Upload-Filters

Bypassing Classical Concepts of Censorship?

by **Amélie Pia Heldt***

Abstract: Protecting human rights in the context of automated decision-making might not be limited to the relationship between intermediaries and their users. In fact, in order to adequately address human rights issues vis-à-vis social media platforms, we need to include the state as an actor too. In the German and European human rights frameworks, fundamental rights are in principle only applicable vertically, that is, between the state and the citizen. Where does that leave the right of freedom of expression when user-generated content is deleted by intermediaries on the basis of an agreement with a public authority? We must address this question in light of the use of artificial intelligence to moderate online speech and its (until now lacking) regulatory framework. When states create incentives for private actors to delete user-content pro-actively, is it still accurate to solely examine the relationship between platforms and users? Are we facing an expansion of collateral censorship? Is the usage of soft law instruments, such as codes of conduct, enhancing the protection of third parties or is it rather an opaque instrument that tends to be conflated with policy laundering? This paper aims to analyse the different layers of the usage of artificial intelligence by platforms, when it is triggered by a non-regulatory mode of governance. In light of the ongoing struggle in content moderation to balance between freedom of speech and other legal interests, it is necessary to analyse whether or not intelligent technologies could meet the requirements of freedom of speech and information to a sufficient degree.

Keywords: Freedom of expression; censorship; democratic legitimation; upload-filters; prior restraint

Recommended citation: Amélie Pia Heldt, Upload-Filters: Bypassing Classical Concepts of Censorship?, 10 (2019) JIPITEC 57 para 1.

A. Introduction

1 Considering that user-generated content constitutes both speech in constitutional terminology as well as the basis for many social media platforms'[1] business models, its regulation poses many challenges. Social media platforms, or to put it more generally, *intermediaries*, rely on user-generated-content to attract other users. To sustain their attention and, by extension, revenue from advertisers, social networks are dependent on the activity of users on the one hand and on a clean, confidence-inspiring environment on the other. Examples such as the decline of MySpace[2] or the almost non-existent moderation policy at

* Amélie P. Heldt is a junior researcher and doctoral candidate with the Leibniz Institute for Media Research/Hans-Bredow-Institute, Hamburg, and currently a visiting fellow with the Information Society Project at Yale Law School.

1 In this article, "intermediaries" is used as a generic term for "social media services, platforms and networks". They will be used as synonyms for Internet-based applications that rely on user-generated-content to create online communities to share information, ideas, personal messages, etc. Definition retrieved from <https://www.merriam-webster.com/dictionary/social%20media> accessed 23 January 2019.

2 Stuart Dredge, 'MySpace – what went wrong: "The site was a massive spaghetti-ball mess"' (2015) <https://www.theguardian.com/technology/2015/mar/06/myspace-what-went-wrong-sean-percival-spotify> accessed 10 December 2018.

4chan have led to the assumption that a minimum level of content moderation is inevitable. Because of the immense amount of uploaded content that they have to negotiate, social networks fall back on technology to detect and, at times, remove illegal or undesirable content.

2 Deleting a post is, first of all, subject to the intermediaries' community guidelines, but content deletion can also be interpreted as (collateral) censorship if its legal basis is a law or even an agreement (such as a code of conduct) between intermediaries and legislators. Examining automated content deletion via upload-filters raises questions about the technology used, as well as the normative framework of intermediaries when they act on grounds of so-called "soft law". First, this paper will provide an overview of the protection of speech under German Basic Law and the European Convention on Human Rights (ECHR). Second, the increasing use of upload-filters in content moderation – especially to counter terrorist propaganda via user-generated content – will serve as a use case. This type of automated speech regulation could potentially be classified as censorship under certain conditions, an examination of which will constitute the third section of this paper.

B. Protection of freedom of speech and the notion of censorship

3 Social media platforms aim at connecting people globally, inevitably linking various jurisdictions through their contractual relationship with users. Freedom of expression and the notion of censorship are relevant in this context because users might feel violated in their freedom of expression when the content they have uploaded is deleted or blocked. In order to assess whether the use of filters for content moderation purposes is in accordance with our human rights framework, we need to first examine the scope of protection.

I. Under art. 5 German Basic Law

1. Broad protection of free speech

4 In Germany, freedom of speech is protected by art. 5 (1) Basic Law; this clause provides a relatively broad scope of protection. It protects freedom of expression and information as well as important ancillary rights to access means of expression and information, including the whole communicative process and all types of speech, regardless of its topic and its commercial worth.[3] Freedom of speech protects factual claims and value judgments and is considered fundamental to German democratic understanding.[4] This protection under art. 5 (1) Basic Law is, however, not boundless; there are limits to speech through general laws, youth protection, and the honour of third parties (art. 5 (2) Basic Law). Limiting fundamental rights by law is not an essential characteristic of freedom of speech: in German constitutionalism, only very few fundamental rights are guaranteed unconditionally, most can be restricted by law if the restriction is proportionate.

