DOI: 10.38146/BSZ.SPEC.2021.6.2

Boglárka Meggyesfalvi

Policing harmful content on social media platforms

Abstract

Social media content moderation is an important area to explore, as the number of users and the amount of content are rapidly increasing every year. As an effect of the COVID-19 pandemic, people of all ages around the world spend proportionately more time online. While the internet undeniably brings many benefits, the need for effective online policing is even greater now, as the risk of exposure to harmful content grows. In this paper, the aim is to understand the context of how harmful content - such as posts containing child sexual abuse material, terrorist propaganda or explicit violence - is policed online on social media platforms, and how it could be improved. It is intended in this assessment to outline the difficulties in defining and regulating the growing amount of harmful content online, which includes looking at relevant current legal frameworks at development. It is noted that the subjectivity and complexity in moderating content online will remain by the very nature of the subject. It is discussed and critically analysed whose responsibility managing toxic online content should be. It is argued that an environment in which all stakeholders (including supranational organisations, states, law enforcement agencies, companies and users) maximise their participation, and cooperation should be created in order to effectively ensure online safety. Acknowledging the critical role human content moderators play in keeping social media platforms safe online spaces, consideration about their working conditions are raised. They are essential stakeholders in policing (legal and illegal) harmful content; therefore, they have to be treated better for humanistic and practical reasons. Recommendations are outlined such as trying to prevent harmful content from entering social media platforms in the first place, providing moderators better access to mental health support, and using more available technological tools.

Keywords: online policing, harmful content, social media, content moderation, online safety

Introduction

Social media content moderation is an important area to explore as the number of users and the amount of content are rapidly increasing every year. As an effect of the COVID-19 pandemic, people of all ages around the world spend proportionately more time online. While the internet undeniably brings many benefits, the need for effective online policing is even greater now, as the risk of exposure to harmful content grows. This essay aims to understand the context of how harmful content is policed online on social media platforms and how it could be improved. It is intended in this assessment to outline the difficulties in defining and regulating the growing amount of harmful content online, which includes looking at relevant current legal frameworks in development. It is noted that subjectivity and complexity in moderating content online will remain by the very nature of the subject.

It is discussed and critically analysed whose responsibility managing toxic online content should be. It is argued that an environment in which all stakeholders maximise their participation and cooperation should be created in order to effectively ensure online safety.

Acknowledging the critical role human content moderators play in keeping social media platforms safe online spaces, consideration about their working conditions are raised. They are essential stakeholders in policing harmful content; therefore, they have to be treated better for humanistic and practical reasons. Recommendations are outlined such as trying to prevent harmful content from entering social media platforms in the first place, providing moderators better access to mental health support, and using more available technological tools.

Problem areas in policing harmful content on social media platforms

The growing amount of harmful content on social media platforms

In recent years, there has been an increasing concern about harmful content on social media platforms, widely available to an ever-rising number of users. At the beginning of 2021, there were around 4.2 billion active social media users, an almost 14% growth compared to a year before (URL5). Social media is a collective term used for community-focused websites and applications that facilitate the creation and distribution of information through interactive digitally-mediated technologies (Munk, 2021). It includes social networks (e.g., Facebook),

media sharing networks (e.g., Youtube, Instagram, Vimeo) and forums, all of which provide technology for people and organisations to share various types of content like text, images, videos, polls, announcements, links and live streams. According to Caplan (2018), the new challenge of managing content that can be publicly disseminated by anybody, from anywhere and at any time, derives from the fact that platforms of such size and information-density as Facebook and YouTube were unprecedented before.

As highlighted by the Online Harms White Paper (2020), the United Kingdom's new regulatory framework that aims to improve citizens' safety online, it is crucial that all actors take responsibility and cooperate to make the internet a safer place, and online spaces are not surrendered *'to those who spread hate, abuse, fear and vitriolic content'* (URL4). International institutions, such as the European Commission, also expressed the need to effectively manage the growing spread of harmful content online, including harassment on social media, and fake news like false information on the COVID-19 pandemic (European Parliament, 2021).

