DISPARATE CONTENT MODERATION

MAPPING SOCIAL JUSTICE ORGANISATIONS PERSPECTIVES ON UNEQUAL CONTENT MODERATION HARMS AND THE EU PLATFORM POLICY DEBATE.

Naomi Appelman

University of Amsterdam
Institute for Information Law
DSA Observatory
funded by The DSA Observatory
October 2023

SUMMARY

Through a series of expert interviews this report seeks to map several civil society perspectives on disparate content moderation. The aim is to deepen our understanding of how organisations working for the interests of or with marginalised groups: [1] conceptualise the causes and effects of disparate content moderation (downranking, shadowbanning, blocking or (refusal to) removing), and [2] relate to and are involved in the EU platform policy, specifically the Digital Services Act.

The hope is to contribute to centring perspectives that are not always the focus of the EU digital rights debate, and the DSA largely overlooks. In doing so, this project wants to help in charting out a course of action for academia and NGO's that appreciates the intersectionality and unequal distribution of content moderation harms and is grounded in solidarity.

RESULTS

1. Disparate content moderation is seen as **a result of wider systems of social oppression** that are reproduced both on the platform and by the platform itself through its content moderation. Within its corporate logic, stigmatized content could be conceived us as a risk and suppressed to minimize that risk, without (sufficient) consideration for the harm that this causes.

2. There is a large **heterogeneity in both experience and impact** that seem to be intersectional in the that they depend on the wider context of someone's life as well as the different broader societal oppressions someone is subject to.

3. Content moderation can cause a **broad range of harms** beyond the most known ones such as deplatforming, removal, or shadowbanning. The norms themselves, the vagueness or seemingly arbitrary application, data harms as well as the affordances of a platform are all instrumental to content moderation harms.

4. Especially groups that already have an adversarial relation with the law seem to experience unjustified deplatforming and content removal the most, such as sex workers and abortion activists. This seems to affect **their relation with the platform**, considering it more of an active political actor, rather than an unreachable corporation.

5. The **strategies people develop** to deal with these hams, relate to several factors: (1) the type of harm experienced, either mainly by platform action or by harassment, (2) the impact this has, whether or not professional for example, and (3) the wider relationship with the law.

6. In line with existing research, **major hurdles to access to justice** are the lack of clarity on how content moderation works and what the norms are, as well as a lack of response from platforms to notifications, a lack of explanation as to why content moderation actions were taken and, finally, available procedural routes are unclear and inaccessible.

7. Crucial is **the support from an organisation** in finding the right procedural route as well as broader support in navigating and dealing with the platform and the harm. These organisations are also important in leading possible collective actions to offload the individual people experiencing this harm.

8. Success in challenging and remedying content moderation harms is often dependent on having **contacts within or access to the platform**. This contact is dependent on the voluntary cooperation of the platforms, and, besides the real threat of arbitrariness, this might also feel to support organisations as if it could limit them in their advocacy.

9. Willingness to engage with, or trust in formal procedures, both legal and with the platforms, must be understood within the context of wider **criminalization and legal stigmatization**.

BRIEF CONCLUSIONS

The differences in experience and impact of content moderation are not explicitly recognized in the DSA but could be included in the codes of conduct or risk assessment requirements.

Intersectionality of content moderation harms as well as societal context must be considered by policymakers, academics, as well as by NGOs in campaigning for change. This means avoiding one dimensional policies and working towards broad coalitions.

A solely legalistic approach is insufficient to support communities who are apprehensive, with good reason, about formal procedures.

Further research is needed on what type of support organisations and funding structures fits specific contexts best, and how to ensure platforms engagement with these victim support organisations is not voluntary and precarious.

Platforms should ensure that the content of otherwise criminalized groups that is legal and does not violate platform norms of otherwise criminalized groups, so-called "grey zone content", does not get caught up in content moderation actions.

CONTENTS

INTRODUCTION

The disparate way in which platforms moderate online content and behaviour, often with the stated aim to create safe online spaces, can function to exacerbate existing harm or create new ones. Think for example of the over- and under-removal of specific content or shadowbanning. Crucially**, these content moderation harms are not equally distributed and disproportionately affect marginalised or vulnerable communities** (Benjamin, 2019; Glitch, UK, 2023; Marshall, 2021; Siapera & Viejo-Otero, 2021).

Through a series of expert interviews this project centres civil society perspectives on disparate content moderation in the EU policy debate. The aim is to deepen our understanding of how EU CSOs working for the interests or with often excluded groups: (1) conceptualise the causes and effects of disparate content moderation practices (i.e. downranking, shadowbanning, blocking or (refusal to) removing), and (2) relate to and are involved in the EU policy debate.

DISPARATE CONTENT MODERATION

This disparate impact is visible throughout the whole of content moderation policies and systems as they determine who is protected and who is considered a threat to 'online safety'. On the one hand there is clear evidence that over-removals disproportionally harm marginalised groups such as sex workers or Black people (EFF, 2019; Haimson et al., 2021). Similarly, a range of qualitative research documents confirm the experiences of marginalised groups such as LGBTQ+, people of colour, or non-binary people with platform policies and content moderation enforcement mechanisms that disproportionally target them and their content(Are & Briggs, 2023; Duffy & Meisner, 2022; Smith et al., 2021; Thach et al., 2022; Tiidenberg, 2021).

Not all content is moderated equally. Often, marginalised groups experience more moderation actions (removals, deplatforming, shadow-banning), receive less protection from platforms against harassment and hate.

Simultaneously, extensive research shows how content moderation systems and policies are not equipped to deal with, or even exacerbate the online harms faced by marginalised groups, such as racist hate speech (Díaz & Hecht-Felella, 2021), or the harassment of women online (Megarry, 2014). Moreover, the intersectionality of these harms and their impact is clear(Gilbert, 2023). The early-internet-optimism about the equalizing effect of online communication has, by now, given way to the realisation that not only social media but also how content is moderated can both reflect and exacerbate existing inequalities.

EU'S ONE-DIMENSIONAL ACCESS TO JUSTICE?

One thing that is clear, is that when confronted with problematic content moderation practices, both people's needs as well as their legal options differ widely according to, amongst others, the type of harm they experienced, context, and intersecting identities (Appelman et al., 2021; Hoboken, van et al., 2020).

It is unclear whether the recently adopted DSA will offer the necessary access to justice for disparate content moderation as it does not explicitly recognize this heterogeneity.

Rather, it takes a blanket approach to remedying harmful online content and it is uncertain whether the appeals procedures do not suffer the same shortcomings of existing notice and takedown procedures. These barriers to access to justice are compounded by the fact that only individual rights-infringements are often recognized as a legal 'harm', rendering more collective or subtle forms of harm invisible (Brown, 1995; Griffin, 2022; Kiss, 1995; Williams, 1991) or making it impossible to contest some content moderation actions.

BROADER APPROACH TO 'ONLINE SAFETY' & 'DIGITAL RIGHTS'

One way to push the policy discussion forward it to map how civil society organisations that have the stated aim to work for the interests or with marginalised groups relate to this policy debate on disparate content moderation. With the increased importance of online communication over the past few decades, a broader range of social justice organisations are relating and contributing to the digital rights policy debate, expanding the range of this debate. However, this movement is still ongoing and far from complete.

Crucially, there is a clear history of the law in general, and regulation on online 'safety' specifically, targeting and disadvantaging marginalised groups such as sex workers or people of colour (Are, 2022). This justifiably **greatly impacts trust in the law or formal procedures to serve their interest**. At the same time, there are potentially several openings in the DSA and the upcoming Directive on combating violence against women and domestic violence, that, in their implementation and enforcement, can still be made to serve these interests. For example, the systemic risk provisions (Arts. 34-35) or the provisions on terms and conditions (Art. 14(4) DSA) already acknowledge a collective dimension. For the DSA to contribute to a healthy digital environment for all, it is essential to map the different needs of different groups regarding content moderation, and foreground social justice or human rights organisations in the debate about the implementation and enforcement of the DSA**.**

Simultaneously, **there are legitimate concerns for the way in which civil society participation is structured** (which organisation, under what conditions etc.) of both elite capture or entrenching platform power and serving corporate interests (Dvoskin, 2021; Táíwò, 2022). As such, it is vital to understanding how organisations that have as their aim to advocate on behalf of marginalised communities online view content moderation and the policy debate, as well as to what extent and under what conditions they would want to engage with it.

The entry point for mapping the positions of CSOs on the protection offered by the DSA will be their perspective on practice and experience with content moderation rather than the legal framework or human rights. What online harms and solutions do these organisations see and how does this relate to the current policy debate?

CONTRIBUTION OF THIS REPORT

The aim of this project is to create an exploratory collection and systematization of the discourse on disparate content moderation based on a series of interviews, in the form of a short report which can inform advocacy, research and the EU policy debate.

The main questions guiding this project and the interviews are how EU CSOs working for the interests or with marginalised or vulnerable groups: (1) conceptualise the causes, effects and solutions of disparate content moderation practices (i.e. downranking, shadowbanning, blocking or (refusal to) removing), and (2) relate to and are involved in the DSA policy debate.

Ultimately through this project I hope to offer a modest contribution to identify intersections and foster possible solidarities in addressing disparate content moderation and; to map what avenues in the current EU policy landscape campaigning or advocacy and research can be pursued to address disparate content moderation that are grounded in the intersectionality, diversity and unequal distribution of content moderation harms.

