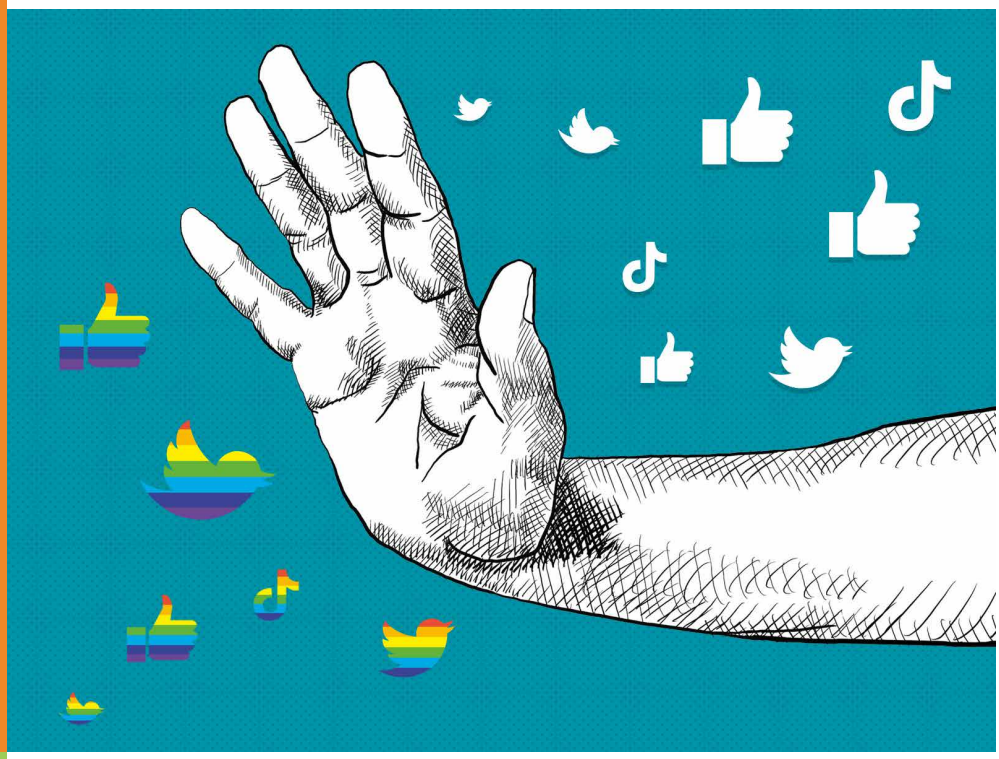
HEINRICH BÖLL STIFTUNG
DEMOCRACY

E-PAPER

The state of content moderation for the LGBTIQA+ community and the role of the EU Digital Services Act



CHRISTINA DINAR

Published by Heinrich-Böll-Stiftung European Union
and Heinrich-Böll-Stiftung Washington, DC, June 2021

The author

Christina Dinar is a freelance lecturer and trainer on anti-discrimination strategies online, anti-racist conflict resolution, digital competences, democratic debating culture and democracy promotion amongst young people.

As Deputy Director at the Centre for Internet and Human Rights Dinar led the research project Democratic Meme Factory, studying the impact on counterspeech on youth. As a project manager at Wikimedia Germany, she worked to enhance diversity in the community of Wikipedia content editors. She also co-developed the concept of digital streetwork, which is a social practice targeting youth online communities, mostly through social media.

Dinar is interested in the effect of digitalization to social norms and practice – especially to youth culture. Christina Dinar studied social work as well as theology, cultural studies and gender studies in Berlin/Jerusalem and lives in Berlin.

Abstract

Social media platforms play a crucial role in supporting freedom of expression in today's digital societies. Platforms can empower groups that have previously been silenced. However, platforms also host hateful and illegal content, often targeted at minorities, and content is prone to being unfairly censored by algorithmically biased moderation systems. This report analyzes the current environment of content moderation, particularly bringing to light negative effects for the LGBTIQA+ community, and provides policy recommendations for the forthcoming negotiations on the EU Digital Services Act.

Contents

List of abbreviations

AI	Artificial Intelligence
CSAM	child sexual abuse material
DMA	EU Digital Markets Act
DMCA	Digital Millennium Copyright Act
DSA	EU Digital Services Act
GDPR	General Data Protection Regulation
GIFCT	Global Internet Forum to Counter Terrorism
LBGTIQA+	Lesbian, bisexual, gay, trans, intersex, questioning/queer, asexual
NetzDG	German Network Enforcement Act
ODIHR	Office for Democratic Institutions and Human Rights
OSCE	Organization for Security and Co-operation in Europe
sCAN	Specialized Cyber-Activists Network
ToS	Terms of service
ugc	user-generated content
VLOP	Very Large Online Platform

1. Background and objectives of this report

Open, public, and rational discourse is often considered the heart of democracy, and social media platforms have provided key infrastructure for exactly this purpose. Social media, freed from the constraints of traditional media gatekeepers, facilitates content created and shared by users themselves. It has enabled social movements and political change[1], especially for those suffering from structural exclusion and suppression. With the help of social media, movements like #MeToo[2], #BlackLivesMatter[3], and #FridaysForFuture[4] created unprecedented and powerful global impacts.

However, not all user-generated content amplifies movements for equity and gender democracy. Social media is a space "of feminist activism [but] at the same time [a space] of surveillance and punishment for feminist activism and activity"[5]. Social media is also a place for populist right-wing content promoting authoritarian and minority-threatening ideologies. This type of content is often perceived as "free speech," even when it harms democracy and destabilizes democratic institutions, as in the case of the storming of the United States Capitol in January 2021. Now able to bypass journalists and standards of reporting, right-wing counterculture flourishes in an "alternative platform"[6] and can quickly spread misinformation and violent ideas.