However, defining what harmful content means can be difficult. In its Digital Services Act proposal (2020), the European Commission stated that it was commonly agreed by stakeholders that defining 'harmful' content should be a subject of ensuing regulations as *'this is a delicate area with severe implications for the protection of freedom of expression'* (European Commission, 2020). The British government also avoided giving an interpretation of the term in the Online Harms White Paper, instead published a non-exhaustive list of harms in scope. The list indicates what types of harmful content or activity had a 'clear' or 'less clear' definition, for example, harassment and cyberstalking falling in the previous, advocacy of self-harm in the latter category (URL4).

While there is no widely accepted, clear definition of what harmful content is, in simple terms it is any content that causes a person distress or harm. This approach, however, is rather subjective and associates a vast amount of content both illegal and legal, making it difficult for people to classify. Perceiving or experiencing distress depends on numerous aspects, including the cultural and religious beliefs, age, and the individual level of sensitivity of a person. For categories such as harassment and fake news, there will always be edge-cases, depending on interpretation, dealing with *'examples where someone's background, personal ethos, or simply their mood on any given day might make the difference between one definition and another'* (URL9).

Policing social media platforms is challenging

Even if it is challenging to define what harmful content is, it is important that it is removed immediately. Regulations and processes need to be in place to ensure that victims are protected, the negative impact is minimised and further harm (e.g. secondary victimisation) is prevented. However, lately, there are more and more questions arising about whose responsibility it is to keep the online space safe and police harmful content on social media platforms.

Policing these platforms is problematic for a number of reasons, including the *'volume of the number of posts that need to be policed; the inter-jurisdictional nature of users; the lack of international cooperation and information-sharing protocols; the ease and anonymity by which the content can be disseminated; and varying legal definitions'* (Williams, Butler, Jurek-Loughrey & Sezer, 2021).

Although there is currently a number of legislations in the pipeline, such as the above mentioned Digital Service Act and the Online Harm Bill, due to the technical complexity and dynamism, and high political sensitivity (Llansó, Hoboken & Leersen, 2020) there are no easy solutions to how effective policing of social media should look like in the future.

Governments in the past preferred self-regulatory approaches, trying to introduce non-binding, voluntary forms of co-regulation, being cautious to introduce determinative regulations regarding harmful content (Llansó et al., 2020). However, since 2018, policymakers in the European Union and the United States of America have started to ask *'increasingly tough questions about how tech giants handle online content'* and push them *'to take greater responsibility for illegal, hateful and false information'* (URL8). Over the past years, the public debate climaxed on the responsibilities and liability of social media companies facilitating the *'mass diffusion of any type of content'* (Bertolini & Cherciu, 2021), leading towards an end of an era of self-regulation and the placement of *'significant legal and practical responsibility on online companies'* (URL4).

It is problematic how social media companies moderate harmful content.

There are several existing strategies for managing content on social media platforms. The way they choose to handle this task and responsibility can depend on their size and the amount of content generated on their platforms. As noted by Gillespie (2018) *'moderation requires a great deal of labor and resources: complaints must be fielded, questionable content or behavior must be judged, consequences must be imposed, and appeals must be considered'*. Giving an example of the extent, in its online transparency report Google revealed that

YouTube removed nearly 35 million videos during 2020, of which almost 2 million were not automatically flagged (URL3).

Caplan (2018) identifies three main types of strategy used by social media companies: artisanal, community-reliant and industrial. Artisanal strategy is followed by platforms such as Vimeo and Patreon, which choose to have smaller teams of moderators operating in-house, doing the job manually with relatively little use of automated technologies. Community-reliant strategy operators, for example, Wikipedia and Reddit, are reliant on their populous volunteer base to respond to moderation needs in their free time, following the previously established content moderation policies of the platforms. The largest global social media companies, such as Google and Facebook, apply the industrial strategy, trying to maximise the use of machine-learning tools and artificial intelligence (AI), operationalising their rules, and often having a significant amount of their policy enforcement work outsourced.