5 The restrictions allowed by constitutional proviso in art. 5 (2) Basic Law are themselves bound to certain requirements: in order to prevent state influence on speech targeting laws, the German Federal Constitutional Court (FCC/BVerfG) elaborated the principle of interdependency (so-called "Wechselwirkungslehre"); this means that not only should the laws restricting speech be in accordance with the scope of protection, but their case-related use needs to be reasonable and adequate when it comes to freedom of speech.[5] This doctrine is, on the one hand, a guarantee for a moderate application of speech-restricting laws and, on the other, it adds a certain complexity when balancing freedom of speech with other rights.

2. Limits to free speech

6 According to the FCC, any law restricting speech needs to serve a higher constitutional purpose than the freedom of expression. It also has to be proportionate and neutral as to the content of the opinion expressed.[6] For obvious reasons, laws according to art. 5 (2) Basic Law shall be as general as possible as to avoid any connection between the purpose of the law and opinions expressed. This means that statements may be punishable by law, but only in order to protect other rights and not to forbid certain opinions.[7] The law may never forbid an opinion due to a concrete political, religious, or ideological position. With this strict criterion, art. 5 Basic Law can guarantee that freedom of expression is only restricted by an opinion-neutral regulation.

7 For example, publicly calling for an unlawful action is penalised just as it would be if it was an incitement under section 111 German criminal code (StGB); i.e. it

3 Jurisprudence of German Federal Constitutional Court: BVerfGE 90, 241, 247.

4 BVerfGE 85, 1, 15; BVerfGE 5, 85, 205.

5 BVerfGE 7, 198, 208 f.

6 BVerfGE 124, 300.

7 BVerfGE 124, 300, 322.

bears the same legal consequence as committing the unlawful action itself. Calling for unlawful action can be considered as expressing an opinion, which makes sec. 111 StGB a speech-restricting law when the speaker is addressing an audience and calling upon them to commit violence. To fulfil the "publicity" criterion the speaker needs to be targeting an indeterminate number of potential recipients, not an individual or specific audience member (in contrast to an individual address such as a private message).[8]

8 At first glance, the use case of this paper – automated filtering and removal of online terrorist propaganda – does not violate the protection of fundamental rights. Uploading a video with a specific message which incites violence is highly likely to meet the requirements of criminal offences. Posting a video on a social network that calls for violence, a "holy war", or for the support of specific terrorist actions is covered by sec. 111 StGB because the internet and social networks in particular may be considered as "public space[s]".[9] To summarise, one cannot be punished for defending a religious belief by expressing his or her opinion but, rather, for calling on others to harm "all non-believers". Restricting this type of speech is therefore in line with the scope of protection outlined in art. 5 (1) Basic Law, unless its enforcement violates the ban on censorship.

3. Uncompromising ban on censorship

9 In German constitutional methodology, restrictions of art. 5 (1) Basic Law have to comply with the so-called restrictions of restrictions ("Schrankenschranke"), amongst others the ban on censorship which is enshrined in art. 5 (1) 3 Basic Law and cannot be subject to adaptations. According to the prevailing opinion in German constitutional jurisprudence and scholarship, censorship can only be the consequence of the obligation to submit a medium to a state agency for *prior* approval of the publication *before* it is produced or distributed.[10] The addressees of this rule are restricted to government agencies, that is, only state-driven actions are forbidden by art. 5 (1) 3 Basic Law and, in principle, the actions of private individuals or entities are not affected under its purview.[11] It shall be referred to as pre-censorship, in contrast to reviewing and possibly deleting content *after* publication or distribution. The majority of scholars are reluctant to extend the ban on this type of pre-censorship to non-state-driven actions.[12]

10 However, this formal and quite conservative interpretation might be subject to changes in the context of online intermediaries.[13] In view of increasing cooperation between tech companies and public authorities,[14] some have argued against this narrow interpretation of censorship that leaves no space for the examination of pre-censorship by private entities.[15] According to Justice Hoffmann-Riem (former judge at the FCC), controlling content on the internet (e.g. by filtering) is only covered by contractual freedom to the extent that it affects persons who have contractual relationships with the respective provider and have thereby consented to control and filtering. Furthermore, the state's duty to protect could require precautions which make it possible to use the infrastructures that are important for the general provision of communications without a framework that is similar to censorship.[16] Löffler, too, believes that the free development of intellectual life can only be guaranteed if the prohibition of censorship also addresses non-state institutions and private instances that have a significant influence on intellectual life.[17] When looking at the power private entities have over our digital communications' infrastructure, holding on to the classical definition of strictly state-driven censorship appears questionable.