In this environment, social media companies frequently cast themselves merely as hosting services that enable free speech—yet they are not neutral platforms. These websites moderate or curate the content that users see, often in poor, discriminatory, or opaque ways that rely on simplistic technical solutions.

Thoughtful content moderation is crucial for building a healthy, safe, and inclusive internet, especially as the number of social media users grows. Today, 46% of the population in eastern Europe uses these platforms, and that number rises to 67% in northern Europe.[7] Yet content moderation has not always received the attention it deserves. The European Commission's draft legislation[8] on regulating high-risk AI, released in April, does not include content moderation, even though many content moderation systems rely on artificial intelligence.

This report will show how current content moderation policies are filled with inconsistencies and double standards that often hurt marginalized communities. The report will then address models of content moderation and current EU regulatory approaches, focusing especially on the aims and obstacles of the proposed Digital Services Act. It will provide recommendations on how to build content moderation practices that protect marginalized communities. Ultimately, the report argues that a stronger democratization of content moderation is necessary to build infrastructure that helps gender democracy become a key part of internet culture.

1.1. Power structures in content moderation

Moderation practices address how platform providers deal with user-generated content. Platforms organize, curate, amplify, shadowban (make a user's contributions invisible to others without the user's knowledge) and can eventually take down content when it is illegal or against their community guidelines.

Many of the large social media platforms were created and grew big in the United States. As a result, they – and their content moderation practices – are influenced by permissive attitudes toward free speech, which date back to a 1927 Supreme Court ruling[9] in which Justice Louis Brandeis wrote that opinions are part of the "marketplace of ideas" in which ideas had to compete for acceptance and popularity.

These market-driven values still exist in the logic of these platforms' infrastructure. Likes, shares, and algorithmic amplifications of social media content operate as part of the "marketplace of ideas," where the ideas compete and the most-"liked" ideas are most visible and "win." This belief in the marketplace of ideas and self-regulating speech leads to a reliance on quick technological solutions.

Human content moderators, tasked with upholding the community guidelines, reproduce stereotypes around gender and race. Algorithms then learn norms from the humans categorizing the content, thus biasing the systems. Ultimately, the "marketplace of ideas," like other digital public spaces, is inherited from long-established power structures that are inconsistent, full of double standards, and that favor the white, male perspective – often at the expense and safety of marginalized communities.

For example, Twitter did not ban the tweets of former US President Donald Trump, but it did suspend an account retweeting the same content[10] without holding the @POTUS[11] account. In Europe, the EU Commission-funded project sCAN[12], a network to analyze and identify reasons for hate speech, has also identified European politicians promoting hate online and benefitting from a different standard of moderation.[13] The terms of service do not seem to apply to the social media accounts of the powerful and influential.

This is not the only way that power structures are deeply embedded in technology. Instagram has long been criticized for censoring female nipples[14] but not male nipples, and therefore promoting puritanical cultural values around bodies.[15] In a study on content moderation and sexism, Gerrad and Thornham show that the Instagram algorithm typically shows similar content, rather than anything alternative or contrary.[16] The platforms are setting parameters of acceptable social conduct, which affects society's most marginalized groups.[17]

Another example of how content moderation can increase marginalization comes from TikTok, which is becoming popular among young people in Europe. According to a report[18] from the German site Netzpolitik.org, the platform's moderation algorithm classifies LGBTIQA+ content as "risky"; this content is then geotagged and suppressed in so-called "Islamic countries." The Netzpolitik.org investigation also showed that content from plus-sized users and users with disabilities is suppressed in order to "protect" these vulnerable and marginalized groups – a move seen as paternalistic and exclusionist by those affected, who believe they do not have enough representation as it is. There is also an element of victim-blaming at play: Even if these users are more likely to be targeted by mobs and trolls, TikTok should focus on sanctioning the perpetrators, not punishing the creators.

1.2. The cases of Salty and PlanetRomeo

Salty[19] is an independent, donation-based and membership-supported newsletter for women, trans, and nonbinary people. The newsletter is regularly sent out to members of the community and is connected to a website that functions as a publishing platform. Salty is very present on Instagram[20] and Twitter[21] and has over 50,000 newsletter subscribers and about 3 million monthly impressions across platforms.

When the Salty Instagram account tried to publish ads featuring transgender and non-binary people of color, Instagram rejected these ads and marked them as "escort service."[22] After not getting through to Instagram with a complaint, Salty used their Twitter account to call out this development and subsequently received an invitation to meet with Facebook and help the company be more inclusive in their content moderation practices. Facebook never followed up on the invitation and the meeting did not happen, but the content of Salty members is still frequently taken down.

Salty gathered these posts into a report[23] on algorithmic bias in content policing on Instagram and Facebook. The report collects the experiences of 118 people, many of whom identify as LGBTIQA+, people of color, plus-sized, or sex workers. Many had experienced content being taken down or being "accidentally" getting booted from platforms.

One of the problems, according to Salty's analysis, is vague communication on the part of platforms regarding which community guidelines a post violated. Nudity and body positivity posts, and posts by BIPOC, are often taken down or shadowbanned or marked as pornography, seemingly without explanation. Many people did not understand exactly why, but felt that it was connected to their skin color or the fact that their bodies did not fit a heteronormative standard.

Furthermore, women-led businesses also had issues with their products being banned. "It is also infuriating, because we see endless ads on the same platforms for erectile dysfunction medication, penis pumps, and 'manscaping' razors. Why are penises normal but the female and non-binary body considered a threat?"[24] asked a person who was not able to advertise breast pumps on Instagram.