Mapping how these CSO see online harms and possible solutions as well as how they relate to the EU policy debate can help identify tensions or intersections and can hopefully contribute to fostering solidarities in activism, advocacy and research to challenge disparate content moderation.

After outlining the method and setup of the study, the report reviews respectively the research in content moderation harms and the current EU policy developments. Subsequently, I will present the results of the interviews, first regarding online harms, second regarding access to justice, and finally regarding social justice organisations perspective on the EU platform policy debate. Finally, based on these results possible routes forward will be discussed.

RESEARCH SETUP

Research question:

How do EU CSOs working for the interests of or with marginalised or vulnerable groups:

(1) conceptualise the causes, effects and solutions of disparate content moderation practices (downranking, shadowbanning, blocking or (refusal to) removing), and

(2) relate to and are involved in the DSA policy debate?

METHOD: INTERVIEWS & WORKSHOP

The study consists of **semi-structured expert interviews,** all conducted between April and June 2023 either via video-call or in person. The interviews were transcribed and analysed using a **qualitative thematic analysis method based on a combination of inductive and deductive coding** informed by the existing literature on disparate content moderation using Atlas.ti (Berg & Lune, 2012; Fereday & Muir-Cochrane, 2006). General results were anonymised. However, due to the variety of experts representing different groups some results could be identifiable and, before publication, were checked with the interviewees.

In total, the series consist of around **ten interviews of 45 minutes**. The interviews were divided in several sections: a general section discussing the interviewees experience and expertise, one focussing on online harms and strategies to deal or resist them, one directed at discussing existing and needed access to justice and, finally, a section discussing their relation to the policy debate.

The interview questions were drafted based on an **exploratory literature review** on the existing empirical research and analysis of the relevant legal framework. The results of the interviews were discussed during a **workshop** with the different organisations interviewed as well as other relevant organisations and researchers.

The study passed the Ethics Review, as well the Data Protection Review and received approval for its Data Management Plan from the Faculty of Law of the University of Amsterdam in March 2023. All participants have given their written consent to participation in the study as well as the use of their data.

INTERVIEWEE SELECTION & PARTICIPATION

As the focus of the project is on understanding the disparate content moderation practices the interviewees were mainly EU focused interest groups and social justice organisations that, to some level, engage with the debate on online harms. Interviewees are identified through existing contacts in the researcher's network, identifying the relevant NGOs in the EU, and referential sampling.

Specifically, the selection of the ten different organisations was informed by: (1) the objective to have a wide range of groups represented, (2) organisations that have some affinity with the issue of harmful online content but are not primarily digital rights organisations. Finally, (3) organisations that relate to policy work on an EU level as some part of their work.

The organisations interviewed worked on themes related to: sex work, LGBTQ+ rights, trans rights, women's rights, abortion access, anti-discrimination, safety of female journalists, investigative journalism, and access to justice for online harms. Dedicated digital rights organisations were invited for the workshop and asked for feedback.

Of course, this is a limited sample, and many crucial perspectives are not represented. For example, Muslim, migrant, or anti-racist organisations are clearly missing. As such, the character of this study is exploratory.

The interviewees are **experts from civil society or interest groups working with/for marginalised groups in their professional capacity**. In the research design I considered it important to ensure that participation is contributes to and not solely extracts from these organisations. The hope is for the workshop, bringing together participants, digital rights organisations, and researchers, to contribute to the organisations' agenda's as well as coalition building, and for them to be able to use the report in their work. Still, the project did not involve financial compensation for their time, expertise, and knowledge.

European Sex Workers' Rights Alliance

Abortion Dream Team

Arabi Facts Hub

Dikke Vinger

DATAWO

Coalition Against Online Violence

Transgender Europe

Hate Aid

Women's Link Worldwide

CONTENT MODERATION HARMS

A growing body of scholarship from sociology, communication and media studies as well as legal research discusses the themes of (1) harms a result of content moderation, (2) access to justice for these harms and (3) the specific or disproportionate impact of content moderation on a particular marginalized group. This section will provide a brief overview of the most important lines in this scholarship as a basis for this report as well as a reference for further study.

DISPARATE CONTENT MODERATION

Often online harms are conceptualised as what people experience on the platform, such as harassment or insults (Banko et al., 2020; Scheuerman et al., 2021). Content moderation is then seen as a way to offer remedies for or prevent this online harm (Goldman, 2021). Policy debates mostly discuss how this can be done in such a way to respect freedom of speech while providing online safety.

Content moderation harms refers to the harms that a caused by platform (in)action in any part of their content moderation systems. Think of unwarranted shadowbanning; unjustified removals; sudden downranking; discriminatory rules, demonetisation without cause; not responding to complaints etc. The focus is on what platforms do or fail to do rather than how users harm each other.

However, a clear line of scholarship has shown that the way in which platforms moderate or fail to moderate is by no means neutral. This is referred to as the 'politics of platforms' or the 'governance by platforms' (Gillespie, 2015, 2018; Gorwa, 2019). That the way in which platforms conduct their content moderation can, itself, cause harm is the focus of this report. It is useful to differentiate two types: horizontal and vertical content moderation harms.

Horizontal refers to how the way in which content is moderated opens the door to, or is ineffective with regard to harm between 'users', such as hate, threats, harassment, organised attacks etc. In other words, how platforms facilitate and mediate 'online harms'. Think here, for example, of not removing harmful content even after reporting or, on the other side, facilitating malicious flagging (Are et al, 2023). Vertical content moderation harms refers to how content moderation actions can cause harm when they are applied discriminatory without cause. Examples are deplatforming (Are & Briggs, 2023; Sitaraman, n.d.), over-removal (Haimson et al., 2021; Smith et al., 2021), shadowbanning (Are et al., 2023; Duffy & Meisner, 2022; Myers West, 2018), downranking (Gillespie, 2022), or demonetisation (Caplan & Gillespie, 2020; Goantă & Ranchordás, 2020). Most of these vertical content moderation are what Zeng and Kaye refer to as 'visibility moderation' (2022).

In the following sections, we will dive in to the unequal and discriminatory distribution and impact of these content moderation harms. For now, to understand *how* content

moderation can cause harm, it is relevant to look at the both the norms as well as the processes by which these are enforced.

First, the norms that govern a social media network are drafted by the platforms themselves in their terms and conditions and what are often called the community guidelines. Content moderation is based primarily on these norms, and often only if something is not covered by them legal speech norms, which are often less strict, are applied (Article 19, 2018; Celeste, 2019; Elkin-Koren et al., 2022). Several problems can arise here. First, **the norms themselves can be discriminatory** by, for example, not considering sexist harassment as a problem or over-removing Arabic content under the banner of extremist content but not white extremism (Díaz & Hecht-Felella, 2021; Gerrard & Thornham, 2020; Glitch, UK, 2023; Marshall, 2021). Clearly, this can lead to an over-policing marginalised users' content while leaving content that harms them untouched. Another well-known issue is **vague norms as well as a lack of transparency** as to what the rules are to begin with, to the extent that Sarah Roberts claims that " obfuscation and secrecy work together to form an operating logic of opacity" (2018). Finally, Robyn Caplan points out that the large social media networks "**collapse contexts** in favour of establishing global rules that make little sense when applied to content across vastly different cultural and political contexts around the world" (Caplan, 2018).

Second, content moderation is more than policy and procedures as **these disparate content moderation practices are technologically mediated**. Beyond the simple fact that we can only 'be' on social media platforms through a device and a screen, the enforcement of content moderation policies is done by and through a range of algorithmic systems, from matching, image recognition, natural language processing, each with their own dynamics (Gorwa et al., 2020). The types of systems as well as their implementation affect the entire sociotechnical practice of content moderation and introduce their own logic and harms (Balayn & Gürses, 2021; Binns et al., 2017; Cobbe, 2021; Dias Oliva et al., 2021; Llansó, 2020; Noble, 2018; Tufekci, 2015).

PROBLEMS WITH ACCESS TO JUSTICE

In trying to understand disparate content moderation harms, it is crucial to consider platform actions and their impact through the lens of access to justice. Looking at the available remedies and the different routes people can take as well as their accessibility is crucial as, first, some of these harms are the consequence of a lack of accessible and effective procedural routes and, second, the impact of some others could be softened with adequate access to justice.

In this latter category falls the lack of adequate and accessible ways to challenge the removal of content or the deactivation of an account (Sitaraman, n.d.). There, for example, a large amount of anecdotal evidence of people being unable to retrieve their account expeditiously via the formal channels and having to resort to media pressure before the account is reinstated (Are & Briggs, 2023; Thach et al., 2022) . Another widely noted problem is people not receiving a clear explanation of why their content was removed or their account blocked. This lack of a full understanding of

how their actions or content relate to the content moderation measures give rise to both feelings of powerlessness and frustration (Are, 2022; Zeng & Kaye, 2022), and feed user theories and narratives on what 'the algorithm' does (Cotter, 2023; Karizat et al., 2021; Myers West, 2018). This is especially the case with communities of people who are economically dependent of the platform. Sophie Bishop calls the practice in which influencers share knowledge and pool resources to understand how to maximise visibility on the platform 'algorithmic gossip' (Bishop, 2019).