II. Freedom of speech in the ECtHR jurisprudence

1. Protection under art. 10 ECHR

11 The jurisprudence of the European Court of Human Rights (ECtHR) on matters of freedom of speech and its protection under art. 10 ECHR has a rich tradition. Between 1959 and 2012 the court

8 Federal Court of Justice: BGH, NStZ 1998, 403, 404.

9 Karl-Heinz Ladeur, ´Ausschluss von Teilnehmern an Diskussionsforen im Internet – Absicherung von Kommunikationsfreiheit durch "netzwerkgerechtes" Privatrecht´ [2001], MMR, 787, 791.

10 BVerfGE 33, 52, 71; BVerfGE 47, 198, 236.

11 Herbert Bethge, Art. 5 Basic Law, *Grundgesetz-Kommentar*, (2014), para 135.

12 Bethge (n 11), para 133; Ansgar Koreng, *Zensur im Internet*, (2010), 235.

13 Christoph Grabenwarter, Art. 5 Grundgesetz, in Maunz/Dürig (eds.) *Grundgesetz-Kommentar* (2018), para. 119.

14 Michael Birnhack, Niva Elkin-Koren, 'The Invisible Handshake: The Re-emergence of the State in the Digital Environment' (2003), 1, Virginia Journal of Law & Technology, 49-52.

15 There is also an ongoing discussion about whether platforms should be bound to the human rights framework through a horizontal binding effect. This is however not the core issue of this paper because it rather focusses on the state acting *through* the platforms in a non-transparent manner, instead of platforms *acting as* public actors.

16 Wolfgang Hoffmann-Riem, Art. 5 Grundgesetz, *Alternativkommentar-Grundgesetz* (2001), para 95.

17 Martin Löffler, ´Das Zensurverbot der Verfassung´ (1969), 50, NJW, 2225, 2227.

asserted 512 infringements of art. 10 (1) ECHR[18] and has shaped a solid case law in balancing freedom of speech and personality rights, which deserves special mention. That being said, the jurisprudence of the ECtHR exists in harmony with the German constitutional understanding of freedom of speech mentioned above: expressions of opinion are protected as long as they do not incite violence. The scope of protection of art. 10 (1) ECHR is similarly broad: it protects the freedom of opinion and of expression and takes into account all opinion and expression of opinion regardless of subject matter, intellectual veracity, or social utility, including trivial, entertaining, commercial, absurd, as well as aggressive and offensive statements.[19] In other words, speech cannot be restricted in accordance with art. 10 (1) ECHR as long as it does not endorse the use of violent procedures or bloody revenge, nor justify the instruction of terrorist acts or potentially incite to violence due to profound and irrational hate towards certain people.[20]

2. No absolute ban on censorship

12 One difference between art. 5 Basic Law and art. 10 ECHR lies in the more restrictive interpretation of the ban on censorship. According to art. 10 ECHR, interventions that constitute censorship are not inadmissible per se. Rather, they must satisfy the principle of proportionality whereby the particular severity must in any case be taken into account.[21] The prohibition of censorship is to be derived – although not explicitly mentioned – from the prohibition of intervention by the authorities in accordance with art. 10 (1) 2 ECHR.[22] Accordingly, it is not surprising that interventions are only permissible within narrow limits and that the ECtHR carries out a detailed review of corresponding measures.[23] So-called "prior restraints"[24] are only permissible if they do not result in a complete prohibition of publication, if the information is less than current, if rapid court proceedings on prohibition orders are possible, and if complex issues of fact and law are clarified in the process.[25] The court has established in numerous cases that prior restraint is not prohibited *per se*,[26] which is the crucial difference when comparing it to art. 5 (1) 3 Basic Law. Nonetheless, the general protection and interpretation of freedom of speech by the FCC and the ECtHR is largely similar, especially when it comes to state-driven restrictions of fundamental rights, be it freedom of expression or media freedom.

C. The rise of upload-filters in content moderation

13 As mentioned above, the vast amount of data constantly uploaded onto social media platforms makes it almost impossible to manage without the help of technological solutions. Algorithms sort, filter, and prioritise content in order to present what is most relevant for each specific user. In this context, different types of filtering and sorting solutions have been developed. Results may be displayed according to a user's behaviour, his or her location, or his or her self-selected preferences, or simply not displayed because of possible infringements on rights or guidelines. When it comes to technological progress, questions regarding the compliance with freedom of speech proviso arise as artificial intelligence takes over the tasks of content reviewers. Practitioners must be aware of the risks and the opportunities that this development towards a machine-only moderation entails. Taking a closer look at upload-filters will reveal that they are not yet capable of moderating content according to our human rights framework,[27] but could nonetheless be deployed accordingly with further technological improvements.[28]

18 Matthias Cornils, Europäische Menschenrechtskonvention, Art. 10, *BeckOK Informations- und Medienrecht* (2016), para 3.

