TARGETING TARGETED HARASSMENT: PROBLEMS WITH CRIMINALIZATION AND PLATFORM LIABILITY

Juhana Riekkinen

University Lecturer in Legal Informatics, University of Lapland, Faculty of Law
Yliopistonkatu 8, PO BOX 122, 96101 Rovaniemi, FI
juhana.riekkinen@ulapland.fi; https://research.ulapland.fi/en/persons/juhana-riekkinen

Keywords: ***Harassment, Freedom of Expression, Criminal Law, Liability, Content Moderation, Platforms***

Abstract: *While there are plenty of ways to tackle content-related harms in the online environment, the phenomenon of targeted harassment of individuals presents particular regulatory challenges. Regulatory means that may be justified and effective against clearly illegal content (e.g., child sexual abuse material or certain forms of hate speech) may unduly restrict freedom of expression and/or be inefficient in relation to targeted harassment. In this paper, I explore why this subcategory of harmful online speech may be especially difficult to prevent through the traditional means of criminal law and rules governing liability for third party content. Although these challenges are considerable, I identify approaches that may help to counter targeted harassment or to reduce its harms.*

1. Introduction

The socio-technical development that has led to the current network society has had many implications for freedom of expression. It is not a new notion that the Digital Age and especially the Internet have fundamentally changed the conditions of speech, and among other things, made it easier for individuals to get their messages seen or heard.[1] The flip side is obvious: the technologies that have democratized speech have also amplified harmful and illegal content and activities, such as hate speech, defamation, and sexual abuse.[2]

A harmful phenomenon that has recently gained attention relates to targeted harassment of individuals by masses of people with the help of computer networks, in particular on social media platforms. I start from the premise that this phenomenon is a significant problem that causes harm and suffering to the individuals subjected to it, discourages expression and societal participation of its victims, and potentially undermines the functioning of political and legal institutions. The objective of this paper is to explore what could be done about this problem, which has some key differences in comparison to other speech-related (or content-related) online harms. I argue that these differences render targeted harassment a particularly hard regulatory nut to crack, and well-meaning but misguided regulatory attempts may not only fail to solve the problem, but also endanger freedom of expression in the online environment.

First, I present an overview of the characteristics of targeted harassment as an online phenomenon. I will then proceed to discuss regulatory strategies that are often used or proposed to counter content-related harms, and argue why these strategies are less suited for tackling targeted harassment than for combating other forms of harmful and illegal content. In the chapter concerning criminal law, I focus on Finnish legislative proposals,

[1] See, e.g., Balkin 2004, pp. 6–9.

[2] These dual effects have been noted by, e.g., the European Court of Human Rights (ECtHR) in Delfi AS v. Estonia, Grand Chamber Judgment of 16 June 2015, § 110.

but this discussion highlights issues that bear relevance for other national legislators who consider trying to solve this complex issue with the blunt instrument that is criminal law. In the chapter concerning platform liability, the perspective is more generally on the European and international levels. Finally, I try to outline actions that could be taken to prevent or reduce targeted harassment and its harms to individuals.

2. What is "targeted harassment of individuals"?

In this paper, I discuss a phenomenon that I call *targeted harassment of individuals*. To be precise, what I aim to discuss is what has been dubbed *maalittaminen* in Finnish public and legislative discourse. The Finnish expression derives from military vernacular and refers to designating a target to be attacked (the stem *maali* means 'target' or 'goal'). Varying definitions have been proposed for this somewhat hazy concept, and other possible translations for some forms or elements of *maalittaminen* include *online shaming*, *cyberbullying*, and *cyberstalking*.[3] In any case, an important aspect of *maalittaminen* is that there are at least two (groups of) actors: (1) the one(s) designating the target(s), i.e., the initiator(s); and (2) the ones carrying out the attack, i.e., the participants. In my view, *targeted harassment* is the term that best reflects this aspect, which is also a key factor in the analysis that follows, especially in relation to the personal reach of criminal liability.

In practice, a target may be designated by simply mentioning the name, personal details, or contact information of a person in a context where the receivers of this message are likely to react in a certain way, or by hinting, suggesting or clearly stating that certain kinds of messages should be sent to the target. The content of the initiating message itself may be completely neutral or highly aggressive, and very vague or extremely explicit – context is everything. The initiator may disseminate their message via many forums and channels (public websites, generally used social media platforms, closed discussion forums, mailing lists, group chats, etc.). The subsequent "attack" may, likewise, consist of many types of messages and acts via multiple channels, including social media, instant messaging, e-mail, physical mail, and even in-person aggression. It should be noted that in addition to textual messages by the participants and other traditional forms of speech, ways of participation may include actions specific to social media, such as liking and sharing posts.