In the other category, where a lack of effective access to justice is itself harmful, is the widely complaint about situations in which platforms do not respond to a notification (Vaccaro et al., 2020). These include situations in which the platforms do not respond at all, or with an unsatisfactory explanation of why harmful content can stay up (Myers West, 2018).

As mentioned in the introduction, these issues related to a lack of accessible, safe and effective ways to challenge these content moderation harms has to be understood in the context of the historical relation between some marginalised groups and formal authority as well as the law. For example, their existence is itself seen as a threat to 'online safety', such as is the case with sex workers (Are, 2022). Moreover, long and persistent histories of marginalisation as well as the law as a tool for oppression can impact trust in formal procedures. For these and other reasons, not everyone is willing to report these harms with either the platform or follow other formal routes or feels these will not produce any result.

The 'automatic bans' , opaque processes, and impenetrable platform decision-making processes not only add, as mentioned, a layer of powerlessness (Are, 2022) they also drive affected people to develop extensive strategies to deal with 'the algorithm' (Ganesh & Moss, 2022; Vitak et al., 2017). Widespread examples of these strategies changing language to make it undetectable for the content moderation system, or having multiple accounts in order to minimize the effect of deplatforming (Gerrard, 2018; Gillett et al., 2023; Karizat et al., 2021; Vitak et al., 2017). However, another clear strategy is to disengage, by either leaving the platform completely or to stop engaging with the topics deemed controversial (Posetti et al., 2021).

As will be discussed further in the next chapter, the recently adopted EU Digital Services Act could address some of these harms with its stricter procedural and transparency obligations. However, an added dimension to these issues with access to justice for content moderation harms is that a significant percentage of users does not seem to be familiar with the most accessible reporting mechanisms (Appelman et al., 2021).

IMPACT ON SPECIFIC GROUPS

There is a rich literature the specific experiences with and impact of online harms on specific groups such as LGBTQ+ people (Dinar, n.d.; Scheuerman et al., 2018), women (Jane, 2016; Posetti et al., 2021), people of colour, specifically Black women (Glitch, UK, 2023; Marshall, 2021), sex workers, and marginalised influencers (Duffy & Meisner, 2022; ESWA, 2022). Concretely, these harms can manifest in a variety of

ways and the actual impact seems to differ greatly according to the specific context. These harms can clearly impact people socially, in their ability to connect to others, as well as in the extent to which they can participate in public debate (Posetti et al., 2021; Scheuerman et al., 2021). However, content moderation harms can also influence someone's professional life and engender real economic costs.

Clearly, a crucial aspect of these content moderation harms is that they are both not equally distributed and that they disproportionally impact already marginalised or vulnerable groups. Marginalization often refers to the continuous social processes and experience of being placed in the 'margins' of a society, which connects to access to resources, power and capital (Alexander et al., 2015; Hall, 1999; Pearce et al., 2020). To ascertain who occupies the centre and who the margins is, as phrased by Clark-Parsons and Lingel, "a slippery process that reflects and is shaped by a researcher's perspectives and values" (2020, p. 2). The focus of this study is not on a specific group, identity or perspective, but wants to cut across the experiences of both historically oppressed groups that are currently, to different degrees, marginalised in public debate. Necessarily, this report provides a non-exhaustive overview and does not aim to compare different experiences. Furthermore, the analysis itself is, of course, based on the interviewed and, as such, shaped by the limited selection of organisations and perspectives and should be appreciated in that light.

In order to grasp the variety of harms and contexts as well as how they can be compounded by systemic forms of oppression, there is a growing call for an intersectional approach to content moderation (Allen, 2022; Gilbert, 2023) as well as a greater role for affected people in the design of content moderation systems (Are et al., 2023). Primarily based on Black Feminist theory, the central idea to this approach is that these harms exist within a historical context of identity and class based social exclusion, and that these different 'axes' can intersect to amount to greater marginalisation. As such, Gilbert argues that (2023, p.2):

"if moderation is going to avoid reproducing harm, it must account for power."

EU POLICY DEVELOPMENTS

With the increased importance of online communication for modern social, professional and economic life, a broader range of organisations are relating to the EU digital rights debate. More and more organisations and their communities are confronted with online harms and the issues with content moderation central to this report. Simultaneously, there has been a growing awareness of the intersection of technology with discrimination and broader social injustice, paired with conscious efforts to expand the focus beyond privacy, data protection and freedom of expression (EDRi & DFF, 2023). However, it is not clear how this broadening of the field is translating to policy.

Simultaneously, EU platform regulation has been developing rapidly, with more and more responsibility placed on social media firms to guarantee 'online safety'. The most important of these is the new Digital Services Act (DSA) which will function as a framework law setting out the general responsibility of platforms on which other legislation targeting specific online harms.

DIGITAL SERVICES ACT

Although the DSA brings a slew of new obligations, what does not change is system of **liability for illegal content**. Platforms are not liable for illegal content posted by their users if they have no 'knowledge' of it and remove the content quickly when they do, for example when a user reports the content.

Most important for this report are the due diligence obligations and the risk mitigation measures. These **due diligence obligations (left box)** include more transparency and procedural obligations in how platforms conduct their content moderation. Amongst other, platforms will have to clearly explain their rules, provide a statement of reasons for any content moderation action (including downranking and shadowbanning), and have specific faster procedures for trusted flaggers.

The **risk obligations (right box)** mean platforms must conduct risk assessments, implement measures to mitigate these risks, and have this audited by independent organisations. Platforms must check for the risk of their service for illegal content, fundamental rights, civic discourse/elections, and gender-based violence (Article 34). The mitigation measures can include almost anything from the design, the terms & conditions, or how they moderate content

DUE DILIGENCE OBLIGATIONS

- **Transparency in terms and conditions (14)**
- **Statement of reasons for content moderation (17)**
- **Trusted flaggers (22)**

RISK MITIGATION MEASURES

- **For Very Large Online Platforms**
- **Identifying risks (34) and implement measures to mitigate them (35)**
- **Risks for illegal content, fundamental rights, elections and gender-based violence;**
- **Subject to independent audits (37)**

(Article 35). How this will be implemented in practice is still quite unclear.

As to **access to justice** (box to the left) the DSA codifies the notice and action procedures and obligates platforms to respond. Moreover, the big platforms also must create an internal complaints handling mechanism for most content moderation actions. People will also have access to independent out of court dispute settlement bodies to bring these complaints to and they can file a wide range of complaints at their national supervisory authority, the digital services coordinator.

The design of the big platforms, their terms and conditions, and how they moderate are still mainly up to the social media firms themselves. Although the DSA can potentially mean a great improvement, a lot depends on the implementation and enforcement, especially when it comes to content moderation at scale.

DSA ACCESS TO JUSTICE

1. **Notice and Action Mechanism (article 16)**
2. **Internal Complaint Mechanism (article 20)**
3. **Out of Court Dispute Settlement (article 21)**
4. **Complaint at the Digital Services Coordinator (article 53)**

OTHER EU POLICY IN-THE-MAKING

Another relevant development is the aforementioned Directive on combating violence against women and domestic violence which in its current draft also looks at online gender-based violence. Specifically, it includes a provision (Article 25) requiring member states to have an interim judicial procedure, which are generally quite quick, to remove non-consensual intimate images or stalking or harassing material.

Moreover, in the negotiations over the upcoming AI act, the European Parliament has voted to designate big social media platforms' recommender systems as a 'high risk AI system' which means it will have to comply with the technical, transparency, and oversight requirements of the Act. If this makes it to the final text, this will most likely interact with the risk-assessment and -mitigation obligations under the DSA.

All this again raises the question how the concerns of this broader range of organisations getting involved in the digital rights debate are included to these policy developments, and how they themselves relate to this policy debate. This leads us back to the main question of this report, how do these NGO's see these developments, where can we possibly find solidarities, and to what extent can we still use openings in the policy developments to leverage?

RESULTS: Online Harms & Strategies of Resistance

This chapter reports the results of the interviews specifically on the topic of online harms and strategies of resistance. The interviewees were asked questions on their perspective on and experience with content moderation harms and how it impacts different communities as well as the strategies they have seen people develop in response.

DISPARATE CONTENT MODERATION

All interviewees, to different degrees, agreed that they see differences in how groups are treated on the big social media platforms, either based on their own experience or what they have seen in their communities. One interviewee commented that "sometimes it's very obvious the bias of the platforms".[1] Moreover, the interviewees clearly identified that the disparate treatment they are seeing disproportionally harms already marginalised or vulnerable groups. This is clearly in line with the scholarship on this topic as discussed in the previous chapter. One summarized this as follows:

"The more marginalized they are the more of this [harassment] they get. And at the same time the more disregarded these communities are by the platforms and content moderation policies and the algorithms."[2]

Beyond this general observation that marginalisation is reproduced through these content moderation harms, the interviewees emphasized the experiences their specific communities.

One clear theme that emerged was **the treatment of sex workers** on the platform who disproportionally experience unjustified deplatforming, content removal and shadowbanning – without having violated the platform terms. When asked whether another interviewee thinks the platforms want to provide space for sex workers, they responded with "I think they don't. I think that there's like mostly like the meta group companies like this, very American mindset, puritan, puritanism, static and whatever, you know, has to do with sex and sex workers. No, you know, for them, we don't, even

[1] 3:5 ¶ 16 – 17 in interview 3.
[2] 9:5 ¶ 31 in interview 9.