A similar story comes from PlanetRomeo, a social media network and dating site for gay, bisexual, and transgender people. Amsterdam-based PlanetRomeo operates in three different languages and has 1.8 million users online. Besides being a dating platform, the app provides information on sexual health[25] and has also formed the PlanetRomeo Foundation[26], which supports LGBT+ rights all over the world.

PlanetRomeo's app has been removed from the Google Play store 11 times since 2013, without notice, for being "sexually sensitive." Because other dating apps, which mainly target a straight audience, are not categorized as "sexually sensitive," PlanetRomeo believes that content moderation tools automatically see queer content as more inherently sexual.

Apps like PlanetRomeo are important to smaller and marginalized communities. Yet dating apps that mostly target straight audiences, and which have higher download numbers, will probably have privileged access in the app store. Interestingly, experiences of exclusion are not new for PlanetRomeo. After the platform expanded internationally in 2012, the name changed from GayRomeo to PlanetRomeo[27], as domains that include the word "gay" are blocked in some countries. All this shows that inconsistencies applied through content moderation are still a reflection of current power and privilege.

2. Different approaches to content moderation

As already noted, risks of discrimination are high in content moderation because both false positives (e.g. abusive messages that pass the search) and false negatives (content that should stay up) are common. Today, major platforms are still at the beginning of their capacity to develop different modes of content moderation.[28] They work with different measurement tools, including downranking, reducing visibility, adding labels alerts or supplementary information, or nudging (active warning before the publication of content). As pressure on companies increases, companies invest more in these tools. Facebook, for example, is experimenting with up and down-voting of postings[29] as well as features that allow a choice between algorithmically-ranked or chronologically sorted feeds.[30]

Table 1 – Approaches to content moderation (following the model of Pershan 2020[31] based on Kaplan 2018[32])

Approach	Moderation	Tools	User involvement	Platforms/ Impact
Industrial Moderation or Decision-factory	Up to 10,0000 employees around the world Many moderators are third parties or contractors Moderation teams are separated from design and policy teams Moderation system formal, systematic, policy guided (factory-like) Moderators labour time kept under surveillance/tracked	Majority of content is filtered by automatic tools Most participate in hash sharing collectives (PhotoDNA, GIFCT)	Trusted Flagger programs with div. civil society actors (e.g. fact-checking)	Facebook, Twitter, YouTube, TikTok impact Helps to process enormous amount of data Exploits content moderation workers Often opaque policies Rules are consistent but lack context and local scale Very Distant relationships between user and platform Little time to make decisions

Approach	Moderation	Tools	User involvement	Platforms/ Impact
Artisanal	Moderation teams range between about 5 and 200 Moderators exchange/coordinate with other teams, mostly employed by the platforms themselves Content moderation decisions are weighed and function on a case-by-case basis (“manually”)	Limited use of artificial intelligence, most content is examined ex post (not filtered) Platforms may participate in hash-sharing collectives but in a passive/non-strategic role	More time is taken per post and users are considered more holistically/within the history of their online activity The process for flagging content is similar to that of industrial platforms	Patreon, Vimeo, change.org Easier to adapt cultural context and balance decision More time for decision-making Higher costs for the platforms Potential to work with volunteer moderators from users’ sight Employed by company and can deal within their own structure
Community-reliant	Multi-layered model with a core team of a few salaried staff and then degrees of volunteer participation and responsibility (“onion layers” around the piece of content) Some transversal policies, but pages/ group establish their own rules and respective moderators are responsible for enforcing these rules (“federal”) Volunteer moderators are not remunerated	Less use of AI, though there are automated tools available to users and moderators to use as they see fit	Any user may become a moderator Moderation responsibility can be increased over time User flagging varies; often, users can bring complains directly to moderators, in some cases	Wikipedia, Reddit, Mastodon Decisions are made with community, more transparency Power structure often establishes between the user groups and volunteers, communication Can provide inside to cultural sensitivity

The three forms of content moderation laid out in **Table 1** each have specific benefits and drawbacks. The industrial type creates internal moderation consistency but leads to false positives. There is little communication between the users and the content moderation teams regarding takedowns, and account blocking is often not communicated clearly. Plus, labor exploitation is an enormous issue with this form in particular. Humans in the loop must be extremely sensitive about the cultural context, as researcher Robyn Kaplan mentions in her interviews with content moderators[33], and are often tasked with viewing traumatic and psychologically damaging material. Yet the labor in content moderation is not paid well.

The artisanal approach is usually the type of moderation a platform begins with. Moderators often learn with a handbook and develop moderation practices from there. There are also artisanal forms that work with specific users that support the team by flagging content and giving notice. The artisanal approach leaves more space for looking at moderation cases that are on the fringes of community standards, but can make decisions inconsistent.

The community-reliant model is often cited as the most democratic approach. Users work using their own sets of rules, as on Wikipedia. (Wikipedia is a purely community-led platform that only works with volunteers participating in the project.) Community members create and moderate all their content alone, without the help of paid professionals, and the approach is seen as a chance to connect users with their content rights.

Wikimedia's Dimi Dimitrov says[34] that being open-source and allowing anyone to look into the requests for takedowns[35] helps keep the requests and notices down. Self-governance of the community allows members to discuss publicly whether content should get taken down, so employees don't have to do it. Over the period from 2012-2018, Wikipedia has had 2,942 takedown requests, but only one has been granted, according to Wikimedia – demonstrating that open policies can create a lower workload.