19 ECtHR, Cholakov v. Bulgaria, 20147/06, para 28.

20 ECtHR, Sik v. Turkey, 53413/11, para 105.

21 Matthias Cornils, Europäische Menschenrechtskonvention, Art. 10, *BeckOK Informations- und Medienrecht* (2016), para 67.

22 Gilert-Hanno Gornig, Äußerungsfreiheit und Informationsfreiheit als Menschenrechte,(1988), 317.

23 ECtHR, Ekin v. FRA, 39288/98, para 58.

24 The ECtHR uses "prior restraints" as a synonym for pre-censorship without fully endorsing the definition in the constitutional jurisprudence of the US Supreme Court, but rather as a "general principle to be applied in this field", see ECtHR, Observer and Guardian v. The United Kingdom, 13585/88, ftn. 6.

25 Christoph Grabenwarter, Katharina Pabel, *Politische und gemeinschaftsbezogene Grundrechte. Europäische Menschenrechtskonvention*, (2016), para 39.

26 ECtHR: Observer/Guardian v. The United Kingdom, 13585/88; Markt Intern Verlag/Beermann v. Germany, 10572/83; Yildirim v. Turkey, 3111/10.

27 Filippo Raso and others., *Artificial Intelligence & Human Rights: Opportunities & Risks* (2018), Berkman-Klein Center for Internet & Society; Viktor Volkmann, 'Hate Speech durch Social Bots' [2018], MMR, 53; Ansgar Koreng, 'Filtersysteme werden nicht lange auf Urheberrechte beschränkt bleiben' [2016], iRights <irights.info /artikel/eu-urheberrecht-content-id-filter/28046> accessed 20 January 2019.

28 Martin Husovec, 'The Promises of Algorithmic Copyright Enforcement: Takedown or Staydown? Which is Superior? And Why?' (2018), 42 Colum. Journal of Law & the Arts, 53, 84.

I. Upload-filters: sorting content before publication

14 In the context of intermediaries, one of the main functions of algorithms is to sort content in a user-oriented way and present it differently depending on a user's profile. When it comes to combating criminal content online in conjunction with algorithmic decisions, the focus is on intelligent filters, such as upload-filters. Upload-filters constitute a subcategory of content-control software.[29] Their function is to recognise certain content, hash[30] it and then - if required - automatically delete it. This means that the entirety of the content uploaded to a platform by its users (user-generated content) is routed through the service provider's cache.[31] Until now, this approach followed a two-step procedure referred to as *notice and take down* (NTD) or *notice and stay down* (NSD), whereas upload-filters act before publication, i.e. while the uploaded content is not yet visible to other users. If a violation is discovered by the filter, the content will not be published at all. Hence, the decision-making process bypasses any human intervention; here, only the filter is doing the work of moderation. The remaining "human in the loop" is the initial programmer of the filter, so, in theory, no additional content moderators will review the content (in contrast to NTD processes that make use of human moderators).

15 One area of application for upload-filters is to search for unlawful content; however, the criterion of illegality is not inherent to the definition of upload-filters because the question of how and what is filtered depends on the initial programming. Beyond that, the system can be self-learning to the extent that, despite small changes to the original content, it still recognizes certain content as a rights or legal violation.[32] Bypassing the mechanism becomes increasingly difficult if the core content is the same. By marking the content as illegal, the filter, through machine learning processes, is trained to recognise it as such and continue to do so further along the process. Upload-filters have been a recurring topic in the discussion on upcoming EU regulation. The two main areas of use are against copyright infringements and terrorist propaganda, which will be examined in the following subsection. Regarding copyright infringements, private companies have already been using filters for a long time. Thanks to its Content-ID-technology,[33] YouTube has been able to identify copyright infringements at a very early stage. The filter was operational as soon as copyright holders had registered their intellectual property (with hashes). YouTube claims that, as of 2016, 99.5% of music claims on YouTube were matched automatically by Content-ID.[34]

