidi Tworek and Paddy Leerssen, "An Analysis of Germany's NetzDG Law", A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, April 15, 2019: https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf

60 The EU Code of conduct on countering illegal hate speech online began with the larger actors, Facebook, Microsoft, Twitter and YouTube, and has since been joined by Snapchat, Dailymotion and Jeuxvideo.com.
See the website of European Commission: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

61 Study for the assessment of the implementation of the Code of Practice on Disinformation, May 8, 2020: https://ec.europa.eu/digital-single-market/en/news/study-assessment-implementation-code-practice-disinformation

citing the experience of the Covid-19 'infodemic'[62].

Currently, public policy debate fails to fully grasp the idiosyncrasies and interconnectedness of content moderation across platforms. This failure stems in part from the lack of comparative research, related to gaps and limitations in transparency reporting. We will certainly emerge from this period with insights to inform policy decisions (though how much data platforms share remains to be seen)[63]. Moving forward, future initiatives must not fail to consider moderation in a broader sense, both transversely across platforms, and holistically as a collection of decisions and processes (beyond simply the number of contents removed).

62 *"We note the announcements and modifications made by signatories of the code of practice in regard to fighting disinformation related to Covid-19 on their networks. This demonstrates that, where willingness is present, these actors can deploy solutions at scale to curb harmful content on their networks."*
See the full joint statement published on June 15, 2020:
https://m.contexte.com/medias-documents/2020/06/Declaration_medias_desinformation.pdf

63 See the joint letter signed by civil society organizations and researchers calling on platforms to preserve data on content decisions made during the health crisis.
"COVID-19 Content Moderation Research Letter ", April 22, 2020: https://cdt.org/insights/covid-19-content-moderation-research-letter/

PART II
LIMITATIONS OF THE CURRENT LEGAL FRAMEWORK

THE DOMINANCE OF THE INDUSTRIAL MODERATION MODEL

The legal frameworks, as well as the moderation frameworks specific to platforms, are often constructed around the practices of the first actors to rise to power in the market. These were the first actors to experiment with and systematize moderation principles, and this can be seen as a kind of *first-mover advantage*[64]. Indeed, many of the platforms that now moderate in an industrial manner began with more artisanal moderation capacities and strategies. As their user-base and resources have grown — but before the arrival of strong regulatory pressure around content moderation — these platforms have had the opportunity to innovate and to augment their teams and capacities over time[65]. For example, Google created the copyright infringement detection tool *ContentID*, and Microsoft set the standard on tracking CSAM with *photoDNA*[66]. Reflecting on *ContentID*, James Grimmelmann asserts that it is the advanced capabilities in computing and algorithmic processing that have made such innovations possible, and have allowed for moderation techniques that legislators and regulators would likely not have been able to articulate and impose themselves.[67]

64 See definition from David Gotteland, "Comment surpasser l'avantage du premier entrant", Décisions Marketing, No. 21 (Septembre-Décembre 2000), pp. 7-14, *Association Française du Marketing*. Available online: https://www.jstor.org/stable/40582911

65 Mike Masnick, founder and CEO of Floor64 and editor of the Techdirt blog, does not mince his words on the subject: *"Some of us keep pointing out to the EU that if these laws are designed to go after Google and Facebook, they're going to miss their target quite a bit, because they'll mostly serve to lock in those companies as the dominant providers. That's because they're big enough to manage the regulatory burden, whereas startups and smaller competitors will not be able to and will suffer."* Masnick is referring here specifically to GDPR and AdTech. Foundation for Economic Education, "Google and Facebook Will Just Get Stronger if Regulators Get Their Way, Europe's Experience Shows", August 27 2019: https://fee.org/articles/google-and-facebook-will-just-get-stronger-if-regulators-get-their-way-europe-s-experience-shows/

66 Evelyn Douek, "The Rise of Content Cartels: Urging transparency and accountability in industry-wide content removal decisions", *The Knight First Amendment Institute*, February 11, 2020: https://knightcolumbia.org/content/the-rise-of-content-cartels

67 James Grimmelmann, "The Virtues of Moderation", *Yale J.L. & Tech*, 2015: https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1110&context=yjolt

These advancements in automated moderation technology by dominant players have come to set the standard. In turn, the preeminence of these moderation capacities has led the innovators to consolidate their dominant positions. Many in France were concerned that requirements in the Avia Law would bring this adverse effect — for example, the requirement that platforms remove certain content in 24 hours or even in 1 hour (for illegal content notified by the authorities), or else face a penalty of one year's imprisonment and a fine of EUR 250 000. This is a *de facto* a requirement for industrial-style moderation. The UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, voiced this issue in a letter to the French government in August of 2019: "*I am deeply concerned that strengthening the role of operators of online platforms in the moderation of content on the Internet could further increase 'an excessive concentration of ownership and control over the practices, [which] constitute an abuse of a dominant market position'*"[68]. Indeed, the non-dominant, non-"industrial", platforms tend to lack access to the quickly evolving artificial intelligence and battalions of human moderators operating around the world and around the clock. They do not have the same resources to perform moderation as the industrial platforms and will struggle to abide by new regulatory requirements. A mandate of one hour and 24-hour take down would oblige them to put in place teams around the clock in addition to algorithmic filtering tools — and to establish these new measures quickly. Meanwhile, as Kaye expresses, the dominant actors tend to own and oversee the now critical moderation technology: digital fingerprinting technology like *ContentID*, and access to hash databases like the GIFCT. Evelyn Douek has exposed the dangers of these content sharing data bases established by the largest internet companies – which she dubs "*content cartels*"[69] — in relation to transparency and accountability, competition, and even effectiveness; categories like spam, CSAM, and copyright