I don't think, we don't even exist."[3] This seems especially the case for female sex workers: "Yes, we recognise this. We can see that many femininities, like for instance sex workers or femininities, who circulate just sexual, consensual sexual content, are treated differently from the perspective of content moderation."[4] Interviewees reported that **fat people similarly feel unwelcome** based on the disparate content moderation they have experienced. To them it seems as if "the platform loses quality or loses value if a lot of fat people are visible".[5]

Other interviewees similarly reported that **women often experience disparate content moderation**. One of the support NGO's, clearly sees "that many marginalized group groups reach out to us, and we can also tell that the majority is women."[6]

Another theme that emerged was the **disparate treatment of non-English and non-Western content**. "But you know, I would say the response is different if you're from France than if you're from like a French speaking country in West Africa, for example."[7] Similarly, another interviewee commented: "this is an equality in in my opinion it's when you are focusing on the English content only on the social media"[8]

Although the interviewees primarily spoke to their areas of expertise and their communities many of them **emphasized the intersectional nature** of these types of harms. For example, "also in an intersectional feminist way of thinking, because I can see the difference in white femininities and refugee femininities or migrant femininities they are under severe risk, so I can imagine that probably the same instances are happening in other countries in other Member States."[9] Another interviewee commented:

"You can see that we are suffering exactly the same oppressions and struggles there. It's a very intersectional fight for me."[10]

PLATFORM (IN)ACTION

Moving on to the concrete content moderation actions, the interviewees all gave several examples of the type of online harms we see that are in line with the scholarship as discussed in the previous chapter and add some additional depth.

Blocking of accounts or deplatforming, was mainly mentioned in the context of sex work, NGO's working on reproductive health, journalists, transgender activists, and

[3] 3:40 ¶ 180 in interview 3.
[4] 1:1 ¶ 18 in interview 1.
[5] 4:32 ¶ 88 in interview 4 *[translation by author].*
[6] 5:3 ¶ 11 in interview 5.
[7] 7:15 ¶ 37 in interview 7.
[8] 10:7 ¶ 20 in interview 10.
[9] 1:39 ¶ 80 in interview 1.
[10] 6:22 ¶ 65 in interview 6.

non-Western accounts.[11] “That happens again and again happens and happens to everywhere, to everybody you know. I think I don't know any sex worker that hasn't been taken their account down at least once. They are coming for us.”[12] Another interviewee commented that, with regard to TikTok, “we also see it with other people that we work with. Their accounts are blocked temporarily or that their content is just blocked, especially with activists in gender topics.”[13] And another "Yeah, that happens a lot. So, when the account gets taken down the journalist has to recover that account”[14] Blocking of accounts was also mentioned in the context of female journalists getting their accounts hacked.[15]

Most interviewees mentioned and gave concrete examples of a double standard in the **removal of content**. One clear instance was when a women’s rights group publicly challenged the sexist comments made by a famous athlete and their response was taken down while “the sexist tweets were not taken down, you know. This is totally contradictory, and I think that this is an obvious example where we can see that how content is treated differently by some groups of people.”[16]

Platform (in)actions mentioned:

Unjustified deplatforming or content removal, vague and unclear norms, lack of transparency, lack of appreciating context, shadowbanning, not removing harmful content, data harm, malicious reporting and harmful platform affordances.

For sex workers, fat people, and Black people interviewees connected this mainly to nudity. One interviewee commented “So then they are being censored by these algorithms like more.”[17] When talking about a fat activist often posting pictures showing skin on Instagram, one interviewee explained: “there are a lot of thin people on Instagram who take half-naked pictures really a lot. That's never a problem and often those photos are even more extreme than the ones she takes. So, it's just weird that those photos are deleted, and I wonder whether that's because people report the photo or whether the platform itself registers and deletes it in a certain way.”[18]

This last point connects to another clear theme that emerged: **harm emerging from the platforms’ norms and automated enforcement**. Interviewees mentioned three modalities. The first is that interviewees reported that the platform norms are **vague, unclear, or unknown**. “usually is like it violates our code of conduct, and for us, it's actually quite vague what it means, because sometimes it means you can revert it just because somebody complained it's an algorithm, so they block you like it's very unclear.[19] Similarly: “the problem is all these things are very untransparent […] so it is quite unfair to expect, you know, users to even if they want to adhere and accept

[11] Interviews 2, 3, 4, 6, 8, and 10.
[12] 3:11 ¶ 33 – 34 in interview 3.
[13] 5:9 ¶ 23 in interview 5.
[14] 7:40 ¶ 123 in interview 7.
[15] Interview 7.
[16] 1:5 ¶ 20 in interview 1.
[17] 8:21 ¶ 49 – 50 in interview 8.
[18] 4:31 ¶ 35 in interview 4 *[translation by author].*
[19] 6:6 ¶ 15 in interview 6.

your rules terms of reference. It's not logical to expect them to basically accept those rules because it's not clear."[20]

A second way in which the norms can create harms is with regard to the substance of the norms themselves and the discriminatory or problematic **values embedded in these platform rules**. "Platforms have struggled to understand what constitutes sexual harassment"[21], and "social media platforms, they tend to just say sex work, pornography, it's illegal even if it's not illegal because it's not socially acceptable."[22]A third and closely connected way in which the interviewees mention platform norms and their application creating harm is their **inability to appreciate context**. In line with existing scholarship, one interviewee remarked: "What we do see is a lack of cultural understanding around particular issues with regards to online abuse, for example, particular words or expressions or phrases which, when translated into English, don't sound that bad, but in the native like the original language, are particularly offensive. You know, they don't have staff to understand the cultural context.[23]

Further, all interviewees reported **shadowbanning** to be a significant problem, as indicated by the quote to the right.[24] This is directly connected to the opacity and lack of explanations: "They suddenly realize their visibility massively decreases and they don't know why. But then it usually happens as retaliation -it's their interpretation- "[25]

"There are more sinister ways that platforms impose these kinds of harms, which is, not really in your face, like deplatforming you directly know, deleting your profile or deleting your pictures, but things like shadow banning which you can't really prove that it's happening, but you know that it's happening"

As already mentioned in this chapter, not **removing harmful content** is one of the content moderation harms the interviewees identify.[26] Related to this is how interviewees note that platforms insufficiently act against malicious and organised reporting.

Within the same theme, in more general terms, more than half of the interviewees emphasized how they feel platforms in a broader sense **facilitate and mediate the harm between users**: "these platforms provide the space for the harm".[27] This is repeatedly phrased in terms of platform responsibility: "they do not do anything in order to avoid any further harm",[28] "that horizontal harm it's basically enabled by the platform policies and or the lack of the

[20] 8:17 ¶ 35 in interview 8.
[21] 7:18 ¶ 45 in interview 7.
[22] 8:24 ¶ 63 in interview 8.
[23] 7:17 ¶ 41 in interview 7.
[24] 8:15 ¶ 34 in interview 8.
[25] 2:8 ¶ 45 – 47 in interview 2.
[26] Interview 1 and 3.
[27] 1:10 ¶ 48 in interview 1.
[28] 1:11 ¶ 50 in interview 1.

platform policies on you know certain topics […] in that sense, everything is the responsibility of, you know, the platform",[29] and "generally I think in terms of harassment, receiving harassment on the Internet, there is a lot more that they could do for that."[30] One way in which platforms facilitate online harm in the broader sense identified by interviewees is the design and subsequent **platform affordances**, such as the possibility of taking screen shots and collecting evidence,[31] or removing harmful comments under your own posts.[32]

Then, finally, half of the interviewees mentioned **data harms**. This refers both to the dangers associated with making personal information public as well as issues around platform data collection and mandatory phone number registration.[33]

MODALITIES OF IMPACT & STRATEGIES

As to the diverse types of impact these harms have, again this study's findings seem to be in line with most scholarship outlining the different ways in which these platform actions create social, economic and political harm. But also psychological, as one interviewee stated: "trying to see the terms, how to avoid the algorithm is really tiring. Is this something that it's also another cost and another way to burn out people.[34] Two interviewees also emphasized that online harms can also result in real-life physical harm.[35]

Clearly, the actual harm that people experience and how it impacts them is not static or one dimensional but depends on an entire range of factors. What came out clearly in the interviews are the issues people face when they are using the platform professionally and are subject to these content moderation harms. Especially sex workers as well as female journalists seem to experience (various levels) of **professional precarity** due to being dependent on the platform. For example, when discussing deplatforming one interviewee commented: "The journalist has to recover that account, and that's time, and time for them is money because quite often they're freelancersthat's the way they connect with their audiences."[36] And another: "So yes, it basically means resources as well. And if you want to engage. It's [...] a lot of work actually to do it in a way that you understand how the platforms work and what are the pluses and minuses of it."[37]

Another crucial dimension is the extent to which people are denied the ability to occupy space in public debate or media outside of the platforms. As one interviewee put it succinctly: "we are marginalized in the society. We're also marginalized in the digital

[29] 8:10 ¶ 29 in interview 8.
[30] 7:22 ¶ 61 in interview 7.
[31] Interview 1.
[32] Interview 4.
[33] Interview 10, 8, 6, 5, 3.
[34] 3:9 ¶ 23 in interview 3.
[35] Interview 10 and 6.
[36] 7:40 ¶ 123 in interview 7.
[37] 6:10 ¶ 28 in interview 6.

space."[38] For example, Instagram's removal of the swimwear pictures of a fat activist compounds the problematic media culture in which fat people are invisible against which the activist is fighting.[39] In a distinct but similar way abortion activists as well as sex workers are **invisibilised in public debate** due to both the stigma as well as the unsafe legal status. This means the venues in which they can advocate, campaign and form community are already limited, and the content moderation harms can function to close of the most effective medium.