Wikipedia's articles are high-quality and up-to-date. However, these community-moderated spaces also lack equality, inclusiveness, and an internal participatory culture, as the gender gap in Wikipedia shows.[36]

All platforms have to deal with inconsistent forms of moderation, policy changes, problems of context, and discussion around the users' right to freedom of speech. All platforms prioritize the balance between context-sensitivity and consistency differently. Often, the decision depends on their resources and organizational dynamics. Most companies keep their content moderation policies hidden – making them appear to be the supposedly "objective rulers" of their content governance.[37]

2.1. PlanetRomeo's community model for moderating content

PlanetRomeo[38], which was covered earlier in this paper, relies on community-based moderation to help flag suspicious accounts. PlanetRomeo community members rate pictures before they are uploaded to dating profiles. This has proven to be very efficient in combination with automated content moderation tools. (The staff have tried to work with A.I.-related tools to filter content but the community-reliant approach of oversight and rating seems to be more efficient for them.) Their moderation team mostly comes from similar communities, which helps with understanding the cases.

PlanetRomeo also has a team of paid employees that frequently deals with cases of discrimination cases (for example, between two community members), such as outings, blackmailing with unauthorized pictures, digital violence like lovescams[39] and sometimes doxing.[40]

PlanetRomeo also works with social councils of the LGBTIQA+ community, such as the Gay Consultation Berlin. PlanetRomeo helps people that come to Gay Consultation Berlin and have been affected by unauthorized outings on PlanetRomeo profiles. Unauthorized outings are especially relevant for queer refugees, who were blackmailed in their home countries (where homosexuality is considered a crime) and then threatened in refugee shelters in Germany. This partnership demonstrates that civil society actors and counselling structures can work well together.

2.2. Technical solutions and their challenges

Automated content moderation relies on machine learning techniques. Many platforms work with a range of technical measures to help content moderation – for example, using artificial intelligence to detect problematic content. Facebook claims to detect nearly 70%[41] of hate speech posts prior to human reporting, but releases very little information[42] about how the platform is doing, so we have few insights into its automated learning system.

Big Tech companies have set up the GIFCT[43] (Global Internet Forum to Counter Terrorism) database with hashes[44] (which function like digital fingerprints) that help remove content classified as CSAM or terrorist content, threats, and cyber harassment. This prevents the re-uploading of content material that has been hashed. Others use a Microsoft software called PhotoDNA.[45]

There is an obvious need for transparency regarding clearing automated illegal material, and the GIFCT has been criticized for a lack of transparency and democratic oversight.[46] Publicly releasing an anonymized database of legal complaints and requests for removal of online materials can be a first step. The Lumendatabase[47] serves here as a good example.

There are other technical challenges when it comes to designing fairness into algorithms. For example, "most current approaches for algorithmic fairness assume that the target characteristics for fairness – frequently, race and legal gender – can be observed or recorded. Sexual orientation and gender identity are prototypical instances of unobserved characteristics, which are frequently missing, unknown or fundamentally unmeasurable."[48] In short, queerness and gender fluidity are not easily mapped by code, as they have an ever-changing nature that is not always compatible with categories and technological systems.

3. The regulatory framework in the EU on content moderation

3.1. From the e-commerce directive to the digital services act

The European Commission wants to assign greater responsibility to today's powerful platform providers with the EU Digital Services Act[49] (DSA). The goal of the DSA is "to enable all EU-citizens to exercise the rights guaranteed to them by the Charter of Fundamental Rights, in particular the right to freedom of expression and information, the freedom to conduct a business and the freedom from discrimination."[50] DSA also aims to regulate Big Tech companies, minimize exploitative practices on users, and address the growing violence, hate speech, and misinformation online.

The proposed DSA touches upon the regulation of content moderation of so-called "Very Large Online Platforms" (VLOPs), meaning those that have more than 45 million users in the EU. Without naming them, the DSA addresses the largest and most dominant companies in the information technology industry, including Amazon, Alibaba, Facebook, and Google.

It contains obligations to disclose how algorithms work, transparency regarding decisions for removed content, and how advertisers target users as well as liabilities for marketplaces. If platforms do not abide by the law, they risk fines of up to 6% of their annual global revenue. The law is designed to regulate the economic and legal role that big platforms have in European society. Exemptions are provided for small businesses, which should also apply to platforms that are not businesses, e.g. Wikipedia.

The legislation will define new rules for the future of the internet and will impact internet governance not only in Europe, but globally. The DSA will set standards alongside and similar to the General Data Protection Regulation[51] (GDPR), which, after being passed in 2016, led to a rise in the standard of data protection for other countries around the world. Similarly, the DSA will also apply to platforms based outside the EU, at least if they target their services at users within the EU.

3.2. A crucial distinction: Harmful vs. illegal content

Mechanisms of exclusion and discrimination can work very subtly. One of the crucial distinctions that the DSA will need to grapple with is the difference between content that is clearly illegal (e.g. rape or death threats, as well as CSAM or extremist messages) and content that is "just" harmful but can still make it difficult for marginalized groups to participate and communicate freely without being harassed or their content being censored.

One way to differentiate between harmful and illegal content is to take a closer look at hate speech. In 1997, the Council of Europe's Committee of Ministers described the scope of hate speech through the media as follows:

"The term 'hate speech' shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin."[52]

Hate speech can be illegal and lead to violence (e.g. incitement). However, hate speech can also be legal – and legal hate speech is called "dangerous speech."[53] Dangerous speech refers to a form of communication that is technically protected by freedom of expression but still lowers the barriers for users to start practicing illegal hate speech. The concept of dangerous speech has its roots in genocide research. According to researcher Susan Benesch, the media atmosphere can play a key role in normalizing hostile social behaviors toward minorities.