68 David Kaye, "Mandat du Rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression" UNHCR reference OL FRA 6/2019, August 20, 2019: https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL_FRA_20.08.19.pdf

69 Evelyn Douek, "The Rise of Content Cartels: Urging transparency and accountability in industry-wide content removal decisions", *The Knight First Amendment Institute*, February 11, 2020: https://knightcolumbia.org/content/the-rise-of-content-cartels

might have clear enough parameters, but categories like hate speech and bullying will test the limits of this method. As policy makers and platforms continue to decide on hashable, scalable categories of toxic content — (categories supposedly precise enough to be addressed *ex ante* though databases) — close attention must be paid to the contents of these databases (information which is not currently available), as well as to their management: which platforms own them and have active decision making roles, and which platforms are passive recipients of this technology.

Today, co-regulation remains a bilateral process between the major platform operators and governments. As would be expected, the larger actors have more lobbying capacity to influence regulation. Meanwhile, the representatives of the "other", non-dominant platforms are not well integrated into policy debates. These platforms tend to have smaller public policy teams, if they have policy representatives at all. In France, smaller platforms lament, along with civil society, not having been meaningfully consulted in the legislative process during the 2018 French Fake News Law (the Information Manipulation Act), and most recently in the Avia Law on cyberhate[70]. As an example of just how critical lobbying can be in content regulation: in Germany, the gaming industry lobby succeeded in having obligations removed from the initial draft of the NetzDG[71].

The primacy of the industrial moderation model also has social and political consequences. Moderation frameworks built around these industrial modalities tend to favor *ex ante* moderation and a more conservative approach to borderline content. This increases the chance that false-positives remove legitimate content. Despite calls by platforms, civil society, and even states against "general monitoring"[72], the recent regulatory trend appears to push platforms towards just that, as strict deadlines, obligations and sanctions — like removal of terrorist content in one hour[73] — effectively oblige automated content filtering[74]. In its June 18th decision, France's Constitutional Council identified the risk that the Avia law would encourage platforms to take down legitimate content, thus infringing excessively on freedom of expression. The court found that, lacking specific grounds for exemption from liability, the penalties for failing to act on reported content in the prescribed time period would encourage platforms to take down content reported to them *"whether or not it is manifestly unlawful"*[75].

The current legal framework built around the industrial approach reinforces the dominance of the already-dominant players — those with the technical capacity, human resources, and lobbying power — while pushing all platforms towards increasingly automated, *ex ante* moderation practices that risk censoring legitimate content. A new regulatory approach is called for, one that accounts for diverse approaches and protects fundamental rights, including by avoiding the trap of general monitoring and excessive curtailing of the freedom of expression.

70 Testimony from participants during the seminar held on February 14, 2020 by Renaissance Numérique

71 Hartleb, Florian. "Lone Wolves: The New Terrorism of Right-Wing Single Actors", *Springer*, 2020.

72 D9+ Non-Paper on the creation of a moderation regulatory framework for the provision of online services in the EU: https://www.gov.pl/web/digitalization/one-voice-of-d9-group-on-new-regulations-concerning-provision-of-digital-services-in-the-eu

73 The French Cyberhate law voted in May 2020 is the latest regulation to impose a 1 hour takedown of "terrorist content". To see the text online: http://www.assemblee-nationale.fr/dyn/15/dossiers/lutte_contre_haine_internet

74 Hannah Bloch-Wehba, *Automation in Moderation,* Cornell International Law Journal, Forthcoming, last revision: April 29, 2020,: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3521619 Renaissance Numérique joined with other members of French civil society in an open letter to French lawmakers, "Lettre ouverte relative à la proposition de loi visant à lutter contre la haine sur Internet", July 2019: https://www.renaissancenumerique.org/publications/lettre-ouverte-relative-a-la-proposition-de-loi-visant-a-lutter-contre-la-haine-sur-internet