Moreover, the way in which these content moderation harms impact someone seems to be directly related to the **strategies they employ in response**. On such response is **disengagement,** often referred to as **the chilling effect** of these harms. Interviewees mentioned people leaving a platform, searching for an alternative or avoiding certain topics. When discussing the chilling effect of these harms, one commented "So I mean it works. That's the thing about online abuse. It really does work."[40] Similarly, another stated "there are lots of strategies they use, even you know, some of them are quite bleak. Like just self-censor and you know not do anything that's you need to do."[41]

Interviewees also mentioned **resignation**, acceptance that they will continue to experience these harms, as they do not always have choice to leave the platform or disengage.[42] For example when they are professionally on the platform or when they try to reach communities via the platform. Regarding female journalists: "I think you know most of them are just resigned to the fact that they'll just continue to get harassment."[43] Or women's rights activists "For many of them just deleting their Facebook pages is not seen as an option."[44]

There are also people who actively try **to challenge or subvert** content moderation. This ranges from reporting the issues to the platform, seeking media attention, or to more subversive strategies such as gaming the algorithm through adopting adjusted language or through having multiple accounts.[45] With regard to abortion activism: "We must indeed focus on subversion and basically knowing how they work. So, it's sort of mouse and cat […] game."[46] And sex work: "trying to use the platform's own infrastructure against it. They do, but it is always a trial, and you know, learn. Sometimes it doesn't work."[47] Doing this takes a lot of effort, "it's a full-time job basically",[48] and takes its toll.

---

[38] 3:41 ¶ 153 in interview 3.
[39] Interview 4.
[40] 7:28 ¶ 67 in interview 7.
[41] 8:31 ¶ 74 in interview 8.
[42] Interviews 2, 3, 5, 6, 7, and 9.
[43] 7:25 ¶ 65 in interview 7.
[44] 2:12 ¶ 65 – 66 in interview 2.
[45] Respectively, interviews 3, 6, 8, and 10, and interviews 3, 6 and 8.
[46] 6:27 ¶ 73 in interview 6.
[47] 8:33 ¶ 78 – 79 in interview 8.
[48] 3:42 ¶ 100 in interview 3.

Finally, other strategies mentioned is to **collect evidence** of the harm,[49] and to create **broader awareness and to teach others** in the community about how to avoid these harms.[50]

LESSONS LEARNED

CONTENT MODERATION HARMS & STRATEGIES

1

First, large **heterogeneity in both experience and impact** that seem to be intersectional in the sense that they depend on the wider context of someone's life as well as broader societal oppression. Content removals have more impact if you have no other way to reach people. Deplatforming does more harm if your dependent professionally on the platform.

2

As to the harms themselves, already this small group of interviewees indicated a **broad range of harms** beyond the most known ones such as deplatforming, removal, or shadowbanning. The norms themselves, the vagueness or seemingly arbitrary application, data harms as well as the affordances of a platform are all instrumental.

3

All organisations interviewed indicated that they and their communities experience online harassment. However, not all communities indicated experiencing removals or deplatforming as much. Especially, those groups that already have an adversarial or strained relation with the law seemed to experience this the most, such as mainly sex workers and abortion activists. This seemed to affect **their relation to the platform**, considering the platform more of an active political actor, rather than an unreachable corporation.

4

As to the **strategies people develop** to deal with these hams, this seems to relate to several factors: (1) the type of harm experienced, either mainly by platform action or by harassment, (2) the impact this has, whether or not professional for example, and (3) the wider relationship with the law. This is clearly seen in, for example, sex workers developing sophisticated and extensive strategies to avoid content moderation harms or abortion activists who try to avoid counselling on the platforms.

[49] Interview 1 and 2.
[50] Interview 2, 7, 1, and 5.

RESULTS: Access to Justice & Systemic Causes

This chapter reports the results of the interviews specifically on the topic of access to justice and the systemic causes underlying content moderation harms. The interviewees were asked questions on their perspective on the underlying causes for disparate content moderation as gaining insight of the broader context in which these harms arise can inform possible solutions. Interviewees were also asked about their experience with the available routes and remedies to deal with / end these harms, and what effective improvements would be.

CAUSES OF DISPARATE CONTENT MODERATION

When asked what they think the wider causes for disparate content moderation is, all interviewees connected to broader systemic causes outside just the platform itself. All mentioned the corporate nature of the platform and/or wider societal injustices.

Most importantly, interviewees emphasized how, with the dynamics of disparate content moderation, **platforms seem to reproduce existing social inequalities** and forms of oppression: “it's a reflection of society, you know, I think that the that whatever happens in the online space is, is a reflection of what happens everywhere”,[51] and “It's purely normative and purely in a kind of societal sphere and not technological.”[52] One interview commented:

“they replicate systems of oppression that are already present in society”.

In this context, the same interviewee emphasized the intersectionality of these harms “you can see we are suffering exactly the same oppressions and struggles there (…) I do think it's very intersectional agenda of the stigmatized practices, if you analyse

[51] 3:18 ¶ 55 – 56 in interview 3.
[52] 6:36 ¶ 97 – 101 in interview 6.

them from public health perspective, discrimination and human rights standards perspective."[53]

Further, interviewees also underlined the role of law in these dynamics and how the main obstacle they are facing is **the illegality of their work**. For example, "criminalized communities like sex workers whose whole lives are considered illegal"[54] and nothing will change "as long as the very big politics of the European Union is paradigmatically against sex work"[55] Or the issues facing access to abortion activists have to be addressed on the level of "more liberty in abortion and freedom and decriminalization."[56]

Most interviewees also mentioned **the wider context of capitalist structures** and how the corporate incentives, or the 'business model' is a driving force behind the unequal content moderation. "It's just basically the larger umbrella of capitalism (…) That as long as they can brush it under the carpet, because that's like it's more, leads to more money or political votes or whatever, this is going to be happening."[57] And: "I think this sounds very obvious or not very common, but I think it's the profit."[58] Especially with regard to content on sensitive or stigmatized topics, interviewees tie these profit motives to risk:[59]

"You're such a small fish it seems in the capitalist scheme of it. That they just very risk averse. So usually they would just say, oh, it's not worth our risk."

More practically, interviewees also mentioned that the enormous scale of operation as well as the lack of perceived diversity among developers as factors.[60]

ACCESSABILITY & EFFECTIVITY PROCEDURES

Many of the procedural problems brought up by the interviewees are in line with scholarship on platform access to justice as well as the policy debate around the DSA. The main problem interviewees kept emphasizing is **a lack of response from the platform** after either reporting harmful content or challenging a content moderation decision: "they haven't received a reply on this report. Nothing was taken down"[61] and "People try to report and try to do something about it, but many of them are frustrated

[53] 2:18 ¶ 95 in interview 2.
[54] 8:39 ¶ 101 in interview 8.
[55] 6:34 ¶ 93 in interview 6.
[56] 6:44 ¶ 127 in interview 6.
[57] 6:34 ¶ 93 in interview 6.
[58] 1:28 ¶ 103 – 109 in interview 1.
[59] 6:18 ¶ 53 in interview 6.
[60] Interviews 2, 3, 4 5, 9, 10.
[61] 1:8 ¶ 28 in interview 1.

in the aftermath when they see that nothing happens, they don't receive a reply."[62] Another issue was people getting lost in automated complaint procedures: "It's very hard to get human attention on those platforms (…) they automatize their jobs as much as possible and when you reach out to a support person is hard to get them"[63] and "it's very difficult at the end to get to somebody that's an actual human being in a decisive position.[64]

Further, as also indicated in the previous chapter, a lack of information and explanation about why content moderation actions are taken is also an obstacle: "I'm just feeling like this is happening and I don't even know. Why or what to do about it?"[65]

"[People experiencing online harm] want to do something about it, but then all these obstacles come. […] it feels insecure, it also feels […] like [a] David against Goliath situation and they say things like "OK, there are these platforms with their powerful lawyers, they have all the money. How could I succeed?"

A major problem that features in most interviews is the **inaccessibility of the procedures offered by platforms**: "Maybe they're out there, but then they're in a format that is not accessible. It's like information is out there, but then you look at it and then you don't understand what you're looking at. People cannot really access it does not take into account that people are different. It's not very accessible. I would say it's not easy to find.[66] And "you don't really have a redress mechanism either you know and they are not transparent about how these mechanisms work."[67] Compounding this is the fact that these procedures seem to change quite rapidly which poses a challenge for toolkits and support documents: "maintaining it and updating it is always an issue. That's because resources go out of date."[68]

All these issues form significant obstacles for both people and organisations to address content moderation harms when then arise. Especially for individual people, this can be **extremely disheartening**, as can be read in the quote to the left.[69]

AVAILABLE NGO SUPPORT

The **support of an organisation** with helping people navigate the different procedural routes as well as broader support is deemed very important and emphasized by most interviewees: "they need some legal aid and they don't have the funding (…) we have to give immediate support and immediate remedy for their situation, which is why we

[62] 5:14 ¶ 59 in interview 5.
[63] 9:18 ¶ 101 in interview 9.
[64] 6:15 ¶ 49 in interview 6.
[65] 2:27 ¶ 116 in interview 2.
[66] 2:22 ¶ 103 in interview 2.
[67] 8:26 ¶ 63 in interview 8.
[68] 7:7 ¶ 25 in interview 7.
[69] 5:14 ¶ 59 in interview 5.

give emotional support in the 1st place, but also cyber security, counselling, communication, counselling."[70]

But, for example, mediation through an NGO can also provide legitimacy and get platforms to take the problem seriously.[71] Especially as one interviewee indicated that much responsibility of addressing harms in the current legal framework is placed on the individual: "the fact that there is no liability unless something is reported already tells us about what is expected and who is in charge of creating the liability and [of] somehow flagging illegal content. It's all up to the users."[72] Connected to this, several interviewees also emphasized the importance of collective actions to address issues that are either beyond the capacity of individuals or where do you not want to burden them with.[73] One also expressed the hope that the DSA could go some way in providing this possibility.