It often seems to be taken for granted that targeted harassment necessarily entails organized co-operation in which a group of people commit a number of (speech) acts that are linked by their purpose. No doubt, those who engage in these activities may share opinions, ideologies, or even concrete goals, and their combined actions may reach a result that most participants aim for, or are at least willing to accept. However, co-operation comes in many degrees, shapes and flavors, especially so in the online environment. Some online groups are tight-knit, some loosely woven. People who direct speech at a common target within short intervals of each other may have very different inner motivations and goals. While some participants in a (perceived) group effort may intend to psychologically harm the target, with the aim of manipulating their actions or simply out of malice, others may feel that they are expressing legitimate political criticism. Indeed, their messages, taken separately, may only amount to exactly that, and yet be experienced as part of an overwhelming wave of hate speech by the recipient, thus contributing to the harmful effects that others may be seeking to inflict.

In targeted harassment, as in online communities generally, strict organizational hierarchies or command structures are typically absent. This is a notable deviance from the military analogy suggested by the terminology. Further, the lack of formal hierarchy or organizational structure presents challenges for assigning liability. The blurring of liability is typical for this mode of operating, and it may reflect an intentional choice by the initiators of the harassment campaign. Instead of carrying out a criminal attack (such as sending a death threat or publishing a defamatory message) directly, the initiator may wish obscure their own role in harmful

[3] See Illman 2020, pp. 11–12. The authors of a recently published book chapter addressing the criminalization of this phenomenon opted for the term online shaming, noting that "[this] kind of shaming [– –] is close to or partly overlaps with *doxing*, *trolling*, *virtual mobbing* or *flaming*, and even mere *gossiping*" (Koivukari/Korpisaari 2021, p. 477).

behavior by merely hinting or implying a target to a receptive audience that they can reach in the online environment, and still obtain the result that they wanted (such as the silencing of their ideological opponent, or instilling fear of personal harm in a public official, thus making them unable or unwilling to carry out their tasks). Using other people to commit intentional criminal acts is not a new invention, and criminal law has principles and provisions designed to deal with this. However, the dynamics of the online environment reflected in this phenomenon make it increasingly complicated, in theory and in practice, to assign legal responsibility for actions carried out by a mass of people who often are virtual strangers to each other. The traditional criminal law requirements of complicity may not be met, or are at least difficult to prove.

There is one more characteristic that I deem important to clarify. The kind of targeted harassment that I discuss is targeted, in the first line, at an individual: a natural person, a human being. Targeted harassment of individuals should be distinguished from efforts to direct criticism, hostility, or hate speech at groups or institutions. Granted, the border may be fuzzy; harmful messages may be sent both to organizations and individual public servants or employees simultaneously. Yet, if there is no individual person as a target, the mechanisms of harm and some of the relevant legal questions are quite different. I do not include these situations in my analysis.

While the *direct* target of targeted harassment, as understood here, must be an individual, there may well also be an *indirect* institutional target. Sometimes it has been suggested that targeted harassment would always have the (indirect) goal of affecting the decision-making processes or operations of public institutions, such as administrative authorities, courts, and political institutions.[4] While in many instances this may even be the *primary* goal of the initiators (and possibly of many other participants), as a matter of definition, I do not subscribe to this point of view, as this would either exclude or mischaracterize a significant portion of harassment that is taking place. Although some actors may harass a judge with the clear and conscious intention of affecting their decision-making in an ongoing case, or their and their colleagues' decision-making in general, some hateful and harassing messages directed at judges may simply reflect the senders' frustration or outrage at some real or perceived injustice. Further, there is no shortage of cases where reporters, university researchers, or even ordinary private persons who have appeared in the mass media in negative light have been subjected to online harassment campaigns, although they have no affiliation with any public institution. While the psychological harms and the dynamics of the phenomenon remain similar, in some instances one would be hard-pressed to find that even the initiators of harassment have a clear purpose of affecting the workings of public institutions, or even those of any private organization (such as the employer of the targeted individual). While influencing institutions through the intimidation of individuals certainly is one motivation behind targeted harassment, it is not the only one. Equally, even though it is tempting to try to find some sense in all human actions, we should recognize that sometimes anger and hate have no rational purpose or goal.