75 *"compte tenu des difficultés d'appréciation du caractère manifestement illicite des contenus signalés dans le délai imparti, de la peine encourue dès le premier manquement et de l'absence de cause spécifique d'exonération de responsabilité, les dispositions contestées ne peuvent qu'inciter les opérateurs de plateforme en ligne à retirer les contenus qui leur sont signalés, qu'ils soient ou non manifestement illicites. Elles portent donc une atteinte à l'exercice de la liberté d'expression et de communication qui n'est pas nécessaire, adaptée et proportionnée. Dès lors, sans qu'il soit d'examiner les autres griefs, le paragraphe II de l'article 1er est contraire à la Constitution."* Décision n° 2020-801 DC du 18 juin 2020, Loi visant à lutter contre les contenus haineux sur internet: Available Online: https://www.conseil-constitutionnel.fr/decision/2020/2020801DC.htm

RETHINKING THE INDICATORS THAT INFORM REGULATION

In order to begin to regulate the practice of content moderation, indicators must be established to identify both the actors that fall under the regulation as well as the activities and aspects that should be monitored. Today's regulatory efforts aiming to capture the industrial platforms end up catching other actors into their nets, while failing to grasp the complexity of the problem. It is therefore necessary to question the indicators that currently inform the public policy that in turn determines the legal framework and moderation modalities for these platforms.

Importantly, our current indicators also vary across countries and legislative acts. There is therefore a need to harmonize these indicators and new regulatory approaches at the European level. Though Germany and France have attempted their own legislations before the arrival of the Digital Services Act, it is far less efficient and authoritative for platforms to impose patchwork moderation policies across individual member states. A minimum of harmonization is essential to addressing toxic content. A shared understanding and approach to the problem can be achieved in part through the framework of process-focused indicators.

A central question for many is how to look beyond the concept of user threshold, or the number of in-country users on a platform at which point the service will be subject to regulation (in France, the number generally evoked is 5 million). This concept is inapt, as this figure alone does not illustrate the moderation challenges faced by the platform. Quality rather than quantity criteria are necessary. For example, does the platform put in place safeguard measures? Does its business model encourage virality? Does the user experience draw on "dark patterns" to intentionally mislead users? The question of the user threshold was debated and finally adopted within the German NetzDG (2 million registered users in Germany), but it also bears mentioning that there is no threshold in *Section 230*, the most analogous regulatory framework in the United States. If user threshold is not an ideal or universal concept, can we reflect instead around the business models, the platform design, or the basis of other specific characteristics? The question of appropriate and effective indicators in not a simple one in this space, where regulation must account for the wide variety of platform operators which face very different moderation challenges, as shown.

We therefore need agile indicators that let us measure the responsiveness of platforms to the real problems they face, problems which evolve. Regulation that may be useful at a given time with respect to a specific issue may not be able to solve future challenges. The Facebook Mission carried out in 2019 in France focused on algorithmic content ordering, which is justified when thinking about the moderation issues of Facebook and YouTube, which are often linked to the curation and virality of toxic content. But there are platforms that do not rely on algorithmic curation and still have toxicity and virality. In this way, there will always be limitations to regulation established around any single technical feature — content ordering, live streaming, etc. — attributes which are not necessarily or exclusively the source of toxic content per se.

Below, indicators are offered to assess the performance of platform content moderation. These indicators seek to look beyond the concept of user threshold and instead question moderation processes and practices. It should be noted, however, that not all improvements in the area of moderation will pass through regulation. As this analysis has aimed to show, overly specific regulation risks causing unintended harm to certain platforms due to their idiosyncrasies, the diversity of services and the rapid evolution of their characteristics and uses. The indicators proposed here, classified into five areas, place the emphasis on general principles rather than on explicit methods. Regulators can use these indicators to measure the engagement of platforms in content moderation without imposing restrictive methods on how platform operators must achieve this engagement. Methods are proposed beneath these general principles, suggesting policies that platforms could choose to implement to adhere to these principles. Of course, these suggestions should not be viewed as definitive, but as a starting point for collaboration; we are at the very beginning of this process and the table below seeks to open future discussion rather than restrict it.

INDICATORS TO ASSESS THE PERFORMANCE OF MODERATION AMONG PLATFORMS IN EUROPE

TRANSPARENCY AND ACCOUNTABILITY

- The implementation of transparent and well explained moderation principles and processes.
- The publication of clear, comprehensive and regular transparency reports.
- Insight into individual cases that expose the logic and merits of decisions (qualitative as well as quantitative insights).

Platforms may choose to:

- Provide clear and accessible conditions of service in all languages in which the service is offered.
- Share appropriate data with researchers, including critical algorithms and other decisions that inform content flows.
- Allow access to "raw" aggregated data for analysis (as aggregated figures are difficult to verify).
- Open their APIs directly to researchers and regulators.

OVERALL INVESTMENT IN MODERATION

- Maximize the amount of resources invested in moderation in relation to the capacity of the operator.
- Investment in human moderation and in the humans behind it (safety, remuneration).
- Investment in knowledge sharing across teams.