Dangerous speech has become digital, creating a slippery slope that can lead to violence and "real world" consequences. Dangerous speech is relevant to content moderation because it sits on the fence between illegal and illegal speech and is often not easy to capture and categorize. Take the example of a comment like "I am bringing the fire accelerant!" on a website for a refugee shelter. The context makes it clear that the comment is a threat. However, it is not illegal to make such a comment.

Ultimately, the term "illegal content" is very limiting. Differentiating between legal and illegal takedown procedures (DSA (2) Art. 14 (2)) does not appear sufficient to protect marginalized communities who are targets of not only illegal hate crime but also of legal hate speech and dangerous speech and are harmed through that.

Trying to solve these problems of digital violence and discrimination through hate crime law and adjusting what is criminal won't help regulate the harmful and dangerous speech that is visible every day on social media. Users that commit hate crimes and create illegal posts need to be taken into account, but further measurements like the proactive promotion of diversity, tolerance, and counterspeech need to be equally enforced.

3.3. Lessons from NetzDG

In January 2018, Germany became one of the first EU Member States to pass a law regulating speech on social media. NetzDG created a catalogue of offences[54] that harm individual and collective rights. It applied high penalties to social media platforms with more than two million registered German users that did not take down content relevant to incitement or insult within a 24-hour to seven days' notice.

But civil society and economic actors[55] worried that the proposed law – The Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act, NetzDG[56]) – did not meet the requirements to adequately protect freedom of expression. Just a few months later, the Russian Duma passed a similar law. Reporters Without Borders, an NGO, stated that the NetzDG law had served as a model for non-democratic states such as Russia to limit freedom of expression using social media.[57]

Then, UN rapporteur for freedom of expression David Kaye released a statement saying that NetzDG creates too much of a burden for companies and that the lack of a judicial oversight is not compatible with international human rights law. He also pointed out that, in reaction to such laws, platforms are likely to over-regulate content and will therefore overblock controversial content.[58]

The NetzDG entered into law two years ago. It provides useful lessons that can be applied to the DSA:

- Be wary of putting all responsibility in the hands of the platforms' content moderation and legal teams. This is what NetzDG did and this led to the overblocking of content, according to a recent study from Liesching et al 2021.[59, 60]
- Avoid creating double standards in reporting structures. NetzDG led to the establishment of one reporting structure according to state regulation and one structure according to community standards (linked with the platform's ToS). But the reporting mechanism of NetzDG was so difficult to find in the interface that many users still preferred to report according to community standards.
- Regular transparency reporting needs clear standards with comparable numbers that are openly accessible.
- Penalty fines cannot be the only way to punish big platforms into taking responsibility. Currently, only Facebook has received a fine of 2 million Euros[61] (in 2018) and has still not paid. Instead, the company is investing in court appeal procedures.[62] This demonstrates the unwillingness to take responsibility. Overall, penalties are not high enough to change behavior or responsibilities.
- Reward social media actors that are transparent and come up with good examples of solving content moderation and enhance their impact by supporting them.

4. Policy recommendations for the EU Digital Services Act

To fight discrimination, solutions need to be embedded in structures that help communities participate on platforms without being harassed or censored. The revised EU regulatory framework for online content moderation, which will be part of the forthcoming EU Digital Services Act, should actively involve civil society, educational programs, and measures to a create a sustainable ecosystem that strengthens diversity:

4.1. Involve users and civil society

- Formalize connections between existing support structures (e.g. women's centers, LGBTIQA+ support groups) and the Digital Services Coordinator (DSA, Art. 2 (l)) to help the groups most vulnerable to experiencing digital violence. Grant these organizations verified accounts and trusted flagger rights to help appeal against unjustified content takedowns, shadow banning, and account blocking. This concerns doxing as well as identity theft or releasing unauthorized information about a user's sexual orientation.
- Decentralize decision-making through Social Media Councils. Tie decision-making of controversial cases to a decentralized and independent Social Media Council (members are users, including experts and affected communities) according to the rules of fair discussion open for a public debate. Social Media Councils should function as an independent and external body. To tackle harmful and dangerous speech the Social Media Council could also work to estimate and discuss the borders of fringe content (between legal and illegal) as well as normative content (such as nudity). It can publicly issue recommendations for the platforms' moderation policies.
- Require platforms to form and support independent community management that mediates between the content moderation team and the users that feel their content/account has been wrongfully blocked. The Digital Services Coordinator (DSA, Art. 2 (l)) is the national authority within the EU for the DSA framework to coordinate. She can invite and direct discussions beyond the legal framework and that go beyond the DSA-proposed out-of-court dispute settlement on content moderation (DSA, Art.18 (2)).

4.2. Educate and prevent through EU programs

- Inform users/citizens about their rights to report content as part of media literacy programs. Evaluate how much users and citizens are aware of their rights when it comes to reporting of content or where they can seek help in case of discrimination. Education programs should target the everyday social media user in the EU with offerings on the topics of content moderation, online discussion culture, legal and illegal speech, counterspeech.
- Support counterspeech initiatives through citizen-led and educational EU programs that empower civil society, similar to the German initiative "Competence Centre: Hate on the Net"[63] for young people.
- Support research and independent report evaluation that creates a deeper understanding of the way social norms and sanctions are distributed in online communities. Understanding how alternative content moderation is conducted helps prevent the uploading of illegal content and prevents discrimination.
- Ensure psychological support for content moderators, make their working conditions transparent, and assure a certain level of training and education.