3. Criminalization

Criminal law is the *ultima ratio* in regulation: the general principle is that the legislator should only rely on it as a last resort. However, if a problem cannot be solved with less repressive regulatory means, criminal law provides a valuable tool, with which harmful conduct can be banned and those who commit harmful acts can be subjected to punishment. Criminal law has a somewhat strenuous relationship with freedom of expression, as vague criminal provisions or disproportionate sanctions may endanger this important right that is widely guaranteed in international treaties and national constitutions. Nevertheless, freedom of expression is not absolute, and it is well established that criminal law and criminal sanctions can be used to restrict speech *ex post* when certain conditions are met. For example, under the European Convention on Human Rights (ECHR), Article 10(2), limitations on this right are permissible for, e.g., the prevention of disorder or crime and the

[4] In accordance with this view, some Finnish criminalization proposals (discussed in the next chapter) have specifically concerned the targeted harassment of authorities or public officials working for them, not of individuals in general.

protection of the reputation or rights of others, provided that they are necessary in a democratic society and prescribed in law.[5] Defining criminal offences that restrict expression may even be *required* by human rights obligations when speech infringes on or endangers other human rights, such as the right to privacy, the right to life and health, or some other person's freedom of expression.

In most jurisdictions, many forms of targeted harassment and individual acts that can be committed as parts of harassment campaigns are already subject to criminal sanctions. Direct threats of violence, defamation, and invasions on private life are typically punishable under existing, often technology neutral provisions. Still, it may be claimed that the existing criminal law provisions are inefficient or otherwise unsuited to respond to this phenomenon, and that the phenomenon also covers speech that is not currently illegal, but still harmful. In this case, it may be suggested that new criminal law provisions specifically tailored for this phenomenon would be needed. This has been the case in Finland. To understand the problems with – and the limitations of – criminalizing targeted harassment, it is useful to analyze some aspects of the Finnish discussion.[6] At least four different versions of a possible criminal provision have been proposed, but in the following I will focus on two of these: the one proposed by the National Police Board, the Office of the Prosecutor General and the Chief Judges of the District Courts in June 2019,[7] and the one discussed by LL.D., District Court Judge *Mika Illman* in a government-commissioned report on organized online harassment, published in November 2020.[8]

The first proposal would seek to criminalize actions that cause a situation that is conducive to frustrating the activities or decision-making of a public authority. The proposal defines alternative ways by which this situation may be caused: it would apply to harassing or threatening a person in the service of a public authority, or a close person of theirs, and to expressing or spreading unfounded claims related to them. The list is non-exhaustive, as the criminalization would cover creating such a situation "in another comparable manner", and further "knowingly providing a platform" for aforementioned actions. The proposed provision would seek to criminalize an extraordinarily varied range of behavior. First, it would apply not only to initiation of and participation in group activities, but also to actions by single perpetrators, which do not fall under targeted harassment (as defined in the previous chapter) and do not entail similar difficulties in assigning criminal liability. Second, it would seek to include the activities of initiators and participants alike within the same provision, without distinguishing these roles. Third, in addition to the specifically stated activities of harassment, threats, and expressing or spreading unfounded claims, it would seek to criminalize all "comparable manners" of behavior that might have similar effects. Fourth, the provision would include criminal liability of the "platform provider", regardless of the fact that such activities may take place on several platforms and channels, and only parts of the campaign may be visible to administrators.[9] On the other hand, the proposal would not protect all victims of harassment, but merely shield public authorities and those working for them – a problematic approach in respect to the equality of victims and societal power dynamics. With this limitation, the provision

5 The lawfulness test ("prescribed in law") includes the requirement that norms must be formulated with sufficient precision to enable the citizens to regulate their conduct, and so that citizens are able to foresee to a reasonable degree the consequences which a given action may entail. See, e.g., ECtHR, Perinçek v. Switzerland, Grand Chamber Judgment of 25 October 2015, § 131. Generally on this and the other two tests ("legitimacy of the aim pursued" and "necessity of the interference in a democratic society"), see, e.g., European Court of Human Rights (Registry) 2021, pp. 19–24. Under Art. 17, ECHR also recognizes the possibility of denying protection for speech that constitutes abuse of that right.

6 The Finnish proposals have been recently analyzed in detail by Koivukari 2021, pp. 963–992. In line with my analysis here, Koivukari concludes that a comprehensive criminalization is not possible, as the principles of legality and presumption of innocence prevent this. Limitations for criminalization efforts are also set by freedom of expression. Koivukari highlights the impossibility of defining the phenomenon precisely enough to differentiate harmful and malicious shaming from criticism that should be allowed under freedom of expression.

7 National Police Board etc. 2019.

8 Illman 2020, pp. 131–137. Before becoming a judge, Illman was a state prosecutor specialized in free speech cases. His doctoral dissertation (2005) concerned ethnic agitation.