Platforms may choose to:

- Increase investment in human moderation, especially in local human moderators and their ongoing training and well-being, providing good working conditions and appropriate support.
- Strive to bring people from diverse backgrounds and life experiences into moderation: as community managers, experts/advisors, beta-testers for design decisions, etc. This may involve investment in the remuneration of external actors.
- Invest in knowledge sharing within the organization, e.g. connecting designers and policy teams with security teams, etc.

DIALOGUE WITH USERS

- The implementation of systems for mutual consultation with users.
- The implementation of clear, transparent and timely systems for redress and remedy.

Platforms may choose to:

- Encourage user flagging of toxic content.
- Motivate user participation in moderation. For example, through graduated engagement systems.
- Provide human support to users making complaints.
- Establish transparent, effective and timely mechanisms for appeal and remedy, that are easily understandable and easily accessible (within 3 clicks maximum[76]). Possible quantitative metrics for responsiveness to complaints include:
  - the speed at which a decision was made on reported content (including the decision to disable access to content and not necessarily the final decision);
  - the speed with which the competent authorities are informed in the event of illegal content.
- Inform users when a moderation decision is made about their content and include adequate information on what triggered the decision, the specific rule that was breached, how the con-

76 'Burried' reporting tools was noted as a hindrance to user reporting on Facebook in Germany.
Heidi Tworek and Paddy Leerssen, "An Analysis of Germany's NetzDG Law", A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, April 15, 2019: https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf

tent moderation guidelines were interpreted, the measures that will be taken, and clear instructions on how to lodge an appeal.

- Provide educational resources to help users understand the logic behind decisions made.

DEFERENCE TO EXPERTISE AND ATTENTIVENESS TO LOCAL CONTEXT

- Inclusion of civil society and experts from the local context and with relevant expertise.
- Data and knowledge sharing with researchers and other stakeholders.
- Commitment to improve researchers' ability to access data and decision-making logic.

Platforms may choose to:

- Seek the expertise of civil society and experts at multiple stages of the design and moderation process (in the creation of community standards, in product and UX design, in moderation decisions, etc.). Meaningful consultation goes beyond relying on civil society organizations to flag content (as in Trusted Flagger programs) and on experts to fact-check content. At the same time, relationships with civil society should not be imposed on their limited resources.
- Consult with experts in the local context so as to make decisions that are culturally informed and that avoid the problem of 'extra-territorial' decision making.

ALIGNMENT WITH FUNDAMENTAL RIGHTS AND LIBERTIES

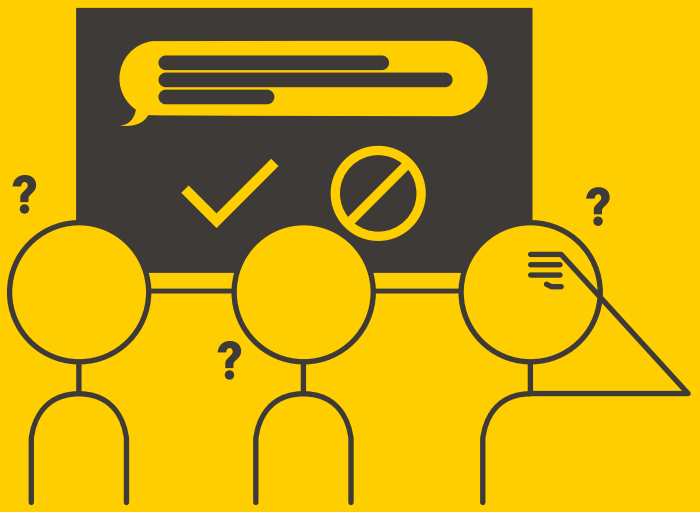
- Implementation of mechanisms for recourse and remedy that respect due process.
- Recovery of erroneously removed content.
- Appropriate processing of user data according to GDPR and applicable legal frameworks.

Platforms may choose to:

- Invest in developing technologies and communication materials for user security.
- Invest in developing technologies and communication materials for education and user resilience.
- Put in place appropriate design measures for different audiences (e.g. for children, journalists, etc.), and develop these resources and technologies in collaboration with the users themselves.

While far from a definitive approach, these indicators hope to push evaluation beyond the concept of user threshold and number of takedowns towards a more process-oriented assessment of platform moderation performance.

PART III
TOWARDS A COLLABORATIVE APPROACH TO MODERATION



CO-CONSTRUCTING MODERATION FRAMEWORKS

A major challenge for content moderation is the application of systems and frameworks at scale, and for diverse users across large cultural/legal/social territories. Context is critical to understanding the meaning of content and how it will be received. Moderation must take into account the local context and rely on appropriate expertise. This strategy is not specific to any of the three moderation typologies, but requires a collaborative approach. Collaboration will necessarily manifest itself differently according to the platform. It may consist, for example, of involving local experts, engaging civil society, working with journalists and fact-checkers, and sharing pedagogic or resilience-building resources that are pertinent in the local context.