4.3. Introduce measures for a sustainable digital ecosystem with diverse platforms

- Promote and reward small actors (e.g. platforms with fewer than 45 million users in the EU) that are performing well in the field of content moderation, instead of creating penalties to lower the power of economically driven platforms that have become too powerful.
- Be aware of the growth in social media users and their development. The EU is expected to have 30 million more internet users by 2025[64], markets might shift under the DSA, and smaller niche platforms might grow big and will have to fall under the law. Those platforms might suppress their growth in order to avoid high penalties.
- Introduce a ratio that guides investment of labor resources in content moderation

$$\frac{\textit{Plattform content uploaded per hour}}{\textit{Hours of work required by moderator}} = \textit{Required labor resources}$$

  This is important especially for the industrial content moderation systems, which tend to minimize cost and maximize profit. The "required labor resources" can be frequently revised and fitted through discussions with digital coordinators on a national level. The guiding principle should be a sustainable digital (informational and social) ecosystem that is guided by principles of human rights and social cohesion.

- Encourage smaller platforms to create alternative models of content moderation that engage stronger community engagement.
- Develop innovative measures in the field of moderation, as well in community engagement. Thus smaller platforms can compete and offer alternatives in the market of social media platforms.
- Make trusted flaggers transparent. The criteria for becoming a trusted flagger should be clear. Trusted flagger status should be issued only for a limited amount of time and should be regularly revised.
- Require platforms to make reporting procedures available easily (3-Click-Rule[65]), in a user-centered design and in plain language. Prohibit dark patterns in reporting structures.
- The actual content moderation labor needs to be in the EU and supervised by people who speak the language.
- Document illegal content transparently, e.g. similar to the GIFCT and hash database (Photo DNA) via an independent agency for research and supervision (e.g. Lumendatabase[66]).
- Create roundtables and make the platforms discuss specific cases that are relevant between platforms (cross-platform abuses, specific events).

4.4. Foster content moderation practices that are more inclusive and accountable

- Require independent oversight from non-governmental structures for content moderation using automated systems. The Ada Lovelace Institute has developed an audit proposal to help standardize the inspection of social media algorithms.[67] The proposal suggests that the current self-regulative model creates asymmetry between the regulator and the public. Therefore, audits on algorithm inspections, as well as third-party access to involve independent expertise, are necessary. The DSA does address aspects of these technical challenges with the platforms (DSA Art.23 1(c)). It will be up to the practitioners to consider and observe the harmful aspects A.I. – based decisions in content moderation.
- Content that is harmful or classified as "dangerous speech" should be reviewed on a case-by-case basis.
- Add "discrimination of content of body positivity, ableism, queer body concepts" as a reporting category.

References

1 Move Me UC Berkley. (2021, May). Introduction to Social Media and Social Movements. *#MoveMe*. https://moveme.berkeley.edu/introduction-to-social-media/ (accessed June 10, 2021)

2 Me Too Movement. (2021). *Wikipedia. The Free Encyclopedia.* https://en.wikipedia.org/wiki/Me_Too_movement (accessed June 10, 2021)

3 Black Lives Matter. (2021). *Wikipedia. The Free Encyclopedia.* https://en.wikipedia.org/wiki/Black_Lives_Matter (accessed June 10, 2021)

4 Pietschmann, F. (2019, June 6). #FridaysForFuture – Kinder fordern ihre Rechte als Teil einer digitalen Zivilgesellschaft. *Amadeu Antonio Stiftung.* https://www.amadeu-antonio-stiftung.de/fridaysforfuture-kinder-fordern-ihre-rechte-als-teil-einer-digitalen-zivilgesellschaft-48063/ (accessed June 10, 2021)

5 Naikamura, L. (2014). Afterword: Blaming, Shaming and the Feminization of Social Media. https://lnakamur.files.wordpress.com/2011/01/nakamura-afterword-feminist-surveillance-studies.pdf (accessed June 10, 2021)

6 Baldauf, J., Dittrich, M., Hermann, M., Kollberg, B., Lüdecke, R., Rathje, J. (2017). Toxische Narrative. Monitoring rechts-alternativer Akteure. *Amadeu Antonio Stiftung.* https://www.amadeu-antonio-stiftung.de/wp-content/uploads/2018/08/monitoring-2017-1.pdf (accessed June 10, 2021)

7 Internet Usage in the European Union – Internet User Statistics, Facebook Subscribers and 2021 Population for the 27 European Union member states. (2021, May 21). *Internet World Stats.* https://www.internetworldstats.com/stats9.htm (accessed June 10, 2021)

8 Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2021, June 3). *European Commission.* https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence (accessed June 10, 2021)

9 Louis Brandeis. (2021, May 24). *Wikipedia. The Free Encyclopedia.* https://en.wikipedia.org/wiki/Louis_Brandeis#Whitney_v._California_(1927)_–_Freedom_of_speech (accessed June 10, 2021)

10 Digital Technology and Democratic Theory. (2021, March 2). *Data & Society Research Institute.* https://www.youtube.com/watch?v=dcyPOjnMx-c (accessed June 10, 2021)

11 Ibid.

12 sCAN project – Platforms, Experts, Tools: Specialised Cyber-Activists Network (2018-2020). (n.d.). *sCAN.* http://scan-project.eu (accessed June 10, 2021)

13 Hamelmann, M. (2019) Hotspots of Hate. *sCAN.* http://scan-project.eu/wp-content/uploads/scan_analytical-paper-3_Hot-Spots_final.pdf (accessed June 10, 2021)

14 Datta, B. (2014, September 16). Never mind the Nipple-Sex, Gender and Social Media *GenderIT.org.* https://www.genderit.org/feminist-talk/never-mind-nipples-sex-gender-and-social-media (accessed June 10, 2021)

15 Schmidt, F. (2020). Netzpolitik Eine feministische Einführung. *Budrich*. p. 20.

16 Gerrard, Y., & Thornham, H. (2020, July 22). Content moderation: Social media's sexist assemblages. *New Media & Society*, 22(7), p.1266–1286. https://doi.org/10.1177/1461444820912540 (accessed June 10, 2021)