Then, the interviewees indicated that matters greatly whether you have contacts at the platform. This is the true for organisations offering support: "Since we have a relationship with the platforms, we could solve it because it was usually in the reporting the context that was missing and also when they know us as trusted partners, we could help with this."[74] But also for individuals or NGO's themselves trying to remedy a content moderation harm: "I remember that when they blocked (…) the page a couple of years ago, then we would go through people who we knew worked for Facebook."[75] The lack of structured access to platforms by NGOs trying to support people is, consequently, also flagged as a major problem: "There is a lack of communication between these NGOs and these initiatives and these big companies that's the problem."[76]

"It's the criminal criminalisation of sex work [...] that really prevents sex workers from looking for any kind of retribution. And because of this constant past experience of being undermined, not being able to get justice in any circumstances, not being able to express yourself [... they] are not really used to going through the usual mechanisms that are available for others. They just don't have belief that any kind of governance that's built within any structure will work for them."

STRUCTURAL HARM OF LAW

One interviewee pointed out that taking legal action can lead to secondary victimization due to the structure of the law: "When they go to the court, they realize that there are many problematic issues in the judicial system, because of the criminal legislation that is based on penalties and not

[70] 5:2 ¶ 7 in interview 5.
[71] Interview 7.
[72] 5:25 ¶ 88 in interview 5.
[73] Interview 1, 5.
[74] 5:5 ¶ 15 in interview 5.
[75] 6:16 ¶ 49 in interview 6.
[76] 10:6 ¶ 20 in interview 10.

on restorative justice. In (…) image-based sexual abuse, the perpetrators are paying the fee and then they go out. Nothing has changed."[77]

Another important point that kept coming up was the trust in formal procedural routes and willingness to use them. Criminalization and structural marginalization leads to **systematic mistrust in formal routes** with the people subjected to this and gives them no reason to believe formal procedures or institutions will not do them further harm, let alone help them, as the quote to the right clearly explains.[78] The interviewees indicated this is mainly the case for sex workers and access to abortion activists.[79]

Another important dimension is how regulatory solutions for a specific type of harm are either one dimensional or seemingly weaponized to exclude other groups. One clear instance is when measures to counter gender-based violence and human trafficking or ensure child safety act to harm sex workers or abortion access content.[80]

LESSONS LEARNED

ACCESS TO JUSTICE

1

Disparate content moderation is seen as **a result of wider systems of social oppression** that are replicated and reproduced both on the platform and by the platform itself through its content moderation actions. Within the corporate logic of the social media firm, stigmatized content could be conceived us as a risk and suppressed to minimize that risk, without (sufficient) consideration for the harm that this causes.

2

In line with existing research, **major hurdles to access to justice** are lack of clarity on how content moderation works and what the norms are, as well as a lack of response from platforms to notifications, a lack of explanation as to why content moderation actions were taken and, finally, available procedural routes are unclear and inaccessible at least to the extent that people are unfamiliar with them and have trouble navigating them when they need to.

3

Crucial is **the support from an organisation** in finding the right procedural route, both legal routes and those offered by the platforms themselves, as well as broader support in navigating and dealing with the platform and the harm. These organisations

[77] 1:27 ¶ 97 in interview 1.
[78] 8:35 ¶ 91 in interview 8.
[79] Interview 3, 6, and 8.
[80] Interview 3, 6, and 8.

are also important in leading possible collective actions to offload the individual people experiencing this harm.

4

Success in challenging and remedying content moderation harms is often dependent on having **contacts within or access to the platform** beyond the standard notification procedures. This contact is dependent on the voluntary cooperation with the platforms, and, besides the real threat of arbitrary treatment, this might also feel to support organisations as if it could limit them in their advocacy efforts.

5

Willingness to engage with, **or trust in formal procedures**, both legal and with the platforms, must be understood within the context of potential wider criminalization and legal stigmatization of marginalised groups. This is also connected to the collection of personal information or visibility in these platform procedures.

RESULTS: PLATFORM REGULATION & SOCIAL JUSTICE

In the final part of the interviews, participants were asked about their relation to the EU platform regulation and broader digital rights debate. As discussed, all organisations interviewed, either in their work or in the support of their communities, are confronted with social media and its possible harms. However, except for the online hate advocacy and support organisations, they are not intimately connected to the policy and regulation. The interviewees highlighted four **obstacles to participation**: lack of capacity, perceived lack of expertise, inaccessible spaces, and not seeing the platform as an actionable actor.

The first and straightforward barrier to participation in the EU policy debate is **a lack of capacity**. For these small organisations it is also a matter of priorities: "there are lots of topics you want to consider as an organisation and you also just have to choose something."[81] Moreover, as indicated clearly by the interviewees, following and engaging this debate requires "time and resources and we are already on top of our capacity."[82] One interviewee also indicated that the reporting obligations put on trusted flaggers in the DSA are excessive: "we will see if we can fulfil these obligations and if not then we will just lose our trusted flagger status."[83] Interviewees saw capacity as also referring to the ability to devise a strategy: "Where I do find this intersection between digital rights and" [84] their organisation's focus.

This connects directly to the second clear barrier: **a perceived lack of expertise**. Interviewees reported clearly that either they or their community feel like they do not know enough about the field: "it's like a whole new area of work, it feels really technical and you're not sure how much you have to contribute."[85] This technicality is also reflected in the level of the language used: "And the language is really difficult, even if you know English, there's a specific language that people use in policy environments which is not accessible.[86]

This lack of perceived expertise and capacity is compounded by the idea that it is nearly impossible to fight these platforms. Often, **platforms are not seen as a challengeable actor**: "I do think it is seen as an important issue, because it really does

[81] 4:18 ¶ 103 in interview 4 *[translation by author].*
[82] 9:14 ¶ 52 in interview 9.
[83] 5:26 ¶ 89 – 94 in interview 5.
[84] 6:25 ¶ 73 in interview 6.
[85] 2:29 ¶ 123 in interview 2.
[86] 8:43 ¶ 117 in interview 8.

have a lot of impact on a large scale. But also, that as an activist issue, it offers very little hope for improvement."[87] This also clearly connects to the 'David versus Goliath" feeling people had in pursuing access to justice for content moderation harms as discussed in the previous chapter.[88]

Despite these different obstacles, more and more social justice organisations are connecting to the digital rights debate and entering these spaces. What they find there, beyond gatekeeping through language and perceived technical complexity, is that these policy spaces themselves are experienced as **very inaccessible and unwelcoming**: "we experienced these spaces as quite white, very Eurocentric, and very inaccessible in general."[89] This sentiment is equally expressed in the quote to the left. [90]

"Many social justice organizations who are not primarily digital rights organizations are entering into this space. […] I think it's been needed for a very, very long time […] But it is definitely not equal footing in terms of accessibility of these spaces".

Though these obstacles are problematic and severe, even with abolishing them not all organisations will want to focus their energies on EU level advocacy. A form of division of labour with different focusses could be very productive, as one interviewee explained: "lobbying or advocacy with EU institutions is also something that we do not do. We work with others who are really good at it, so sometimes we collaborate when […] but it's an area that it's not our expertise or our focus."[91]

However, from the results in the past two chapters we can also clearly see several potential obstacles for wider collaboration, coordination and solidarity within the platform regulation space. These **barriers for wider cooperation** are: differing relations with the platforms, possible tensions in demands, and differing lobbying and legal positions. Each will be briefly discussed in turn. However, even more than the other results, these very much are a function of the specific organisations participating.

From the results on contentment moderation hams, we already saw how experiencing different types of harms gave rise to a **different relationship with the platform**. Specifically, where all organisations indicated that they and their communities experience online harassment, not all experience removals or deplatforming as much. The organisations that experience these active content moderation harms, had a more adversarial relationship to the platforms. Especially those groups that also experience wider

[87] 4:23 ¶ 114 – 115 in interview 4 *[translation by author].*
[88] See p. 25.
[89] 8:43 ¶ 117 in interview 8.
[90] 8:42 ¶ 113 in interview 8.
[91] 2:31 ¶ 136 – 137 in interview 2.

criminalization saw the platforms as a potential negative political actor rather than an unreachable behemoth corporation.