9 Despite the ambiguous reference to platforms, this would not affect major platform providers (such as Facebook or Twitter), as the proposal concerns personal criminal liability, not the liability of legal persons. Thus, it might allow the conviction of an administrator of a website or a Facebook group who failed to take action to prevent harassment.

would concentrate on protecting those with varying degrees of societal power from those who – generally – do not have it. Of course, many initiators of harassment may also wield considerable influence and power.

Criminal offences should be clearly and precisely defined in law. The proposal falls considerably short of this ideal. Without a doubt, the provision would make the field of allowed expression narrower than it currently is. It would also leave the borderline between legal and illegal extremely ambiguous and difficult to foresee. As I see it, the potential efficiency of such a criminal law provision would rely specifically on its broadness, vagueness, and the threat of relatively harsh punishment[10]: the provision would not differentiate between legal and illegal content nor between malicious harassment and bona fide criticism, creating a "chilling effect" that would discourage also legal speech. In Balkin's terms, it would be "old school speech regulation" where the state wants to make sure that all unprotected activity is deferred with no regard on whether it would capture also protected expression.[11] Such a provision would be unlikely to satisfy the "prescribed in law" test set in ECHR Article 10(2), not to mention failing national constitutional tests for restricting fundamental rights.

In his report, Illman rightly rejected the aforementioned proposal (and another, similar but even more vague version) due to incompatibility with international human rights obligations and the requirements of the Constitution of Finland. In particular, he found the proposed provisions too imprecise to fulfill the principle of legality and too restrictive in regard to freedom of expression. However, Illman presented an alternative approach to drafting a criminal law provision on targeted harassment. Instead of trying to capture and criminalize excessive amounts of currently legal expression under the new provision, this approach can be characterized as a collection of existing criminal law norms in a single provision that explicitly recognizes targeted online harassment as a phenomenon. The proposed section on "participation in illegal targeted harassment" would apply to five different types of already unlawful activities when committed in the online environment and in co-operation with others engaging in those activities. The definitions correspond to menace, two forms of defamation, dissemination of information violating personal privacy, and public incitement to an offence.[12] The provision does not strictly follow the division of roles outlined earlier in this paper (initiators and participants), but its list of criminal activities would allow convicting both initiators and other participants in a somewhat foreseeable manner.[13] This provision would also safeguard freedom of expression by including similar exemptions as are already found in the current criminal law sections concerning defamation and dissemination of information violating personal privacy. These exemptions concern criticism that is directed at a person's activities in politics, business, public office, public position, science, art or in comparable public activity, and the presentation of an expression in the consideration of a matter of general importance. In both cases, however, only expressions that do not clearly exceed what can be deemed acceptable are exempted. The clauses do not mean that freedom of expression would be absolute even in the context of important societal debate or political criticism, but they obligate the courts to engage in a balancing act between different (fundamental) rights at stake. Should any sort of targeted harassment provision be added to the law, this kind of limitations to criminal liability would certainly be needed for upholding free speech and related rights.

For most part, Illman's version sidesteps the problems of overbroadness, vagueness, and over-restrictiveness in relation to freedom of expression. However, it would have only limited effect in frustrating targeted harassment. Namely, the provision does not criminalize much – if anything – that is out of reach of existent criminal law provisions. Instead, compared to the existing provisions that it is based on, it contains additional elements

[10] While the basic form of the proposed offence would only carry the possibility of a fine or imprisonment for at most six months, the proposal also includes an (equally vaguely defined) aggravated form, which would carry the maximum sentence of imprisonment for four years.

[11] See Balkin 2014, pp. 2340–2341.

[12] These offences are currently regulated in chapter 25, section 7 (menace), chapter 24, sections 9 (defamation) and 8 (dissemination of information violating personal privacy) and chapter 17, section 1 (public incitement to an offence) of the Criminal Code.

[13] The "incitement" activity would usually apply to initiating a harassment campaign. The other four activities might apply, under specific circumstances, to messages of initiators and participants alike.

that need to be proven beyond reasonable doubt (in particular, the accused's knowledge of other online users' activities, as the provision specifically concerns only activities by a group of people). For law enforcement and prosecution, this means more work and potentially lower conviction rates.

The focus on group activity would, however, allow the courts to consider the harmfulness of one person's actions within the context of a harassment campaign, not just as isolated acts. This might make it possible to convict the author of a message that is insignificant and below the threshold of criminality in itself, but objectively harmful as part of a massive wave of hateful messages. As a way of expanding criminal liability, this is not a trouble-free notion, as differing evaluations of the contextual factors might open the door for restricting and chilling legal speech due to unpredictability and lack of foreseeability. Then again, investigatory and evidentiary challenges might render these potentially problematic aspects moot in practice. As long as the standards of proof were taken seriously, the provision would be unlikely to lead to any increase in convictions. Factoring in the collective nature of targeted harassment would be less problematic in sentencing: if a person's activities are clearly illegal in themselves, taking their context within a harassment campaign into consideration might allow the court to assess the severity of the offence in a more appropriate way. This could be a potential, albeit small, improvement to current law.