During the seminar organized by Renaissance Numérique on 14 February, 2020, several attendees expressed their desire for a discussion group among platforms and civil society organizations in France to share knowledge and debate moderation challenges within the national context[77]. Strengthening the capacity to make moderation decisions at the local level would help to avoid moderation failures that stem from 'extraterritorial' decision-making, or sending a decision abroad to the platform's headquarters - often in the United States - where decision makers may not necessarily understand all the contextual elements necessary for an informed decision. Platform operators do not want to play the role of judge, particularly in areas where they lack expertise, and on sensitive cultural issues like religious clothing, LGBTQ issues, etc. — these issues are delicate to parse. In France, some fear that the blunt approaches provided by the controversial Avia Law are inept. While the association Inter-LGBT denounces "raids" or online attacks on LGBTQ content — behavior the law aims to forcefully address — they fear that LGBTQ content would be taken down unfairly and without sufficient explanation or appropriate recourse under the new law. The association also

77 The creation of an independent supervisory board by civil society actors has been discussed in Germany within the framework of the NetzDG but does not yet exist. Testimony of Christina Dinar, former Project Director at Wikimedia Germany. Interview February 25, 2020.

raised concerns that the law will force minors to out themselves should they wish to make a complaint according to the legal process provided to them under the law. Véronique Godet, co-president of SOS Homophobie, vowed to remain vigilant on the application methods of the text: *"What assurance do we have today that the content we are removing is indeed hateful? For the moment, none"*[78].

The concept of a separate body assembling platforms and civil society would also presumably benefit civil society watchdogs and researchers by giving them access to data and internal decision making. Facebook's Oversight Board, not without controversy, seems to respond to some of these issues. TikTok and Twitch plan to follow this model[79] and likely other platforms will as well. Such "social media councils"[80], when well implemented, may offer benefits to platforms and users by replacing ad hoc moderation decision making with a more transparent, accountable and legally compliant system. However, these structures should not replace the prerogatives of justice in a state governed by the rule of law[81]. Further, such social media councils are not a sufficient tool for achieving massive user participation - these are not mechanisms for bottom-up collaboration and discourse with diverse users. Civil society and users broadly should be able to voice concerns and contribute to platform moderation policies. Inherent to user-generated content-hosting platforms is the notion of the co-creation of value: these spaces cannot simply be viewed as the territories of private companies due to the important role they play in democratic debate and the extent to which end users contribute to their construction[82]. This substantial contribution must be reflected in their governance, in particular in relation to content moderation. The top-down approach of an oversight board and a more democratic, community-driven approach advocated for in this analysis are not mutually exclusive, in fact they may well be complimentary. It bears noting that social media councils risk further consolidating the importance of the largest actors, the first-movers into this new techno-policy space: Facebook, TikTok and Amazon's Twitch. They also may have consequences around freedom of expression by homogenizing the way content is governed, as Kate Klonic warns[83]. Strengthening communication between platforms and users, and involving users broadly in the moderation process, is one way of mitigating these risks.

A space for this kind of communication must first be constructed, and capacities of all parties ensured. Public authorities have the obligation to reinforce the capacities of stakeholders to allow for functional, collaboration and discursive processes. It is the responsibility of public authorities to facilitate intra- and inter-sectoral collaboration and knowledge sharing, to work with civil society, researchers and technical experts to find effective methods, and to share these methods with all actors. The European Digital Media Observatory currently taking shape is well suited to this role at the European level. In France, the 'observatory of online hate' put forward in the Avia Law[84] could play this role at the national level. The latter observatory is proposed under the authority of the *Conseil supérieur de l'audiovisuel* (CSA), the regulator tasked with ensuring the law's implementation. Care should be taken to avoid redundancy and inconsistency across these bodies. Public authorities must ensure that all platform operators, in particular those with fewer resources to devote to these issues, are consulted and accounted for

78 Hervé, Elodie. "Les associations LGBT inquiètes après le vote de la loi Avia contre la haine en ligne", *Têtu*, May 13, 2020: https://tetu.com/2020/05/13/les-associations-lgbt-inquietes-apres-le-vote-de-la-loi-avia-contre-la-haine-en-ligne/

79 See TikTok Newsroom, March 18, 2020: https://newsroom.tiktok.com/en-us/introducing-the-tiktok-content-advisory-council

80 Social Media Councils, from Concept to Reality", Stanford Digital Policy Incubator Conference Report, 1-2 February, 2020: https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/gdpiart_19_smc_conference_report_wip_2019-05-12_final_1.pdf

81 Renaissance Numérique (2020), Regulating digital platforms: why and how?, Available online: https://www.renaissancenumerique.org/publications/regulating-digital-platforms-why-and-how?token=OXtPZ5JJaXc8BTyU-xX9Fw

82 Through both the content they share intentionally and the data they share in the process.

83 See comment by Kate Klonic, "How Facebook's oversight board could rewrite the rules of the entire internet", *Protocol*, May 6, 2020: https://www.protocol.com/facebook-oversight-board-rules-of-the-internet?utm_campaign=The%20Interface&utm_medium=email&utm_source=Revue%20newsletter

84 Consult the most recent version of the text (from May 13, 2020): http://www.assemblee-nationale.fr/dyn/15/textes/l15t0419_texte-adopte-provisoire.pdf

in the creation of regulatory frameworks. To this end, regulators could require an impact assessment to accompany the regulation. They could also strengthen the technical capacities of platforms with fewer resources: one approach would be to direct the fines drawn from regulatory violations toward capacity building among these latter platforms.