OBSTACLES TO PARTICIPATION	OBSTACLES TO COLLABORATION
LACK OF CAPACITY	RELATION TO THE PLATFORM
PERCIEVED LACK OF EXPERTISE	TENSIONS DEMANDS
PLATFORM AS ACTOR	POSITION IN LAW & LOBBYING

These differences in the type of harm people are confronted with and the distinct relationship with the platform that arises from it, reflects in the expectations and demands mate. This can translate to **possible tension between demands** of different groups, of wanting platform to take more responsibility and moderate content more actively or wanting them to refrain from policing their content.

For example, interviewees signalled the need for more transparency on platform norms,[92] but one interviewee also clearly nuanced this: "then the trap here is that I don't advocate that they have to put everything super clear, because this is also a legalistic kind of thinking […] also sometimes is not helpful."[93] Moreover, regulatory solutions for a specific type of harm can be weaponized to exclude other groups. One clear instance are measures to counter gender-based violence, trafficking, or ensure child safety that harm sex workers or abortion access content.[94]

Finally, in the context of access to justice, we discussed how communities and organisations can have different relations to the law as well as institutions due to historical or structural criminalization. In the context of the EU policy debate. This can translate to a significant **difference in lobbying position**, where some communities are taken more seriously than others. One interviewee commented: "[We] build lots of coalitions and push other organizations to take positions on behalf of [our community]. We can create this joint movement basically, which did quite good. […]. [Our community is] really kicking above their weight and that makes us happy." [95] But at the same time:

"it only happened after we took […] mainstream organizations, more acceptable organizations towards our side. So, it's really them they are listening to, it's not really us […]. So that's really annoying."

[92] Interviews 3, 4, 6, 8, and 9.
[93] 6:40 ¶ 119 in interview 6.
[94] Interview 3, 6, and 8.
[95] 8:42 ¶ 113 in interview 8.

CONCLUSIONS

& LESSONS LEARNED

Taking stock of all these results so far, we can discern five clear lines to base both further research and potential policy on, specifically the risk assessment and codes of conduct in the DSA. By centring the perspective of social justice organisations who are confronted with content moderation harms but are not necessarily engaged with the platform regulation debate, the following tentative conclusions can be drawn:

1

First, there is a large **heterogeneity in both the experience and impact** of disparate content moderation harms, especially in the dimensions: type of (in)action, impact, and people's response. **Actions** range from unjustified removals, vague norms, platform design, and a lack of protection against harassment. Their **impact** varies enormously, whether people use the platform professionally, whether they are excluded from or stigmatized in other media, and the broader context of intersecting social oppression. This means context and identity matter immensely for the impact these actions have. **Strategies** developed in response vary from disengagement, or resignation to sophisticated subversion strategies targeting the algorithmic enforcement or challenging and fighting the unjustified content moderation actions.

The diversity in experience and impact of content moderation harms are not explicitly recognised in the DSA. More sensitivity to context could be included in the codes of conduct or risk assessment requirements.

Especially important factors are whether someone (b) uses the platform professionally, (b) is part of a marginalised group that is targeted and impacted most (e.g. abortion, LGBTQI+ or fat activists, and sex workers), as well as (c) is in an region or uses a language underserved by the platforms.

2

Second, in thinking through *how* these different experiences and contexts can be appreciated in concrete policies, it is important to understand how they intersect. Understanding the **broader intersectionality** of these harms means that we have to take seriously how content moderation harms are connected to wider social oppressions and surface in the context of both a business model and a sociotechnical system. Consequently, solutions cannot be fully reduced to only technological or policy terms,

but must be accompanied by a broader vision of the sociopolitical context. Failing to do so means policy solutions can reproduce these harms as one-dimensional policy solutions can exacerbate tensions between **different needs and demands** made of the platform. Crucial in this regard is to also appreciate is the **inaccessibility of the EU platform policy debate** as well as the difference in lobbying capacity for organisations campaigning against criminalisation.

The intersectionality of these harms, how they relate to power as well as their broader societal context must be considered by policymakers, academics, as well as by NGOs in campaigning for change. This means avoiding one dimensional policy and working towards broad coalitions.

3

Third, in this variety of experiences some groups or organisations that experience disproportionate unjustified removals, blockings, or shadowbannings can develop a **negative relationship with the platform**. This can tie in with a mistrust of institutions and formal procedures as a result of historical oppression, stigmatization, and criminalization. As such, a solely legalistic approach with placing more responsibility on platforms, more detailed data gathering on users in a complaint procedure, and more detailed rules does not necessarily offer an accessible option.

A solely legalistic approach is insufficient to support communities who are apprehensive, with good reason, about formal procedures.

4

Fourth, in terms of **access to justice**, at the same time, clear and **accessible procedural routes** as well as clarity on content moderation norms was broadly felt as lacking. Crucial is the **support of an organisation** both as to navigating the procedural landscape and, more broadly, in dealing with the harm. Moreover, success in challenging and remedying content moderation harms is often dependent on having **contacts within or access to the platform** beyond the standard notification procedures which is dependent on the platforms' goodwill. This fosters a lack of transparency, inequality in treatment, and could hold these organisations back in their campaigning against disparate content moderation.

To ensure effective access to justice without over-formalising, platform procedures should be co-designed by people most affected by them.

Further research is needed on what type of organisations and funding structures fits specific contexts best, and how to ensure platforms engagement is not voluntary and precarious.

5

Fifth, there is a vital need for the protection of **grey zone content** from criminalized and marginalized groups. Due to the stigmatization and exclusion faced by these groups such as, depending on context, sex workers, trans activists, or abortion access activists, it is crucial for their freedom of expression and public debate that their content that society does allow is protected. This means extra vigilance is needed to facilitate and protect the content of these groups that is legal and not in violation of platform norms. Social media are a crucial space to advocate and find a community for many of these groups and due to their vulnerable position, extra care should be taken in content moderation and complaint procedures that this content remains.

Platforms should ensure that the content that is legal and does not violate platform norms of otherwise criminalized groups, "grey zone content" is diligently protected and does not get caught up in content moderation actions.

ACKNOWLEDGEMENTS & INTERVIEWEES

INTERVIEWEES

Anastasia Karagianni	Co-founder, DATAWO
Ela Stapley	Founder, Director of Siskin Labs part of the Coalition Against Online Violence founded by the IWMF
Gema Fernández	Managing Attorney, Women's Link Worldwide
Josephine Ballon	CEO, HateAid gGmbH
Kinga Jelinska	Women Help Women, Executive Director/ co-founder; Abortion Dream Team, member/ co-founder
M. Schipper	Activist, Dikke Vinger
Nadya Yurinova	Senior Communications Officer at TGEU
Osama Al-Sayyad	Co-managing editor, Arabi Facts Hub
Sabrina Sanchez	Director, ESWA
Yigit Aydinalp	Digital Rights Programme Officer, ESWA

ACKNOWLEDGEMENTS

I am deeply grateful for the people and their organisations that donated their time and expertise to participate in the interviews for this project. I also want to thank all participants to the September 2023 Disparate Content Moderation Workshop.

Specifically, I want to thank Joris van Hoboken for his feedback and support, Angelina Wagner for her valuable help, as well as the wider support and feedback from the DSA Observatory.

This research project was funded by the DSA Observatory.

SUGGESTED CITATION

Naomi Appelman, "*Disparate Content Moderation: mapping social justice organisations Perspectives on unequal content moderation Harms and the EU platform policy debate.*" (2023) Institute for Information Law, University of Amsterdam, available at: https://dsa-observatory.eu/2023/10/31/research-report-on-disparate-content-moderation/.

REFERENCES

Alexander, J., Savigny, H., Thorsen, E., & Jackson, D. (2015). Introduction: Marginalised Voices, Representations and Practices. In E. Thorsen, H. Savigny, J. Alexander, & D. Jackson (Eds.), *Media, Margins and Popular Culture* (pp. 1–12). Palgrave Macmillan UK. https://doi.org/10.1057/9781137512819_1

Allen, A. (2022). An Intersectional Lens on Online Gender Based Violence and the Digital Services Act. *Verfassungsblog: On Matters Constitutional*. https://doi.org/10.17176/20221101-215626-0

Appelman, N., van Duin, J. M. L., Fahy, R., van Hoboken, J., Helberger, N., & Zarouali, B. (2021). Access to Digital Justice: In Search of an Effective Remedy for Removing Unlawful Online Content. In X. Kramer, J. Hoevenaars, B. Kas, & E. Themeli (Eds.), *Frontiers in Civil Justice: Privatisation, Monetisation and Digitisation*. Elgar. https://papers.ssrn.com/abstract=3961390

Are, C. (2022). An autoethnography of automated powerlessness: Lacking platform affordances in Instagram and TikTok account deletions. *Media, Culture & Society*, 016344372211405. https://doi.org/10.1177/01634437221140531

Are, C., & Briggs, P. (2023). The emotional and financial impact of de-platforming on creators at the margins. *Social Media and Society*.

Are, C., Collingham, H., Carrothers, A. M., & Fox, E. (2023). *Co-designing platform governance policies. Tackling malicious flagging and de-platforming with impacted social media users.* https://blogger-onpole.com/2023/07/new-report-shares-user-centred-social-media-policy-recommendations/

Article 19. (2018). *Side-stepping rights: Regulating speech by contract (Policy Brief)* (p. 56). Article 19. https://www.article19.org/resources/side-stepping-rights-regulating-speech-by-contract/

Balayn, A., & Gürses, S. (2021). *Beyond debiasing: Regulating AI and its inequalities*. EDRi.