Another possible advantage of such a provision could be symbolic value. By specifically recognizing online targeted harassment in the criminal code, the state would send a message to those engaging in such activities. Yet, it is dubious if such an official recognition of the phenomenon would have any real, lasting effect in hindering the harmful behavior it seeks to target. Further, the symbolic value might even end up being negative – if the new provision was shown to be impractical, inefficient, and impossible to prosecute, it could even become a sort of confirmation of impunity of targeted harassment. Speculative symbolic value alone cannot justify creating new criminal offences. Tellingly, even Illman himself ultimately concludes that there is no need for this provision.[14]

4. Platform liability

Platform liability has been a focal point in recent discussions concerning online harms. On both sides of the Atlantic, online intermediaries have traditionally enjoyed wide immunity from liability for third party content: in the US, this has been based on the provision known as "Section 230";[15] in EU law, on the "safe harbor" regime established by the e-Commerce Directive.[16] In Europe, the ECtHR Grand Chamber Judgment in *Delfi AS v. Estonia* can be seen as a departure from the previously restrictive view of intermediary liability,[17] but the ECtHR has been careful not to further expand intermediary liability for user-generated content in subsequent case law.[18] The current EU liability framework is set to be replaced by the Digital Services Act, proposed by the European Commission in December 2020,[19] and EU laws affecting the liability and other responsibilities

[14] See Illman 2020, pp. 137–148. In tackling targeted harassment, Illman sees more promise in increasing the liability of administrators and service providers, and suggests creating a new provision on the criminal liability of administrators in cases where they neglect to remove illegal content.

[15] 47 U.S.C. § 230. Generally, see, e.g., Balkin 2014, pp. 2313–2314, Klonick 2018, pp. 1604–1609, Gillespie 2018, pp. 30–35, and Kosseff 2019.

[16] Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce), Art. 12–15.

[17] The ECtHR found no violation of ECHR Art. 10 in the national courts finding the provider of a major commercial news portal liable for illegal posts in the user comments section of a news article, despite moderation measures that were in place, and eventual (slow) removal of illegal comments.

[18] See ECtHR, Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary, Judgment of 2 February 2016 (finding the platform liable violated ECHR Art. 10). See also Pihl v. Sweden, Decision of 9 March 2017, Tamiz v. the United Kingdom, Decision of 12 October 2017, and Høiness v. Norway, Judgment of 19 March 2019 (rejecting claims against the administrator/platform did not violate ECHR Art. 8). About the Delfi and MTE cases, see Maroni 2020, pp. 255–278. See also Polanski 2018, pp. 873–874.

[19] COM(2020) 825 final.

of intermediaries in specific contexts have been recently enacted.[20] The question in this paper is not what current law states about the liability of platforms or other online intermediaries in regard to targeted online harassment. Instead, I concentrate on whether broader platform liability might help in reducing targeted harassment in the future, and if so, how and at what cost.

The appeal of platform liability derives from the fact that online platforms are in a position to monitor and control massive amounts of content. Additionally, as private actors, they are not currently subject to establishing and adhering to strict procedural standards or safeguards, as is the case in criminal proceedings, and therefore can make decisions faster and in a more flexible way. As a further advantage, platforms are not limited to binary decision-making (legal/illegal, leave/delete), but can also affect the reach of content in more subtle ways. All platforms engage in *content moderation* in one way or another, even if not commanded to do so by law.[21] This is essential for maintaining a service that their users will enjoy. The threat of civil or criminal liability, however, is an efficient way to incentivize platforms to remove certain content that lawmakers wish to have removed.

In cases where content is clearly illegal, regardless of the context (e.g., most forms of child sexual abuse material), large platforms can be expected to cope with removal obligations without compromising legal expression, even within strict deadlines. When the legality of speech depends heavily on contextual evaluation, instead, platforms are put in a difficult situation, as even the largest ones do not have the resources for comprehensive case-by-case review of all user content by legal experts, and have to rely on automated measures, or masses of human content moderators operating without legal training and typically under very strict time constraints. In these situations, if a decision to leave illegal content online is penalized, and a decision to remove legal content has no consequences, the logical course of action for the platform is to systemically over-censor rather than risk under-censoring. Thus, intermediary liability always creates some risk of biased incentives and *collateral censorship*.[22] In the current online environment, censorship by private entities may considerably narrow the field of permissible expression – much more than is possible through traditional state censorship or criminal law restrictions on expression.