II. Use against terrorist propaganda

16 Upload-filters' other area of use is to restrict terrorist propaganda online. Given the increasing risk that social networks and video platforms pose with regards to potential radicalising effects,[35] the EU Commission has proposed a more effective take-down policy for content glorifying violence, especially terrorist propaganda. In 2015, the EU Commission founded the EU Internet Forum which brought together interior ministers of the EU member states, high-ranking representatives of leading companies in the internet industry, Europol, the European Parliament, and the EU Counter-Terrorism Coordinator. The aim was to develop a common approach based on a public-private partnership to detect and combat harmful online content.[36] Against this background, the EU Commission presented its "Code of Conduct on illegal online hate speech"[37] in May 2016 (EU Commission, press release IP/16/1937).[38] The IT companies involved - Facebook, Twitter, YouTube, and Microsoft - committed to take action against illegal hate speech on the internet. Legally speaking, a code of conduct is a so-called "soft law instrument", that is, an agreement on the basis of which companies are bound to the terms, but it has no legislative activity as its basis.[39] The Code of Conduct on illegal online hate speech contains concrete obligations for IT companies, such as verifying the majority of valid reports relating to the

29 <https://en.wikipedia.org/wiki/Content-control_software> accessed 10 December 2018.

30 <https://techterms.com/definition/hash> accessed 10 December 2018.

31 <https://techterms.com/definition/cache> accessed 10 December 2018.

32 Henrike Maier, *Remixe auf Hosting-Plattformen: Eine urheberrechtliche Untersuchung filmischer Remixe zwischen grundrechtsrelevanten Schranken und Inhaltefiltern* (2018), 150.

33 <https://support.google.com/youtube/answer/2797370?hl=en> accessed 10 December 2018.

34 Lyor Cohen, 'Five observations from my time at YouTube' (2017) Official Blog <https://youtube.googleblog.com/2017/08/five-observations-from-my-time-at.html> accessed 10 December 2018.

35 Zeynep Tufekci,'YouTube, the Great Radicalizer', The New York Times, (2018), <https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html?smid=tw-share&referer=https://t.co/aXAthxinwn%3famp=1> accessed 10 December 2018.

36 EU Commission, press release IP/15/6243.

37 'Code of Conduct on Countering Illegal Hate Speech Online' <https://ec.europa.eu/info/files/code-conduct-countering-illegal-hate-speech-online_en> accessed 10 December 2018.

38 <http://europa.eu/rapid/press-release_IP-16-1937_en.htm> accessed 10 December 2018.

39 Michelle Cini, 'The soft law approach: Commission rule-making in the EU's state aid regime', [2001], Journal of European Public Policy, 192, 194.

removal of illegal hate speech in less than 24 hours and removing or blocking access to such content. The first results[40] of the Code's implementation were evaluated in late 2016.

17 In March 2017, the EU Commission introduced the "Database of Hashes", a common database and network developed in collaboration with the four major IT companies who had already agreed to the Code of Conduct. The legal instruments and the technology used for this Database are an exemplary use case for this paper's main argument (which shall be fully elaborated in section D. below). The Database, which is accessible to all participating companies and the intergovernmental authorities mentioned above, collects so-called "hashes" (digital fingerprints) of content that has been marked as "terrorist" or "extremist" by the means of filters. Its purpose is to combat online terrorist propaganda more effectively, that is without the necessity of a human reviewer. But, in so doing this filtering system raises important questions for the exercise of freedom of expression and information.[41] This is mainly due to the "successful" implementation of filtering technology as described above. A few months after the introduction of the Database, representatives of the four IT companies reported that most unwanted content is now deleted before it even goes online. This content includes many videos that are uploaded for the first time and until then not filed with the relevant companies or police authorities and accompanied by a request for deletion.[42] This shows that the Database was fully operational as of late 2017 and contained more than 40,000 hashes for terrorist videos and images.[43] Currently, thirteen companies are associated with the Database which comprised approximately 100.000 hashes by late 2018.[44]

18 YouTube has already been mentioned as an example of a platform that uses filter technologies to prevent copyright infringements. It is also one of the major contributors to the Database of Hashes. This observation is consistent with the assumption that YouTube's recommendation system might lead further down the "rabbit hole of extremism" from video to video,[45] coming to the fore of those working on terrorist propaganda prevention. In an official statement, YouTube explained the use of intelligent filters to combat terrorist propaganda.[46] According to this report, YouTube has removed 7.8 million videos because of their "violative content" from July to September 2018. Through machine learning, it is capable of deleting five times more videos than before. 98% of the videos deleted in 2017 that were related to "violent extremism" were marked by machine-learning algorithms.[47] In this context, YouTube estimates that the human workforce "replaced" by the use of intelligent filters has been 180,000 full-time employees since June 2017. The company also announced its expansion of intelligent filter use to include youth protection and hate speech.