Inevitably, one component shared by all companies in the process of policing harmful content is the necessity for human moderators. Even the most popular social media platforms like Youtube now admit that human review is absolutely critical for them (URL8). Companies invest in technological tools that are becoming sufficiently robust and capable of flagging different types of harmful content (URL6), nevertheless, it will not be able to *'replicate the computing power of an army of human content moderators'*, especially when the content is controversial and *'require local knowledge or cultural cues'* (URL8).

Moderators' working conditions are often not adequate.

While companies following the artisanal strategy might work with in-house teams of not more than 10 members, others have publicly committed to employing more moderators to make their platforms safer. YouTube alone had 10,000 individuals working in such positions in 2018, and Facebook pledged to have 20,000 people in their content moderation and policy teams by the end of the same year (Caplan, 2018). Their job is demanding due to the amount and the nature of the workload they have to review and analyse on one hand, and on the other, because adequate mental and available technological support measures put in place are often lacking (URL6; URL9; URL8). Problems, such as the *'risk of serious shortcomings in the training, working conditions and support provided for content moderators'* are also becoming recognised at the governmental level (URL4).

As Vincent (URL9) summarised *'humans tasked with cleaning up the internet's mess are miserable'*, and it is about time to better explore this area and provide meaningful solutions.

Critical analysis and discussion

Harmful content will stay

In order to better understand how policing harmful content on social media platforms could be more effective, we need to take a closer look at the problems outlined above.

Trends are that social media platforms are gaining ground in the online space, with more people connected, and more content generated every year. There seems to be no reason to expect that with time, harmful content would miraculously disappear or decrease from these online platforms without serious, targeted interventions. If so, it would have probably happened in the past decades. Instead, a change of view is needed to acknowledge that harmful content, at least until now, has been an integral, inalienable part of the content online. Similarly to offline everyday life, there are actors in the online space too, that intentionally or accidentally cause harm or distress to others. Social media companies, users, law enforcement and civil organisations develop in recognising and handling damaging content, but so do criminals and mal-intentioned players in tricking and outsmarting their weaknesses.

Taking into account the vast amount of content produced, defining and regulating what constitutes harmful can have far-reaching effects. For example, in the case of drawing the line between sexually explicit and sexy in a certain way can result in the difference of removing or leaving billions of images accessible online. There will always be harmful content shared online with 'clear' or 'less clear' definitions and 'edge-cases'. The intention behind sharing a particular piece of content might determine whether it is legitimate documentation of a potential war crime or potentially harmful material (URL8). Therefore, the conclusion can be drawn, that by its nature defining what constitutes harmful will remain challenging, and in some cases objectively impossible.

There are no easy solutions to how effective policing of social media should look like in the future

Besides the enormous amount of content, the lack of clear, widely accepted or legal definitions, there are other factors, such as the lack of international cooperation and information sharing, that add to the complexity of policing social media. Powerful actors as companies, governments and supranational organisations, like the European Union, often seem to focus on how to shift the responsibility to another player in the field, rather than developing common strategies.

Social media platforms try to avoid regulating their sites beyond necessity, and frame the users predominantly responsible *'for what they say, read, or watch'* (Gillespie, 2018). Governments and supranational organisations increasingly criticise the companies for acting vaguely and not doing enough, and try to impose stricter rules. The British government, for example, plans to *'establish a new statutory duty of care to make companies take more responsibility for the safety of their users and tackle harm caused by content or activity on their services'* (GOV. UK, 2020). In a recent briefing to the European Parliament on the Digital services act in progress, they stated that concerning illegal and harmful content on platforms *'in the absence of effective EU regulation and enforcement, those platforms set the rules of the game'* (European Parliament, 2021).