The issue of interoperability, for example, will prove an important territory for public authorities to ensure capacity across the full range of ugc-hosting platforms. As policy conversations and technical decisions begin to define the contours around data portability and interoperability, all platforms must be considered to avoid the creation of standards and protocols that exclude or hamper them from benefiting from these developments. Interoperability and data portability intend to empower users to move more freely between platforms (to realize fully a right they are already entitled to under the GDPR), and to level the playing field among platforms. But this alone is not sufficient to establish competition, since the skills and means deployed by digital platforms are the ultimate factors that give them a competitive advantage, not the data by itself. Indeed, not all platforms will favor interoperability, for example, well-established niche operators, for fear that this openness could strengthen the largest actors with the strongest innovation capacity by granting them access to their data[85]. Public authorities must ensure that all relevant platform operators are heard in this future debate, to avoid interoperability becoming a tool to strengthen disproportionately the dominant few.

85 Renaissance Numérique (2020), Regulating digital platforms: why and how?, Available online: https://www.renaissancenumerique.org/publications/regulating-digital-platforms-why-and-how?token=OXtPZ5JJaXc8BTyU-xX9Fw

CULTIVATING A CULTURE OF MODERATION WITH USERS

A collaborative approach with users can assist platform operators in content moderation, particularly given the contextually sensitive aspects. In the case of Wikipedia and Framasoft, both non-profit organisations, this community approach is a financial necessity should they continue to operate at scale. But the merits of this approach should not be seen as financial — indeed, the fact that community moderation is not remunerated raises some concerns. A collaborative approach requires discursive processes, not just the outsourcing of labor. Governance structures are needed to facilitate this participation. Examples of more successful, inclusive moderation suggest that the most effective systems are multidimensional, with many levels of participation around a core moderation team, and with clear and strong communication between these layers. Christina Dinar describes both Wikipedia and the German website Gutefrage.net as “onion” systems in their many layered structures around a central core[86]. Because content moderation has been a challenge since the emergence of the modern web, lessons in community governance should also be taken from these early days (semi-transparent community moderation in blog communities like MetaFilter or Slashdot)[87].

NGOs and activists in France have long voiced the need for citizens to reclaim their online spaces from toxic content through online mobilizations and counter-speech (SOS Homophobie, SOS Racisme, #StopHateMoney, Project Seriously). Similarly, many lament the *bystander effect*, where users witness toxic content without reacting. As a result, user participation is often mentioned in the context of 1. being an “active bystander”, 2. officially flagging content or 3. responding to this toxic content. This kind of user involvement must be part of a broader behavioral shift on online platforms:

86 Testimony of Christina Dinar, former Project Director at Wikimedia Germany. Interview February 25, 2020.

87 See research on MetaFilter: https://metatalk.metafilter.com/24732/Taking-Care-of-a-Fruit-Tree-Moderation-on-Metafilter

'active bystander', 'digital citizenship', user 'resilience' — these buzzwords are often evoked, gesturing at a fundamental reframing of the end user as agent. Online counter-speech is an important development, but as a practical matter, often fails to rise above the fray of toxic content[88]. It is therefore necessary to strengthen and formalize channels for user contribution and to create mechanisms that value this kind of participation from users[89]. Reinforcing existing channels for complaints, recourse and remedy is one immediate way that platforms can re-center user participation. However, regulators share this responsibility for capacity building. As Renaissance Numérique has put forward in a recent note on the "platformization" of digital service regulation[90], regulators could offer a macro-approach to end user participation across digital services by means of a digital platform. Inspired by the logic of the digital platforms themselves, such a system could aggregate feedback and disputed cases from across a range of platform operators; it would streamline and structure the contributions of millions of end users and give them weight in their dialogue with these services, including through the construction of regulatory and moderation instruments (indicators, processes, etc.). Of course, while it is important for platform operators to foster dialogue with users, a collaborative approach should not be imposed in such a way as to further constrain those platforms that are less able to do so. Giving this macro responsibility to the regulatory authorities would ensure against this scenario.

Platforms are responsible for educating and equipping users as part of the construction of a culture of moderation. Moderation is much more than content removal, and it often necessitates pedagogy at the user level. Pedagogy can be integrated in the design and features of the platform, as many platforms are already experimenting with[91]. Pedagogy is also part of transparent and efficient recourse mechanisms: the availability of clear and unambiguous policies helps to reduce the repetition of infractions and increases confidence in the governance of the platform. Platforms have an obligation to provide resources to help users understand the logic behind moderation decisions.