D. Frictions with the notion of censorship

19 The issue with 1) the obligation to use upload-filters to comply with the Code of Conduct, 2) the introduction of the Database, and 3) the collection of data through private companies in a Database accessible to public authorities, is that the distinction between state-driven action and contractual relationships becomes increasingly blurred. When bringing together the human rights framework on freedom of speech including the ban on censorship on the one hand, and the use of upload-filters by private entities such as social media platforms on the other, the question is: is it sufficient to limit our definition of censorship to state-driven action?[48] When public authorities push social media platforms to use upload-filters through "soft law", the effects for the end-user of the platform are identical to when they oblige them to do so by law,[49] because pre-censorship is brought into effect, regardless of the quality of the normative framework used. This phenomenon, referred to as an "invisible handshake", is a contentious one as

40 EU Commission, Code of Conduct on countering illegal hate speech online: First results on implementation, <https://ec.europa.eu/home-affairs/sites/homeaffairs/files/news/docs/first_evaluation_of_the_code_of_conduct_en.pdf>, accessed 15 January 2019.

41 Maryant Fernández Pérez, 'Parliamentarians Encourage Online Platforms to Censor Legal Content', (2017), <https://edri.org/parliamentarians-encourage-online-platforms-to-censor-legal-content/> accessed 15 January 2019.

42 Matthias Monroy,'EU-Internetforum": Viele Inhalte zu „Extremismus"' werden mit Künstlicher Intelligenz aufgespürt', (2017), <https://netzpolitik.org/2017/eu-internetforum-viele-inhalte-zu-extremismus-werden-mit-kuenstlicher-intelligenz-aufgespuert/> accessed 10 December 2018.

43 EU-Commission, press release IP/17/5105, <http://europa.eu/rapid/press-release_IP-17-5105_en.htm> accessed 15 January 2019.

44 EU Commission, Statement/18/6681, <http://europa.eu/rapid/press-release_STATEMENT-18-6681_en.htm> accessed 15 January 2019.

45 Tufekci (n 35).

46 Youtube, Official Blog (2018), <https://youtube.googleblog.com/2018/12/faster-removals-and-tackling-comments.html> accessed 15 January 2019.

47 Youtube, Official Blog (2017), <https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html> accessed 15 January 2019.

48 Jack M. Balkin, 'Old-school/new-school speech regulation' (2013), 127, Harv. L. Rev., 2296.

49 Fernández Pérez (n 41).

it places citizens in an unusual position between private and public law.[50] The difference worth pointing out is that actions taken by virtue of a soft law instrument cannot be appealed in the same way as actions taken by virtue of an administrative act. If decisions related to speech on social media platforms are attributed to community guidelines and not to an act of public authority, the defence capabilities of citizens under that regime will be restricted.

I. Bad filters, good intentions?

20 The analysis above has shown that upload-filters intervene exactly at the point prohibited by the ban on pre-censorship, which is why they are so heavily criticised. But is artificial intelligence really the problem? Should we not summarise the protection afforded by upload-filters as follows: the protection of copyright holders via *Content-ID*, the protection of children via *PhotoDNA*, and the protection of public security from terrorist propaganda via the *Database of Hashes*? Filtering user-generated-content may serve a legitimate purpose (which is why this paper does not aim to question their purposes). Nevertheless, this should not come at the price of unconstitutionality. The intentions behind the use of certain technologies can rarely justify disproportionate rights infringements. This is even more relevant if machine learning is being utilised, as AI amplifies the possibility of losing control over the relevant mechanisms. Today already, the risk of both chilling effects on freedom of expression and collateral censorship is very real when using content-filtering algorithms. In particular, the proportionality of the use of upload-filters is highly doubtful since they operate in a manner that includes a mass and suspicion-independent examination of contents. This is why the use of upload-filters requires more scrutiny when it comes to possible violations of freedom of expression and information.

21 In the case of the German Network Enforcement Act (NetzDG), published reports demonstrated that technology is not yet capable of identifying criminal behaviour in the field of hate speech such as libel and defamation (reports from Facebook, Twitter, Google, YouTube and Change.org available at the German Federal Gazette).[51] Upload-filters still lack the ability to understand content in context or to identify satire in videos,[52] which means that content is often filtered and deleted before being published or made visible to other users even though it might not violate any laws or third-party rights (i.e. legal content). The intermediate conclusion to this section is that the EU impels private companies to use upload-filters which are, technologically speaking, not fit for purpose in meeting the requirements of our common human rights framework.