In the nodal-like structure of stakeholders in online policing, political and economic interests and power relations determine the margins in which all of them can operate. Trying to shift the responsibility of policing harmful content away can delay taking effective actions, and keep the costs of allocating new resources on hold. However, it is only a relatively small and temporary reward compared to the possibility of an extensive, global cooperation of actors, which could serve the overall online safety of all. Therefore, in order to move towards a more effective way of policing, the real question might not be whose responsibility it is to do what part of the job, but rather how to create an environment in which all stakeholders are motivated and benefit from the enhanced safety in online space.

Human moderators' role will remain essential

Moderation is an important part of policing social media platforms. Human moderators and artificial intelligence are key to keep these online spaces safe from disturbing material and bad-intentioned, potentially dangerous actors. Their work includes ensuring that harmful material is removed instantly, such as terrorist propaganda, child abuse material, live streaming of extreme violence, fake news, advertisements of illegal goods and more. Their capabilities and performance have to be maximized because lives can literally depend on it.

In the midst of the growing public concern for more effective solutions and advancing legal regulations globally, companies race to show more results by using automated technologies like machine learning algorithms. However, all companies have a different level of access to these technologies. Some, that prefer the industrial strategy like Facebook, tried to communicate for a while that artificial intelligence would ultimately solve content moderation on their platforms (URL9) and invested heavily in machine learning. However, it can

be argued that the global COVID-19 pandemic has shown that the enhanced use of technology did not ultimately prove to eliminate or visibly decrease the amount of harmful content online overall. AI *'can both miss content (false negatives) and incorrectly flag unrelated content (false positives)'* (URL6). These mistakes, coupled with slowed down response time, were noted by many since the beginning of the pandemic (URL2; URL8). It also led Facebook to acknowledge that relying more on automated tools has limitations (URL2).

The circumstances of moderators will have to improve

Those limitations then are supposed to be overcome by trained human moderators. The challenge is, that unlike machines and algorithms, humans need proper, tailor-made conditions to succeed in a demanding job like content moderation. By human nature, their performance and efficiency depend on their circumstances. AI follows the same rules each time it receives information, but moderators pick, choose and decide all the time, in all kinds of ways (Gillespie, 2018), based on their understanding of company policy, local and cultural knowledge and individual judgement. Therefore, it is essential to provide adequate working conditions for the tens of thousands of people working in these roles, both because they are entitled to it, but also because it is a prerequisite of effective work. Content moderators are key actors in policing social media online, thus it is unsettling that good conditions have reportedly not been provided by social media companies, neither in-house nor outsourced (URL7).

There are several difficulties that can be identified when analysing why adequate working conditions for content moderators are not sufficiently provided. On a micro level of the individuals, content moderators might not be prepared for the work, or have false expectations of what it will constitute. Many of them come from a 'customer support' background with no previous expectations and burn out quickly without proper mental support (URL1).

On a meso, community and organisational level, companies often lack the resources or the will to provide technological tools that could make moderators work easier (URL6; URL1). Lack of diversity is also another problem, because the *'largely white, largely young, tech-savvy* Californians' (Gillespie, 2018) who form the core group of moderators might find it more difficult to deal with some specific content, compared to a background-wise more diverse group. Another factor that can negatively affect their performance, morale and health is the lack of professional management, including leaders who deny break time, fired employees on flimsy pretexts, and changed shifts without warning, unavailable mental support and no sufficient wellness time provided (URL7).

On the macro level of leadership, legal and cultural context regulations are unarguably missing. It obviously has not been of interest for social media companies to advocate for more restrictions and responsibilities for themselves. Legal regulation, as noted by the government of the United Kingdom, it is a duty that should be initiated on a governmental level, that is why they committed to set out codes of practice, *'outline the systems, procedures, technologies and investment, including in staffing, training and support of human moderators, that companies need to adopt'* (URL4). For positive changes on all levels, there is one other element clearly needed, one that has started to gain recognition: namely that there is little understanding of the long-term effects of viewing disturbing images in quantity, on a regular basis. As more and more users create content that needs to be assessed, there is still a lack of basic knowledge of how the most difficult aspects of this work – removing graphic and disturbing content – affect the people doing it (URL7).