17 Gillespie, T. (2010, February 9). The politics of 'platforms'. *New Media & Society Journal*. Vol 12, Issue 3, 2010. https://doi.org/10.1177%2F1461444809342738 (accessed June 10, 2021)

18 Reuter, M., Köver C. (2019, November 23). Cheerfulness and censorship. *Netzpolitik.org.* https://netzpolitik.org/2019/cheerfulness-and-censorship/ (accessed June 10, 2021)

19 Salty's feminism is not millennial pink. It's not brand-safe, snackable, or neatly packaged for retweets. (2021). *SaltyWorld.* https://www.saltyworld.net/whatwestandfor/ (accessed June 10, 2021)

20 Salty.World. (2021). *Instagram.* https://www.instagram.com/salty.world/ (accessed June 10, 2021)

21 Saltyworldbabes. (2021). *Twitter.* https://twitter.com/saltyworldbabes (accessed June 10, 2021)

22 Dickson E. (2019, July 11). Why Did Instagram Confuse These Ads Featuring LGBTQ People for Escort Ads? *Rolling Stone Magazine.* https://www.rollingstone.com/culture/culture-features/instagram-transgender-sex-workers-857667/ (accessed June 10, 2021)

23 Algorithmic Bias Report. (2019, October). SaltyWorld. https://saltyworld.net/algorithmicbiasreport-2/ (accessed June 10, 2021)

24 Ibid.

25 PlanetRomeo. (2021, May 13). *Wikipedia. The Free Encyclopedia.* https://en.wikipedia.org/wiki/PlanetRomeo#Community_information (accessed June 10, 2021)

26 PlanetRomeoFoundation. (2021). https://www.planetromeofoundation.org (accessed June 10, 2021)

27 PlanetRomeo. (2021, May 13). *Wikipedia. The Free Encyclopedia.* https://en.wikipedia.org/wiki/PlanetRomeo#Community_information (accessed June 10, 2021)

28 Pershan, C. (2020, June). Moderation our (dis)content: renewing the regulatory approach. *Renaissance Numérique.* https://www.renaissancenumerique.org/system/attach_files/files/000/000/279/original/RenaissanceNumerique_Note_ContentModeration.pdf?1613557339 (accessed June 10, 2021)

29 Hutchison, A. (2021, April 8) Facebook Tests Updated Up and Downvoting for Comments in Groups. *SocialMediaToday.* https://www.socialmediatoday.com/news/facebook-tests-updated-up-and-downvoting-for-comments-in-groups/598096/ (accessed June 10, 2021)

30 Sethuraman, R. (2021, March 31). More Control and Context in News Feed. Facebook. https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/ (accessed June 10, 2021)

31 Pershan, C. (2020, June). Moderation our (dis)content: renewing the regulatory approach. *Renaissance Numérique.* https://www.renaissancenumerique.org/system/attach_files/files/000/000/279/original/RenaissanceNumerique_Note_ContentModeration.pdf?1613557339 (accessed June 10, 2021)

32 Kaplan, R. (2018). Content or Context Moderation. Artisanal, Community-Reliant and Industrial Approaches. *Data Society.* https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf (accessed June 10, 2021)

33 Kaplan, R. (2018). Content or Context Moderation. Artisanal, Community-Reliant and Industrial Approaches. *Data Society.* https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf (accessed June 10, 2021)

34 A New EU law on notice and action for removal of illegal content online. My Content, My Rights Conference. (2020, October 1). *Greens EFA.* https://www.youtube.com/watch?v=ma4sIiUgSR8&t=404s (accessed June 10, 2021)

35 Requests for content alteration and takedown. (2020). *Wikimedia Foundation.* https://wikimediafoundation.org/about/transparency/2020-1/requests-for-content-alteration-and-takedown/ (accessed June 10, 2021)

36 Ford, H., Wajcman, J. (2017, March 1). 'Anyone can edit', not everyone does: Wikipedia's infrastructure and the gender gap. *Social Studies of Science, SAGE Journals.* https://journals.sagepub.com/doi/10.1177/0306312717692172 (accessed June 10, 2021)

37 Roberts, S. T. (2018, February 7). Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday.* https://firstmonday.org/ojs/index.php/fm/article/view/8283/6649 (accessed June 10, 2021)

38 PlanetRomeo. (2021). https://www.planetromeo.com/auth/signup (accessed June 10, 2021)

39 Romance scam. (2021, May 31). *Wikipedia. The Free Encyclopedia.* https://en.wikipedia.org/wiki/Romance_scam (accessed June 10, 2021)

40 Doxing. (2021, May 12) *Wikipedia. The Free Encyclopedia.* https://en.wikipedia.org/wiki/Doxing (accessed June 10, 2021)

41 Schroepfer, M. (2019, November 13). Community Standards Report. *Facebook IA.* https://ai.facebook.com/blog/community-standards-report (accessed June 10, 2021)

42 Community Standards Enforcement Report. (n.d.). *Facebook Transparency Center.* https://transparency.facebook.com/community-standards-enforcement#hate-speech (accessed June 10, 2021)

43 GIFCT – Global Internet Forum to Counter Terrorism. (2021). https://gifct.org (accessed June 10, 2021)

44 Perceptual Hashing. (2021, June 1). *Wikipedia. The Free Encyclopedia.* https://en.wikipedia.org/wiki/Perceptual_hashing (accessed June 10, 2021)

45 PhotoDNA. (2021, May 15). *Wikipedia. The Free Encyclopedia.* https://en.wikipedia.org/wiki/PhotoDNA (accessed June 10, 2021)