Banko, M., MacKeen, B., & Ray, L. (2020). A Unified Taxonomy of Harmful Content. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 125–137. https://doi.org/10.18653/v1/2020.alw-1.16

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity Press.

Berg, B. L., & Lune, H. (2012). *Qualitative Research Methods for the Social Sciences*. Pearson.

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), *Social Informatics* (pp. 405–415). Springer International Publishing. https://doi.org/10.1007/978-3-319-67256-4_32

Bishop, S. (2019). Managing visibility on YouTube through algorithmic gossip. *New Media & Society*, *21*(11–12), 2589–2606. https://doi.org/10.1177/1461444819854731

Brown, W. (1995). *States of Injury*. https://press.princeton.edu/books/paperback/9780691029894/states-of-injury

Caplan, R. (2018). *Content or context moderation?* Data & Society.

Caplan, R., & Gillespie, T. (2020). Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy. *Social Media + Society*, *6*(2), 205630512093663. https://doi.org/10.1177/2056305120936636

Celeste, E. (2019). Terms of service and bills of rights: New mechanisms of constitutionalisation in the social media environment? *International Review of Law, Computers & Technology*, *33*(2), 122–138. https://doi.org/10.1080/13600869.2018.1475898

Clark-Parsons, R., & Lingel, J. (2020). Margins as Methods, Margins as Ethics: A Feminist Framework for Studying Online Alterity. *Social Media + Society*, *6*(1), 205630512091399. https://doi.org/10.1177/2056305120913994

Cobbe, J. (2021). Algorithmic Censorship by Social Platforms: Power and Resistance. *Philosophy & Technology*, *34*(4), 739–766. https://doi.org/10.1007/s13347-020-00429-0

Cotter, K. (2023). "Shadowbanning is not a thing": Black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*, *26*(6), 1226–1243. https://doi.org/10.1080/1369118X.2021.1994624

Dias Oliva, T., Antonialli, D. M., & Gomes, A. (2021). Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture*, *25*(2), 700–732. https://doi.org/10.1007/s12119-020-09790-w

Díaz, Á., & Hecht-Felella, L. (2021). *Double Standards in Social Media Content Moderation*. Brennan Center for Jus-

tice. https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation

Dinar, C. (n.d.). *The state of content moderation for the LGBTIQA+ community and the role of the EU Digital Services Act*.

Duffy, B. E., & Meisner, C. (2022). Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility. *Media, Culture & Society*, 016344372211119. https://doi.org/10.1177/01634437221111923

Dvoskin, B. (2021). Representation without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3986181

EDRi, & DFF. (2023). *Decolonising the Digital Rights Field in Europe*. https://digitalfreedomfund.org/decolonising/

EFF. (2019). *EFF Project Shows How People are unfairly "Tossed Out" by Platforms' Absurd Enforcement of Content Rules.* Electronic Frontier Foundation.

Elkin-Koren, N., De Gregorio, G., & Perel, M. (2022). Social Media as Contractual Networks: A Bottom up Check on Content Moderation. *Iowa Law Review*, *107*. https://doi.org/10.2139/ssrn.3797554

ESWA. (2022). *The Impact of Online Censorship and Digital Discrimination on Sex Workers*. ESWA.

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods*, *5*(1), 80–92. https://doi.org/10.1177/160940690600500107

Ganesh, M. I., & Moss, E. (2022). Resistance and refusal to algorithmic harms: Varieties of 'knowledge projects'. *Media International Australia*, *183*(1), 90–106. https://doi.org/10.1177/1329878X221076288

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, *20*(12), 4492–4511. https://doi.org/10.1177/1461444818776611

Gerrard, Y., & Thornham, H. (2020). Content moderation: Social media's sexist assemblages. *New Media & Society*, *22*(7), 1266–1286. https://doi.org/10.1177/1461444820912540

Gilbert, S. A. (2023). *Towards Intersectional Moderation: An Alternative Model of Moderation Built on Care and Power* (arXiv:2305.11250). arXiv. http://arxiv.org/abs/2305.11250

Gillespie, T. (2015). Platforms Intervene. *Social Media + Society*, *1*(1), 205630511558047. https://doi.org/10.1177/2056305115580479

Gillespie, T. (2018). Governance of and by platforms. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The SAGE handbook of social media* (p. 30). Sage.

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, *8*(3), 20563051221117552. https://doi.org/10.1177/20563051221117552

Gillett, R., Gray, J. E., & Valdovinos Kaye, D. B. (2023). 'Just a little hack': Investigating cultures of content moderation circumvention by Facebook users. *New Media & Society*, 146144482211476. https://doi.org/10.1177/14614448221147661

Glitch, UK. (2023). *The Digital Misogynoir Report: Ending the dehumanising of Black women on social media*.

Goantă, C., & Ranchordás, S. (Eds.). (2020). *The regulation of social media influencers*. Edward Elgar Publishing.

Goldman, E. (2021). Content Moderation Remedies. *Michigan Technology Law Review*, *28.1*, 1. https://doi.org/10.36645/mtlr.28.1.content

Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, *22*(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 205395171989794. https://doi.org/10.1177/2053951719897945

Griffin, R. (2022). Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4064738

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–35. https://doi.org/10.1145/3479610

Hall, J. M. (1999). Marginalization Revisited: Critical, Postmodern, and Liberation Perspectives. *Advances in Nursing Science*, *22*(2), 88.

Hoboken, van, J., Appelman, N., van Duin, A., Blom, T., Zarouali, B., Fathaigh, R. Ó., Seel, M., Stringhi, E., & Helberger, N. (2020). *WODC-onderzoek: Voorziening voor verzoeken tot snelle verwijdering van onrechtmatige online content*. Institute for Information Law.

Jane, E. A. (2016). Online misogyny and feminist digilantism. *Continuum*, *30*(3), 284–297. https://doi.org/10.1080/10304312.2016.1166560

Karizat, N., Delmonaco, D., Eslami, M., & Andalibi, N. (2021). Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–44. https://doi.org/10.1145/3476046

Kiss, E. (1995). Alchemy or Fool's Gold? Assessing Feminist Doubts About Rights. *Dissent Magazine*. https://www.dissentmagazine.org/article/alchemy-or-fools-gold

Llansó, E. J. (2020). No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society*, *7*(1), 2053951720920686. https://doi.org/10.1177/2053951720920686

Marshall, B. (2021). *Algorithmic misogynoir in content moderation practice* (p. 17). Heinrich-Böll Stiftung.

Megarry, J. (2014). Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. *Women's Studies International Forum*, *47*, 46–55. https://doi.org/10.1016/j.wsif.2014.07.012

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Pearce, K. E., Gonzales, A., & Foucault Welles, B. (2020). Introduction: Marginality and Social Media. *Social Media + Society*, *6*(3), 205630512093041. https://doi.org/10.1177/2056305120930413

Posetti, J., Shabbir, N., Maynard, D., Bontcheva, K., & Aboulez, N. (2021). *The Chilling: Global trends in online violence against women journalists; research discussion paper* (CI-2021/FEJ/PI/1). Unesco. https://unesdoc.unesco.org/ark:/48223/pf0000377223

Roberts, S. T. (2018). Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday*, *23*(3). https://doi.org/10.5210/fm.v23i3.8283

Scheuerman, M. K., Branham, S. M., & Hamidi, F. (2018). Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–27. https://doi.org/10.1145/3274424

Scheuerman, M. K., Jiang, J. A., Fiesler, C., & Brubaker, J. R. (2021). A Framework of Severity for Harmful Content Online. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–33. https://doi.org/10.1145/3479512

Siapera, E., & Viejo-Otero, P. (2021). Governing Hate: Facebook and Digital Racism. *Television & New Media*, *22*(2), 112–130. https://doi.org/10.1177/1527476420982232

Sitaraman, G. (n.d.). Deplatforming. *Yale Law Journal*, *forthcoming*.

Smith, S. L., Haimson, O. L., Fitzsimmons, C., & et al. (2021). *Exclusive: Censorship of Marginalized Communities on Instagram*. Salty. https://saltyworld.net/exclusive-report-censorship-of-marginalized-communities-on-instagram-2021-pdf-download/

Táíwò, O. O. (2022). *Elite capture: How the powerful took over identity politics (and everything else)*. Haymarket Books.

Thach, H., Mayworm, S., Delmonaco, D., & Haimson, O. (2022). (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society*, 146144482211098. https://doi.org/10.1177/14614448221109804

Tiidenberg, K. (2021). Sex, power and platform governance. *Porn Studies*, *8*(4), 381–393. https://doi.org/10.1080/23268743.2021.1974312

Tufekci, Z. (2015). Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency Symposium Essays. *Colorado Technology Law Journal*, *13*(2), 203–218.

Vaccaro, K., Sandvig, C., & Karahalios, K. (2020). 'At the End of the Day Facebook Does What ItWants': How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–22. https://doi.org/10.1145/3415238

Vitak, J., Chadha, K., Steiner, L., & Ashktorab, Z. (2017). Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1231–1245. https://doi.org/10.1145/2998181.2998337

Williams, P. (1991). *The Alchemy of Race and Rights: Diary of a Law Professor*. Harward University Press.

Zeng, J., & Kaye, D. B. V. (2022). From content moderation to *visibility moderation*: A case study of platform governance on TikTok. *Policy & Internet*, *14*(1), 79–95. https://doi.org/10.1002/poi3.287