Recommendations

Having discussed the challenges in policing harmful content, whose responsibility it should be, and the maltreatment of moderators, there are recommendations on how online safety could possibly be improved.

Policing less content

An obvious solution to the challenge of policing an enormous amount of harmful content on social media platforms would be to have less of that kind of disturbing content. Easy as it sounds, it would involve a complex coordinated effort from a group of diverse actors on different levels. Social media companies would need to finally take a firm stand and admit their responsibility in policing their platforms, instead of avoiding to face their role in it. More effective prevention of harmful content appearing online in the first place would require companies looking at and modifying their policies (and having hard discussions regarding the freedom of speech), recoding algorithms, allocating more resources to this mission, and not at least, financial loss caused by less traffic on their platforms (URL6). However, efforts by social media companies would not be sufficient alone. Governments and supranational governmental organisations as the European Union, need to provide clear policies and regulations that companies can follow. It is their responsibility to use their democratically legitimised power to make decisions on distinctions of categories of 'less-clear' and not harmful

content. Defining these categories broader or narrower also has implications on what can enter and stay online, and involving experts would bring the necessary knowledge to the discussions.

Governmental organisations also have to facilitate the creation of an environment, in which companies have incentives and become motivated to maximise the safety of their platforms. The British government has already supported the incorporation of *'existing good practices into their products from the earliest stages of product development to ensure that their products are safe by design'* (URL4). The safety by design approach is gaining recognition in the United Kingdom and the United States and would prove to be beneficial globally.

Besides the above mentioned, governmental organisations also need to ensure on a policy level that users, namely their citizens are adequately supported in recognising harmful content, so they can prevent its creation and spread by their means, or adequately deal with it when encountered. It seems a positive development that the British government has already promised to raise awareness, and develop a new online media literacy strategy (URL4). Civil and educational organisations alongside law enforcement have an important role to play in awareness-raising and prevention too, as they can act as intermediaries between policy goals and people.

Finally, users themselves have to be more active in preventing harmful content through learning about its nature, not posting or sharing damaging material, protecting each other online, and helping fast removal when needed. As 'citizens of the digital space' they also have to act responsibly, with caution and respect toward each other, avoiding to create or amplify distressing content online.

Policing better

In addition to the multi-stakeholder collaboration in the field of prevention, legitimate regulations and 'clearer' definitions, there is room for improvement in dealing with the already existing harmful social media content. As discussed previously, despite the technological developments regarding AI, human moderators will be needed for this challenging job in the foreseeable future. As it is becoming recognised, progress is much awaited in this field and could be done in a number of different areas.

First, besides the humanistic considerations, it is important to enhance moderators mental safety in order to improve the quality of the work they deliver. Preventing burnout, desensitisation and the normalization effect regarding harmful content can have a positive effect on retaining people who are good at this job, therefore increasing the quality of online safeguarding. As more users produce

more content on social media platforms, recruiting and retaining people who can do this hard job of high importance is strategically crucial.

Secondly, it would be recommended to use existing technology and tools better. Leetaru (URL6) argues that there are already sophisticated technological solutions available to moderate content online that should be applied together with human power. Content moderation is a largely manual work, and if automatised technology filtered more content adequately, human moderators would have more time and energy to focus on controversial material that cannot be judged by AI. Likewise filtering, technological tools such as 'turner effects' (e.g., blurring and manipulating disturbing images), the black-and-white view option, playing videos backwards, the muted or sliced view can help to protect moderators' mental health (URL1).

There are many other possibilities to improve work experience, and therefore the productivity of human moderators, but the one support feature that was noticed through all materials reviewed was the need for counselling and human support. Providing quality counselling and mental health advice, together with more real wellness time would most likely provide an improvement for people working in these roles.