Harassment campaigns may include actions and messages that are clearly illegal, and thus not difficult to identify as such. However, many if not most messages – especially those that initiate harassment campaigns – are markedly context-dependent. They may be innocent-looking as such, yet intended to cause harm. Then again, many messages sent during harassment campaigns are, in fact, perfectly legal and should not be restricted, regardless of whether they may cause some inconvenience to a person and regardless of whether illegal messages are being posted simultaneously. This means that a party deciding whether or not to remove messages perceived as, or claimed to relate to, harassment would need to make particularly complex legal and contextual evaluations to be able to distinguish legal and illegal, or even harmless and harmful messages. Should platforms be held increasingly liable for harassment campaigns, there would be a strong incentive to over-censor even perfectly normal discussion, including political and societal debate and criticism that is crucial for a functioning democratic *Rechtstaat*. This problem is highlighted by the fact that some (although not all) victims of targeted harassment are public figures and people in positions of power, who are expected to endure more public criticism than ordinary people.

---

20 Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, and Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online.

21 See Gillespie 2018, pp. 1–23, 207. For a definition of content moderation, see, e.g, Jhaver/Bruckman/Gilbert 2019, p. 4: "Content moderation determines which posts are allowed to stay online and which are removed, how prominently the allowed posts are displayed, and which actions accompany content removals." For a taxonomy, see Goldman 2021, pp. 30–51.

22 See Balkin 2014, pp. 2309–2311 and Balkin 2018, pp. 1176–1177. Cf. Wu 2013, pp. 293–349, who recognizes the need for intermediary immunity to avoid collateral censorship when the incentives of the original speaker and the intermediary are not aligned, but argues that avoiding collateral censorship does not necessitate unbounded immunity.

Upholding intermediary immunity and leaving the moderation decisions to the platforms themselves, to be decided according to platform-specific rules or through some common self-regulatory framework without the risk of heavy sanctions, should reduce the risk of collateral censorship caused by unbalanced incentives. Still, it does not make drawing the line between permissible and non-permissible much easier. In any case, the increased risk of collateral censorship and the complexity of distinguishing between legal and illegal are not the only reasons why platform liability may be a badly suited approach for hindering targeted harassment of individuals. Another one relates to the fact that campaigns often utilize multiple channels, thus crossing platform borders as well as jurisdictional borders. To effectively suppress an ongoing harassment campaign, at least most major platforms would have to enforce similar (if not the same) policies for moderating content that may cause or is in itself harassment. Otherwise, the (re)actions of one platform may not so much prevent as redirect ongoing harassment, and effective mitigation may not be achieved even by means of over-censoring on some platforms. While this does not make platform liability more appealing, it does suggest a need for some common and if possible, global framework, be it regulatory or self-regulatory.

Even more importantly, although harassment campaigns are initiated via messages published to an audience, further harassing messages may be sent directly to the victim as end-to-end-encrypted instant messages (which cannot be subject to content-based *ex ante* moderation), via e-mail, or by utilizing other forms of private, one-on-one communication. A platform cannot mitigate harassment by removing an illegal message from view if it is not published in the first place. This also creates a problem of knowledge: if a platform is only evaluating published posts, they may not be able to place those messages in context with a harassment campaign that is taking place partially in closed channels or offline. This further impairs the platforms' chances of effectively reacting to harassment by timely content removals and other such measures. The threat of liability cannot counter this problem, and will, again, only aggravate the problem of collateral censorship.

Victims of harassment may take steps to shield themselves by limiting their availability in many online messaging services, for example by blocking messages from strangers or by using filters that flag messages for deletion or isolate them in separate folders. The platforms' role is to provide the user with tools that make this possible. However, if the harassment campaign consists of hateful messages sent to a work-related e-mail account, the victim can hardly turn off their e-mail without impairing their ability to work. Similarly, blocking messages sent to a private account may result in the victim missing important personal communications, and harm their ability to maintain their social relations. Service providers can ban those who abuse their messaging services (again, possibly subject to tough contextual evaluation on whether the messages constitute abuse or not), but individual platforms – even the major ones – are in no position to block all potential channels of harassment.