II. Censorship by whom?

22 Part of the complexity in designing regulation for this field is ingrained into its multi-stakeholder constellation. Instead of structuring a bipolar state-citizen or company-user relationship, communication in digital spaces involves state actors, intermediaries, and users/citizens.[53] We have already established that, in classical constitutional law, we understand "censorship" as the consequence of a state-driven action. However, in the context of online communication, numerous variations have emerged. Censorship *by proxy* is when public authorities control communication or censor it through any number of intermediaries.[54] *Collateral* censorship is when public authorities force intermediaries to control their users' communication.[55] This type of behaviour could be subsumed under the notion of censorship because under FCC jurisprudence, for instance, the internet is considered as a "publicly available source". Withholding information, therefore, interferes with the right to access appropriate information that is required by the general public to inform themselves.[56] Nonetheless, such an action would need to be taken by a *state* entity in order to be classified as censorship, not as content moderation.

23 In relation to the upload-filters used within the Database of Hashes to curtail terrorist propaganda, the question arises as to when might state action be considered an indirect encroachment on fundamental rights if it is implemented by private entities. This question has already been discussed for many years: is it an "unholy alliance" or a necessary cooperation between the state and private intermediaries?[57] Some scholars argue in favour of a more modern concept of state action which also includes private behaviour that can be attributed to the state on the basis of its intention - even if that behaviour is not based on a "hard law"

50 Birnhack, Elkin-Koren (n 14), 49ff.

51 <https://www.bundesanzeiger.de/ebanzwww/wexsservlet> accessed 10 December 2018.

52 YouTube, NetzDG Report 2018 <https://transparencyreport.google.com/netzdg/youtube> accessed 15 January 2019.

53 Jack M. Balkin, 'Free Speech is a Triangle' [2018] Columbia Law Review (forthcoming 2018).

54 Seth F. Kreimer, 'Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weakest Link' (2006), 155, University of Pennsylvania Law Review, 11-100.

55 Balkin (n 48).

56 BVerfGE 103, 44, 60.

57 Birnhack, Elkin-Koren (n 14).

regulatory framework.[58] If a legal implementation of an obligation to filter was to emerge out of the current regulatory propositions,[59] the preconditions for state action could be fulfilled.

III. Sound legal foundation required

24 Censorship functions must not be "outsourced" by the state in such a way that it demands censorship-like action by private actors or provides for corresponding legal obligations or the imposition of negative sanctions in the event of a violation.[60] Using intermediaries to fulfil certain functions on the internet is a collateral way of regulating (online) speech. Although the prohibition of pre-publication censorship is intended to protect freedom of speech and a free flow of information, it might be attractive to public authorities to bypass its protective purpose. Here, a rethink is called for: the vast majority of digital communication spaces are privately owned and therefore not the immediate addressees of the ban on censorship. Limiting the latter to state actors is no longer up-to-date as far as guarantees of freedom of opinion and information are concerned. When pre-censorship (according to the definition elaborated above) is directly based on the initiative of the state (in contrast to *strictly private* content moderation), legal reservations should nevertheless be observed as a barrier to a speech restricting behaviour. Basic legal guarantees such as accountability, transparency, or due process can hardly be ensured when the legal basis for 'voluntary' automated content removal is lacking.[61]

25 A soft law instrument such as a Code of Conduct may offer a certain degree of flexibility and room for manoeuvre, whereas laws take longer to come into force and cannot be adapted as quickly. In line with ECtHR case law, all forms of regulation must be defined by law, they must be in pursuit of a legitimate aim, and they must be necessary.[62] Clearly, soft law can at times serve as an adequate means of regulation but when it comes to restricting human rights, regulation by law is preferable as it fosters transparency and empowers citizens to respond.[63] In his report on the promotion and protection of the right to freedom of opinion and expression for the UN, David Kaye argues that obligations to monitor and rapidly remove user-generated content have increased globally and have established punitive frameworks that are likely to undermine freedom of expression even in democratic societies.[64] As a consequence, states and intergovernmental organisations "should refrain from establishing laws or arrangements that would require the 'proactive' monitoring or filtering of content, which is both inconsistent with the right to privacy and likely to amount to pre-publication censorship".[65] In their study for the Council of Europe, the committee of experts on internet intermediaries came to the same conclusion: "States should not impose a general obligation on internet intermediaries to use automated techniques to monitor information that they transmit, store or give access to, as such monitoring infringes on users' privacy and has a chilling effect on the freedom of expression".[66] This leaves no room for confusion and stipulates very clearly that such collateral censorship mechanisms must be avoided.

IV. Relief through a new EU regulation?

26 In September 2018, the EU Commission presented its proposal for a regulation on preventing the dissemination of terrorist content online,[67] which – in a nutshell – transfers the stipulations from

58 Andreas Voßkuhle, Anna-Bettina Kaiser, 'Der Grundrechtseingriff' [2009], Juristische Schulung, 313; Julian Staben, Markus Oermann, (2013)'Mittelbare Grundrechtsreingriffe durch Abschreckung? – Zur grundrechtlichen Bewertung polizeilicher „Online-Streifen" und „Online-Ermittlungen" in sozialen Netzwerken', Der Staat, 630, 637.