88 Renaissance Numérique (2017), Taking action against hate on the internet in a collaborative society, Available online: https://www.renaissancenumerique.org/ckeditor_assets/attachments/210/note_finale_seriously_en.pdf

89 Platforms could regard counter-speech as a form of content moderation, and explore content curation and design decisions to facilitate and promote it.

90 Renaissance Numérique (2020), Regulating digital platforms: why and how?, Available online: https://www.renaissancenumerique.org/publications/regulating-digital-platforms-why-and-how?token=OXtPZ5JJaXc8BTyU-xX9Fw

91 For instance, with automated warnings (messages shown to the author asking them to reflect before posting potentially harmful content). In the context the Covid-19 pandemic and the "infodemic", many platforms are implementing pedagogical material and "nudging" features around misinformation (alerts, redirections to scientifically-supported sources, etc.) "How Facebook can Flatten the Curve of the Coronavirus Infodemic", Avaaz, April 15, 2020: https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/?utm_campaign=The%20Interface&utm_medium=email&utm_source=Revue%20newsletter

CONCLUSION
MODERATION, A TOOL FOR DEMOCRACY

It is necessary to value collaborative processes and to construct participative frameworks for content moderation both within and beyond public policy. There are some moderation practices that regulation can encourage, and other practices that require the participation of civil society or the actions of platforms directly. Beyond a "Duty of Care" among platform operators, a system of governance characterized by democratic and user-centric processes is called for. Meanwhile, policy makers must shape regulation in a way that does not reinforce the dominance of a few platforms to the detriment of others. As this analysis has tried to convey, though it is often the tools and methods of the 'industrial' operators that become the norm across the platform landscape, there are lessons that all platform operators can take from more 'community-based' and 'artisanal' approaches to content moderation. As is clear, further research is needed, particularly comparative cross-platform research on different moderation approaches. It would also be useful to further parse the typologies used here (industrial, artisanal, and community-reliant), for instance further examining moderation on decentralized and open-source platforms, and moderation on platforms supported by advertising or by subscription.

Recent events like the face-off between U.S. President Donald Trump and Twitter, and the Covid-19 'infodemic' have drawn attention to the need for fundamental changes to content moderation, as well as to the democratic issues at stake. These events also show the necessity for these changes to be made through democratic deliberation, and through regulation that is crafted with all stakeholders. Regulation is necessary to reduce the toxicity of our online public spaces and to uphold fundamental rights and democratic debate, but it must not be a cure worse than the disease. Regulation that reinforces, either directly or indirectly, the less transparent, accountable, and user-centric practices of content moderation — among them, *ex ante* automated moderation, centralized and proprietary technology, non-transparent hash-sharing databases, etc. — risks harming not only the quality of our online spaces, but also their variety. Regulatory frameworks must be careful not to further reduce the diversity of platforms available to host a wide range of expression. Other actors must be able to enter, emerge and become sustainable in order to offer users a diversity of spaces to express

themselves and to assemble, to be safe and to be heard. Otherwise, the concentration of ugc-hosting platforms and the homogenization and industrialization of content moderation methods risk exaggerating the very challenges to online expression that content moderation is supposed to address.

FURTHER RESOURCES

- Bloch-Wehba, Hannah. “Automation in Moderation”, *Cornell International Law Journal*, Forthcoming (2020)
- “Regulating digital platforms: why and how?”, Renaissance Numérique (May 2020)
- “Nine Principles for Future EU Policy Making on Intermediary Liability”, Center for Democracy and Technology (April 2020)
- “Cyberbullying: a review of the literature”, Renaissance Numérique (April 2020)
- “Recommendations on Content Governance”, Access Now (March 2020)
- Douek, Evelyn. ”The Rise of Content Cartels: Urging transparency and accountability in industry-wide content removal decisions”, Knight First Amendment Institute at Columbia University (February 2020)
- “Online Harms White Paper: Initial consultation response”, Gov.uk (February 2020)
- Clark, Faris, Gasser, Holland, & Ross, Tilton. “Content and Conduct: How English Wikipedia Moderates Harmful Speech”, The Berkman Klein Center for Internet & Society at Harvard University (December 2019)
- Roberts, Sarah T. “Behind the Screen, Content Moderation in the Shadows of Social Media”, Yale University Press (August 2019)
- “Rapport de la mission ‘Régulation des réseaux sociaux’ – Expérimentation Facebook”, Remis au Secrétaire d’État en charge du numérique (May 2019)
- “Beyond the ‘Big Three’, Alternative platforms for online hate speech”, The sCAN Project (April 2019)

- Heidi Tworek and Paddy Leerssen. “An Analysis of Germany’s NetzDG Law”, A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression (April 2019)

- Newton, Casey. “The Trauma Floor: The secret lives of Facebook moderators in America”, The Verge (February 2019)

- “Fake News, Faire face aux troubles informationnels à l’ère numérique”, Renaissance Numérique (March 2018)