5. Ways forward?

In addition to reactively removing published content, banning (or "de-platforming") abusive users and helping to block channels of private messaging, online platforms may be able to do something more pre-emptive. On social media, algorithms play a role in community-building, as they influence what kind of content and people users come into contact with. Roughly speaking, if a user engages with content related to knitting (e.g., by watching a knitting tutorial video), a typical social media algorithm might recommend that user other knitting related videos, posts, or pages in the future. If the original content that the user engages with is racist or misogynist in nature, the same algorithm might lead the user to more racist or misogynist content, or to groups in which members show interest to that kind of content. The precise effects and consequences of algorithms are difficult to ascertain and verify, but these kinds of algorithms do contribute to building communities, some of which may be potential breeding grounds for harassment campaigns. As gathering-places of like-minded people, they may increase the reach of messages that call for harassment of an individual.[23]

[23] This is related to the discussion on algorithmic radicalization and, to some extent, on filter bubbles. See, e.g., Woods 2021, pp. 83–85 and De Gregorio/Stremlau 2021, p. 438. Cf. Bruns 2019 (criticizing the concept of filter bubble).

Community-building and content-curating algorithms are important features of social media. To completely stop recommending users content based on their interests is not a realistic solution. Still, there surely are ways for platforms to limit the harmful effects of their algorithms. Platforms may even have natural incentives to modify their algorithms in order to create safer, more enjoyable and positive environments for their users. A counterpoint is that prevalent platform business models are typically tied to user engagement and ad revenue measured in clicks and page views, and even engagement with harmful consequences is profitable. Therefore, it may be necessary to regulate how platforms construct their algorithms. This kind of regulation, however, would have little in common with traditional liability rules for individual pieces of user-generated content. Instead, the approach would need to be more *systemic*, and focused on proactive responsibilities and duties of platforms relating to how their services are *designed*.[24] Naturally, there are limits to what can be achieved with this approach – it would not eliminate all online harassment or dissolve all existing problematic communities and subcultures (although, with the right combination of moderation techniques, platforms could definitely make their operation more difficult and/or force them to relocate). Yet, it would be a good start if social media algorithms were aimed at reducing the possibilities of harassment – instead of actively contributing to them.

A possible solution to content-related problems that has recently gained popularity is so called *human rights-based* approach to content moderation.[25] This includes attention to informational and procedural rights of users, such as transparency and the right to review of moderation decision, but also the tying of moderation decisions to substantive international human rights standards, and the selection of moderation measures according to the idea of the least restrictive measure, in line with the tests traditionally applied for restricting human rights. There is undeniable promise and appeal in this, and it may indeed provide a regulatory paradigm that is both more efficient in reducing online harms and more sensitive to human and fundamental rights than the more traditional approaches based on criminal law and the liability of intermediaries. However, the efficacy of this approach is yet to be tested, and also critical remarks have been presented by scholars.[26] Concerning targeted harassment in particular, it is hard to evaluate the practical usefulness of the human rights-based model.

A further, more traditional solution to targeted harassment that I find worth exploring relates to the criminal justice system. Instead of trying to maximize the field of criminalized online behavior by creating new offences or by introducing harsher penalties, existing criminal law should be enforced more efficiently in the online environment. To achieve this, many actions are necessary. For example, any existing procedural roadblocks should be removed,[27] as far as this can be done without compromising the fairness of criminal proceedings. Methods of investigation and evidence collection should be developed, more resources should be directed to investigation, and investigations should focus on low-hanging fruit, such as clearly illegal acts with known perpetrators, and the offenders who cause the most harm. This might go a long way in dissolving the atmosphere of impunity that may currently encourage people to participate in harmful online harassment.

[24] See, e.g., Woods 2021, pp. 77–97. Similarly, Hartzog 2018, pp. 223–229 argues that social media services should mitigate the risk of online harassment through design. In economic terms, the *transaction costs* of harmful speech and access to the intended victim's information should be increased, and the costs of defending against harassment should be lowered. Cf. Douek 2021, pp. 759–833, who proposes systemic *balancing* as an approach to moderation.

[25] See Kaye 2018, pp. 14–20.

[26] For a balanced overview, see Sander 2020, pp. 939–1006. See also Aswad 2018, pp. 57–67.

[27] E.g., some Finnish criminal offences that may apply to targeted harassment are offences the prosecution of which rests with the complainant, i.e., the prosecutor may not press charges without the victim's initiative and permission. For various reasons, victims may be unwilling to engage in prolonged legal proceedings with their harassers. To counter this problem, the Criminal Code was recently amended to give prosecutors the power to independently press charges on menace when the threatening act is committed for reasons related to the victim's work tasks.

6. References

Aswad, Evelyn Mary, The Future of Freedom of Expression Online, Duke Law & Technology Review, Vol. 17, Issue 1, 2018, pp. 26–70.