46 Tech Firms' Counterterrorism Forum Threatens Rights. (2020, July 30). *Human Rights Watch.* https://www.hrw.org/news/2020/07/30/tech-firms-counterterrorism-forum-threatens-rights (accessed June 10, 2021)

47 LumenDatabase. (2021). https://lumendatabase.org (accessed June 10, 2021)

48 Tomasev, N., McKee, K. R., Kay, J., & Mohamed, S. (2021, April 28). Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. *ArXiv.* https://arxiv.org/pdf/2102.04257.pdf (accessed June 10, 2021)

49 Digital Service Act. (2021, June 6). *Wikipedia. The Free Encyclopedia.* https://en.wikipedia.org/wiki/Digital_Services_Act (accessed June 10, 2021)

50 Digital Services Act. (2020, December 15). *European Commission.* p.8. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en (accessed June 10, 2021)

51 General Data Protection Regulation. (2021, May 29). *Wikipedia. The Free Encyclopedia.* https://en.wikipedia.org/wiki/General_Data_Protection_Regulation (accessed June 10, 2021)

52 Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law. (2008, November 28). *European Council.* https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:328:0055:0058:en:PDF (accessed June 10, 2021)

53 Dangerous Speech: A practical guide. (2021, April 19). *Dangerous Speech Project.* https://dangerousspeech.org/guide/ (accessed June 10, 2021)

54 Network Enforcement Act (Netzdurchsetzungsgesetz, NetzDG). (2017, October 1). *German Law Archive.* https://germanlawarchive.iuscomp.org/?p=1245 (accessed June 10, 2021)

55 Declaration on Freedom of Expression. In response to the adoption of the Network Enforcement Law ("Netzwerkdurchsetzungsgesetz") by the Federal Cabinet. (2017, April 15). Declaration on Freedom of Expression. https://deklaration-fuer-meinungsfreiheit.de/en/ (accessed June 10, 2021)

56 Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act, NetzDG) - Basic Information (2017). (n.d.). *Bundesministerium der Justiz und für Verbraucherschutz.* https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/_documents/NetzDG_englisch.html (accessed June 10, 2021)

57 Russian bill is copy-and-paste of Germany's hate speech law. (2019, July 19). *Reporter Without Borders.* https://rsf.org/en/news/russian-bill-copy-and-paste-germanys-hate-speech-law (accessed June 10, 2021)

58 Kaye, D. (2017, June 1). Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. *Office of the High Commissioner for Human Rights.* https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf (accessed June 10, 2021)

59 Liesching, M., Funke, C., Hermann, A., Kneschke, C., Michnick, C., Nguyen, L., Prüßner, J., Rudolph, S., & Zschammer, V. (2021). Das NetzDG in der praktischen Anwendung: Eine Teilevaluation des Netzwerkdurchsetzungsgesetzes (1. Auflage). *Schriftenreihe Medienrecht & Medientheorie:* Vol. 3., Carl Grossmann Verlag

60 Morty, R. (2021). Study: Network Enforcement Act is of little use and leads to over-blocking. *Marijuanapy. The World News.* https://marijuanapy.com/study-network-enforcement-act-is-of-little-use-and-leads-to-over-blocking/ (accessed June 10, 2021)

61 Federal Office of Justice Germany issues Fine against Facebook. (2019, July 3). *Federal Office of Justice Germany.* https://www.bundesjustizamt.de/DE/Presse/Archiv/2019/20190702_EN.html (accessed June 10, 2021)

62 Torrendo, T. (2019, July 20). Facebook fights back against NetzDG fine. *Best Games World.* https://www.bestgamesworld.com/facebook-fights-back-against-netzdg-fine (accessed June 10, 2021)

63 Kompetenzzentrum: Hass im Netz. (n.d.). *Bundesministerium für Familie, Senioren, Frauen, und Jugend.* https://www.demokratie-leben.de/projekte-expertise/kompetenzzentren-und-netzwerke/kompetenzzentrum-hass-im-netz (accessed June 10, 2021)

64 Degenhard, J. (2021, February 1). Forecast of the number of internet users in Europe from 2010 to 2025. Internet users in Europe 2010-2025. *Statista.* https://www.statista.com/forecasts/1145081/internet-users-in-europe (accessed June 10, 2021)

65 Three-click rule. (2020, November 18). *Wikipedia. The Free Encyclopedia.* https://en.wikipedia.org/wiki/Three-click_rule (accessed June 10, 2021)

66 LumenDatabase. (2021). https://lumendatabase.org (accessed June 10, 2021)

67 Inspecting algorithms in social media platforms. (2021). *Ada Lovelace Institute.* https://www.adalovelaceinstitute.org/wp-content/uploads/2020/11/Inspecting-algorithms-in-social-media-platforms.pdf (accessed June 10, 2021)

Imprint

Heinrich-Böll-Stiftung European Union, Brussels
Rue du Luxembourg 47-51, 1050 Brussels, Belgium

Heinrich-Böll-Stiftung Washington, DC, 1432 K St NW, Washington, DC 20005, USA

Contact, Heinrich-Böll-Stiftung European Union, Brussels
Zora Siebert, Head of Program, EU Policy
E Zora.Siebert@eu.boell.org

Contact, Heinrich-Böll-Stiftung Washington, DC
Sabine Muscat, Program Director, Technology and Digital Policy
E Sabine.Muscat@us.boell.org

Place of publication: https://eu.boell.org | https://us.boell.org

Release date: June 2021

Layout: Micheline Gutman, Brussels

Illustrations: Pia Danner, p*zwe, Hannover

Editor: Angela Chen