Conclusion

In this paper, the context of policing online harmful material in social media platforms was analysed from different perspectives. First, the very notion of harmful material and challenges around defining it was explored. As harmful content was found to constitute an integral part of the content on social media platforms by its very nature, it was necessary to discuss whose responsibility it is to police it. The position of a number of relevant actors was looked at, including social media companies, governments and supranational organisations, and users. The conclusion was that in order to effectively deal with, prevent the creation, and minimise harmful content online, all stakeholders must work together in collaboration, taking responsibilities matching their positions, be it making adequate regulations, allocating more funds to moderation, or raising awareness and learning about the problem itself. Finally, the essential need for human moderators was outlined together with their working conditions. It was concluded that AI technology can help a better policing of social media platforms, but not replace humans. Therefore, taking better care of these important players, the human moderators, is essential not only for humanitarian reasons but also in order to enhance online safety.

References

Bertolini, A. & Cherciu, N. (2021). *Liability of online platforms*. Panel for the Future of Science and Technology. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/656318/EPRS_STU(2021)656318_EN.pdf

Caplan, 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

European Commission (2020). *Proposal for a regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC.* European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020PC0825&from=en

European Parliament (2021). *Digital services act.* EU Legislation in Progress, EPRS. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/689357/EPRS_BRI(2021)689357_EN.pdf

Gillespie, T. (2018). *Custodians of the Internet*. Yale University Press.

Hanna, M. (2021). *UK government publishes details of digital regulation to combat 'Online Harms'.* Allen & Overy LLP.

Llansó, E, Hoboken, J. & Leersen, P. (2020). *Artificial intelligence, content moderation, and freedom of expression*. Trans-Atlantic Working Group Working Papers Series. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

Munk, T. (2021). *Online hate speech and harmful material*. Regulatory Practices and Policing Cybercrime. https://mdx.mrooms.net/course/view.php?id=27892

Williams, M., Butler, M., Jurek-Loughrey, A. & Sezer, S. (2021). Offensive communications: exploring the challenges involved in policing social media. *Contemporary Social Science, 16*(2), 227-240. https://doi.org/10.1080/21582041.2018.1563305

Online links to this article

URL1: *SafetyTech 2021 - Resilience and wellness*. 26 Mar 2021. https://www.youtube.com/watch?v=42pNOeTx86c

URL2: *Dave, P. (2020). 'Social Media Giants Warn of AI Content Moderation Errors, as Employees Sent Home'*. WEForum. https://www.weforum.org/agenda/2020/03/social-media-giants-ai-moderation-errors-coronavirus/

URL3: *Google Transparency Report*. https://transparencyreport.google.com/

URL4: *Online Harms White Paper*. Consultation Outcome.
https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper

URL5: *Hootsuite & We Are Social (2021). Digital 2021 Global Digital Overview*. https://datareportal.com/reports/digital-2021-global-overview-report

URL6: *Leetaru, K. (2019). 'The Problem With AI-Powered Content Moderation Is Incentives Not Technology'*. Forbes. https://www.forbes.com/sites/kalevleetaru/2019/03/19/the-problem-with-ai-powered-content-moderation-is-incentives-not-technology/?sh=3e1e849555b7

URL7: *Newton, C. (2019). 'The terror queue'*. The Verge. https://www.theverge.com/2019/12/16/21021005/google-youtube-moderators-ptsd-accenture-violent-disturbing-content-interviews-video

URL8: *Scott, M. & Kayali, L. (2020). 'What Happened When Humans Stopped Managing Social Media Content'*. Politico. https://www.politico.eu/article/facebook-content-moderation-automation/

URL9: *Vincent, J. (2019). 'AI Won't Relieve the Misery of Facebook's Human Moderators'*. The Verge. https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms

Reference of the article according to APA regulation

Meggyesfalvi, B. (2021). Policing harmful content on social media platforms. *Belügyi Szemle, 69*(SI6), 26-38. https://doi.org/10.38146/BSZ.SPEC.2021.6.2