Balkin, Jack M., Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society, New York University Law Review, Vol. 79, Issue 1, 2004, pp. 1–55.

Balkin, Jack M., Old-School/New-School Speech Regulation, Harvard Law Review, Vol. 127, Issue 8, 2014, pp. 2296–2342.

Balkin, Jack M., Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation, UC Davis Law Review, Vol. 51, Issue 3, 2018, pp. 1149–1210.

Bruns, Axel, Filter bubble, Internet Policy Review, Vol. 8, Issue 4, 2019.

COM(2020) 825 final, 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, Brussels 15 December 2020.

De Gregorio, Giovanni/Stremlau, Nicole, Platform Governance at the Periphery: Moderation, Shutdowns and Intervention. In: Bayer, Judit/Holznagel, Bernd/Korpisaari, Päivi/Woods, Lorna (Eds.), Perspectives on Platform Regulation: Concepts and Models of Social Media Governance Across the Globe, Nomos, Baden-Baden 2021, pp. 433–450.

Douek, Evelyn, Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability, Columbia Law Review, Vol. 121, Issue 3, 2021, pp. 759–833.

European Court of Human Rights (Registry), Guide on Article 10 of the European Convention on Human Rights, Freedom of Expression. Updated 30 April 2021. https://www.echr.coe.int/Documents/Guide_Art_10_ENG.pdf (accessed on 14 November 2021).

Gillespie, Tarleton, Custodians of the Internet – platforms, content moderation, and the hidden decisions that shape social media, Yale University Press, New Haven & London 2018.

Goldman, Eric, Content Moderation Remedies, Michigan Technology Law Review [forthcoming], draft 24 March 2021. https://ssrn.com/abstract=3810580 (accessed on 14 November 2021).

Hartzog, Woodrow, Privacy's Blueprint: The Battle to Control the Design of New Technologies, Harvard University Press, Cambridge, MA & London 2018.

Illman, Mika, Järjestelmällinen häirintä ja maalittaminen: Lainsäädännön arviointia, Valtioneuvoston kanslia, Helsinki 2020.

Jhaver, Shagun/Bruckman, Amy/Gilbert, Eric, Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit, Proceedings of the ACM on Human-Computer Interaction, Vol. 3, Issue CSCW, 2019, Article 150.

Kaye, David, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/38/35, 6 April 2018.

Klonick, Kate, The New Governors: The People, Rules, and Processes Governing Online Speech, Harvard Law Review, Vol. 131, Issue 6, 2018, pp. 1598–1670.

Koivukari, Kristiina, Voisiko maalittaminen olla rangaistavaa? Lakimies, Vol. 119, Issue 6, 2021, pp. 963–992.

Koivukari, Kristiina/Korpisaari, Päivi, Online shaming – a New Challenge for Criminal Justice. In: Bayer, Judit/Holznagel, Bernd/Korpisaari, Päivi/Woods, Lorna (Eds.), Perspectives on Platform Regulation: Concepts and Models of Social Media Governance Across the Globe, Nomos, Baden-Baden 2021, pp. 473–487.

Kosseff, Jeff, The Twenty-Six Words That Created the Internet, Cornell University Press, Ithaca & London 2019.

Maroni, Marta, The liability of internet intermediaries and the European Court of Human Rights. In: Petkova, Bilyana/Ojanen, Tuomas (Eds.), Fundamental Rights Protection Online: The Future Regulation of Intermediaries, Edward Elgar, Cheltenham & Northampton, MA 2020, pp. 255–278.

National Police Board etc., Lainsäädännölliset muutostarpeet viranomaisten maalittamiseen puuttumiseksi, Initiative to the Ministry of Justice, ID-19121964, POL-2018-54628, 17 June 2019.

Polanski, Paul Przemysław, Rethinking the notion of hosting in the aftermath of Delfi: Shifting from liability to responsibility? Computer Law & Security Review, Vol. 34, Issue 4, 2018, pp. 870–880.

Sander, Barrie, Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation, Fordham International Law Journal, Vol. 43, Issue 4, 2020, pp. 939–1006.

Woods, Lorna, Introducing the Systems Approach and the Statutory Duty of Care. In: Bayer, Judit/Holznagel, Bernd/Korpisaari, Päivi/Woods, Lorna (Eds.), Perspectives on Platform Regulation: Concepts and Models of Social Media Governance Across the Globe, Nomos, Baden-Baden 2021, pp. 77–97.

Wu, Felix T., Collateral Censorship and the Limits of Intermediary Immunity, Notre Dame Law Review, Vol. 87, Issue 1, 2013, pp. 293–349.