- Echikson, William et Knodt, Olivia. “Germany’s NetzDG, A key test for combating online hate”, Center for European Policy Studies CEPS Policy Insight (December 2018)

- Caplan, Robyn. “Content or Context Moderation?: Artisanal, Community-Reliant, and Industrial Approaches”, Data & Society (November 2018)

- “Taking action against hate on the internet in a collaborative society”, Renaissance Numérique (July 2017)

- Grimmelmann, James. “The Virtues of Moderation”, Yale Journal of Law and Technology (2015)

ACKNOWLEDGEMENTS

The think tank Renaissance Numérique organized a seminar on the challenges of the moderation of toxic content, in order to better understand the problems faced by the “other” platforms — that is to say, platforms beyond the general focus of public policy (Twitter, Facebook, Youtube) and which may favor more “artisanal” or “community-reliant” moderation methods — and to reflect together on possible courses of action. Entitled *“How to integrate all platform operators in the moderation debate”*, the seminar took place on 14 February 2020, and brought together platform representatives, members of civil society, researchers and members of French public institutions. We would like to thank all the participants of this morning, as well as the individuals interviewed as part of this publication.

Robyn Caplan, Researcher at Data & Society and PhD Candidate at Rutgers University, author of “Content or Context Moderation?” (2018) provided insights that were invaluable to this project, from the framing of the seminar to the final publication. A particular thanks is also due to Christina Dinar, Deputy Director of the Centre for Internet and Human Rights, Germany, for her insights shared during multiple interviews and her review of the final publication.

Below is the full list of seminar participants:

- Sarah Durieux, Director of Change.org France
- Baltis Mejanes, Chief of Staff and Parliamentary Counsel to Adrien Taquet
- Charlotte Collonge, Communication and Counter-Speech Officer at the French Interministerial Committee for the Prevention of Delinquency and Radicalisation (CIPDR)
- Elise Fajgeles, Policy Officer at the French Interministerial Delegation for the Fight against Racism, Antisemitism and Anti-LGBT Hate (DILCRAH)
- Lucile Petit, Head of the Department: Audiovisual Media On Demand Services, Distribution, New Services, at the *Conseil supérieur de l’audiovisuel* (French Broadcasting Council)
- Salwa Toko, President of the *Conseil National du Numérique* (French National Digital Council)
- Stéphane Koch, Digital Strategy Consultant
- Clément Reix, Public and Regulatory Affairs at Dailymotion
- Justine Atlan, Chief Executive Officer of e-Enfance
- Enguerrand Leger, Co-founder and Community Manager of GensdeConfiance.fr

- Léo Laugier, PhD student at the Institut Polytechnique de Paris
- Lucien Castex, General Secretary of the Internet Society France
- Camille l'Hopitault, Policy Officer at the International League Against Racism and Anti-Semitism (LICRA)
- Hector de Rivoire, Government Affairs Manager at Microsoft France
- William Schun, Government Affairs Coordinator at Microsoft France
- Laure Durand-Viel, Project Manager "Regulation of digital platforms" at the DGMIC at the French Ministry of Culture and Communication
- Betty Jeulin, intern at Pinsent Masons
- Louise Florand, Lawyer at Point de Contact
- Iris de Villars, Head of the Technology Bureau at Reporters Without Borders
- Benoît Loutrel, French Secretary of State for Digital Affairs, Mission on "*Régulation des réseaux sociaux*" (Social Networks Regulation)
- Jean Gonié, Director of Public Policy Europe at Snap
- Pauline Birolini, Head of the Legal Department of SOS Racisme
- Valentin Stel, Project Manager in the Legal Department of SOS Racisme
- Andrea Cairola, Programme Specialist, Division for Freedom of Expression, Democracy and Peace of the Communication and Information Sector of UNESCO
- Juliette Sénéchal, Lecturer at the Faculty of Legal Sciences at the University of Lille
- Isabelle Jaquemet, Director of Operations at Webedia
- Julien Lopez, Advocate General Gaming & E-sports at Webedia
- Pierre-Yves Beaudouin, President of Wikimedia France
- Willie Robert, Vice President of Wikimedia France
- Lucien Grandval, Public Policy and Communication Lead at Yubo

An interview was held following the seminar with Maïtané Maiwann, Moderator at Framasoft's Mastodon instance, "Framapiaf".

DIRECTOR OF THE PUBLICATION

Jennyfer Chrétien, Executive Director, Renaissance Numérique

RAPPORTEUR

Claire Pershan, Project Manager, Renaissance Numérique

ABOUT RENAISSANCE NUMÉRIQUE

Renaissance Numérique is France's main independent think tank focusing on the challenges of digital transformation in society. Bringing together universities, associations, corporations, start-ups and schools, it aims to develop workable proposals to help public stakeholders, citizens and businesses build an inclusive e-society.

Renaissance Numérique
22 bis rue des Taillandiers - 75011 Paris
www.renaissancenumerique.org

June 2020