59 EU Commission, press release IP/18/5561, 'State of the Union 2018: Commission proposes new rules to get terrorist content off the web' <http://europa.eu/rapid/press-release_IP-18-5561_en.htm> accessed 15 January 2019.

60 Hoffmann-Riem (n 16), para 94; Bethge (n 11), para 135a.

61 Niva Elkin-Koren, Eldar Haber,'Governance by Proxy: Cyber Challenges to Civil Liberties' (2016), 105, Brooklyn Law Review, 161 f.

62 Council of Europe, 'Ethical Journalism and Human Rights' (2011), Issue Paper commissioned and published by Thomas Hammarberg, Council of Europe Commissioner for Human Rights, CommDH/IssuePaper (2011) 1; Andrew Sharland 'Focus on Article 10 of the ECHR' (2009), 14:1, Judicial Review, 59, 63; Linda Senden, 'Soft Law, Self-Regulation and Co-Regulation in European Law: Where Do They Meet?' (2005), 9.1, Electronic Journal of Comparative Law.

63 Tal Z. Zarsky,'Law and Online Social Networks: Mapping the Challenges and Promises of User-generated Information Flows' [2008], Fordham Intell. Prop. Media & Ent. Law Journal, 741, 780.

64 David Kaye, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, United Nations Human Rights Council, A/HRC/38/35, (2018), 7.

65 ibid 64.

66 Council of Europe, 'Algorithms and human rights', Study on the human rights dimensions of automated data processing techniques and possible regulatory implications, (2018), Committee of experts on internet intermediaries (MSI-NET), 46.

67 EU Commission, COM (2018) 640 final <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018PC0640&from=EN> accessed 16 January 2019.

the Code of Conduct to a regulatory framework. The preamble of the proposal mentions that the regulation aims at increasing "the effectiveness of current measures to detect, identify and remove terrorist content online without encroaching on fundamental rights". These "new rules to get terrorist content off the web within one hour" are supposed to increase the speed and effectiveness of the ongoing "voluntary cooperation in the EU Internet Forum". Art. 6 of the proposal governs the implementation of pro-active measures by service providers, including but not limited to, "detecting, identifying and expeditiously removing or disabling access to terrorist content" in art. 6 (2) b. Here, "pro-active" is used as a synonym for automated removal and/or intelligent technologies. In accordance with art. 6 (1) the hosting service providers are required to implement this type of measure whilst taking into account the "fundamental importance of the freedom of expression and information in an open and democratic society".

27 The proposed regulation could produce relief for the issue outlined in this article. Due to the shift from an "invisible handshake" to a more visible governance by proxy[68] the problems regarding an opaque public-private-partnership could partly be solved. This proposal does, nonetheless, raise other questions regarding the respect of fundamental rights such as (amongst others) the right of "competent authorities" to "request the hosting service provider to take specific additional proactive measures" (art. 6 (3)). This adumbrates the quality of future measures and the usage of artificial intelligence for such purposes.

E. Conclusion

28 We are still unaware of the developments of artificial intelligence in the field of digital communication, and machine learning is – by definition – work in progress. In general, we should refrain from designing too many new, made-to-measure regulations in the field of AI research and implementation. Instead, we should be aware of the constitutional provisos that rule our legal system and think about expanding existing concepts such as the proportionality test. According to these requirements, no state action should be hidden – the alliance of state authority and intermediaries must be transparent and recognisable. We need to clarify the legal basis upon which upload-filters or other types of artificial intelligence are being utilised as part of digital communication processes and services. This need is even more prescient when their effects are forbidden by constitution or by constitutional jurisprudence and when the legal instruments used to regulate them do not meet the requirements of the rule of law. Creating a regulatory framework that renders the "invisible" handshake more visible is unavoidable in a democracy. The proposed regulation for the use case of terrorist propaganda could provide an adequate solution to the problem of the lack of the means of defence: where there is a clear regulatory act, citizens who feel violated in their fundamental rights can respond in a court of law. However, this claim is not only valid for freedom of speech and information issues, but for all fundamental rights that might be restricted by a law enforcement by proxy that exists by virtue of a hidden public agenda.

Acknowledgements

The author thanks Professor Wolfgang Schulz for his valuable feedback, Professor Niva Elkin-Koren for her inspiring and very helpful advice, and the participants of the Young Scholars Workshop on AI at the University of Haifa in December 2018 for their comments.

68 Elkin-Koren, Haber (n 61), 108.