Privatising Law Enforcement in Social Networks: A Comparative Model Analysis

Katharina Kaesling*

Abstract

These days, it appears to be common ground that what is illegal and punishable offline must also be treated as such in online formats. However, the enforcement of laws in the field of hate speech and fake news in social networks faces a number of challenges. Public policy makers increasingly rely on the regu-lation of user generated online content through private entities, i.e. through social networks as intermediaries. With this privat-ization of law enforcement, state actors hand the delicate bal-ancing of (fundamental) rights concerned off to private entities. Different strategies complementing traditional law enforcement mechanisms in Europe will be juxtaposed and analysed with particular regard to their respective incentive structures and conse-quential dangers for the exercise of fundamental rights. Propositions for a recommendable model honouring both pri-vate and public responsibilities will be presented.

1 Introduction: Fake News and Hate Speech on Social Networks

The Internet provides platforms for many forms of speech, with social networks emphasising user-generated content (UGC) like tweets, Facebook posts and Instagram pictures and videos. Digitally expressed 'hate speech' and 'fake news' on social networks have been the topic of public debate worldwide. The term 'fake news' has only recently entered colloquial language. While it is applied in different contexts to characterise political sentiments, manipulation and propaganda, use is made of the term here to describe deliberately false factual claims, i.e. disinformation with no viable basis. False claims are susceptible to be proven either wrong or false, which distinguishes them from opinions.

In that sense, fake news, much like hate speech and defamation, are not new phenomena. However, the particularities of the Internet add a new dimension to them.[1] The Web 2.0, i.e. websites designed to allow easy content creation by end users,[2] facilitates the dissemination of defamatory material. The reach of statements made online in social networks is increased by social media functions like sharing and liking posts. Due to these mechanisms, statements can 'go viral', i.e. trigger a snowball effect. They lead to a quick and global spread at no extra cost for the source. These effects largely lie beyond the control of the statement's creator, though they can be wilfully enhanced by different means including bots.

Hate speech is a political term rather than a legal one. It is not a clear-cut concept; it can encompass incivilities as well as insults and defamation. The specific danger of hate speech lies within the disparagement of a particular group of people. Traditionally, the term 'hate speech' refers to expressions inciting hatred, mainly racial, national or religious in nature.[3] Individuals are offended as members of a group, for example by reason of nationality, gender, race, ethnicity, religion or sexual tendencies. Hate speech has been found particularly worrisome by policy makers as it can stimulate further hatred against these groups. It can greatly influence recipients of such messages depending on the speaker's influence, the message's dissemination and the social and historical context and can be understood as call for action against the targeted groups. While hate can be planted both by illegal and undesirable content, the regulation of UGC, however, has to respect the boundaries of the law. These boundaries define the degree to which the exercise of individual fundamental rights such as free speech is limited in order to safeguard other rights such as the general right of personality.

In recent years, the question has shifted from *whether* to regulate online activities to *how* to do it. While John Perry Barlow proclaimed the independence of cyberspace in his 1996 declaration of the same name,[4] the current prevailing opinion is that illegality offline equals illegality online.[5] Substantive law standards are thus also

* The author is research coordinator at the Center for Advanced Study 'Law as Culture', University of Bonn.

1. D. Cucereanu, *Aspects of Regulating Freedom of Expression on the Internet* (2008), at 7.
2. T. O'Reilly, *What Is Web 2.0 – Design Patterns and Business Models for the Next Generation of Software* (2005), available at https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html.
3. *See* H. Darbishire, Hate Speech: New European Perspectives, *Roma Rights*, No. 4 (1999), available at www.errc.org/article/hate-speech-new-european-perspective/1129; F.M. Lawrence, 'Resolving the Hate Crimes/Hate Speech Paradox: Punishing Bias Crimes and Protecting Racist Speech', 68 *Notre Dame Law Review* 673 (1993).
4. J. P. Barlow, *A Declaration of the Independence of Cyberspace*, Davos, Switzerland, 8 February 1996, available at https://www.eff.org/de/cyberspace-independence.
5. *See* UK House of Commons, 'Hate Crime: Abuse, Hate and Extremism Online', *Fourteenth Report of Session* 2016-17, HC 609, at 11 (2017);

doi: 10.5553/ELR.000115 - ELR december 2018 | No. 3

applicable to online contexts. Nonetheless, an online–offline divide cannot be denied when it comes to the enforcement of substantive law, namely, criminal law provisions, in social networks. The special environments of social networks and the often-invoked borderless nature of the Internet pose massive challenges for an effective law enforcement. Particularities of these environments, principally the relative anonymity of users, the fast dissemination of large volumes of UGC across borders and the global activity of platform operators set significant hurdles.

Social networks were initially rather seen as merely opening new means of communication for users without triggering a responsibility for UGC.[6] Faced with the particularities of the Internet, state actors have increasingly opted to assign responsibility to social networks as intermediaries. Sweden already passed a law to that effect in 1998,[7] while there were no 'precise ideas' on the enforcement of ICT law in Germany, France, the United Kingdom and the United States in 2000.[8] The debates on fake news and hate speech emerged later on and recently invited a number of state interventions worldwide.

In Germany, the 'Act to Improve the Enforcement of Rights on Social Networks' was adopted in 2017.[9] It has gained international attention, as it threatens large fines on social networks that systematically breach their obligations regarding the timely removal of illegal UGC. In the United Kingdom and the Russian Federation, the German law has been cited as model for respective legislative projects. The UK Home Affairs Committee of the House of Commons recommended 'that the Government consult on a system of escalating sanctions to include meaningful fines for social media companies which fail to remove illegal content within a strict timeframe'.[10] The Russian Duma advanced a bill considered 'copy-and-paste of Germany's hate speech law' shortly after its adoption.[11]

In Europe and elsewhere, traditional law enforcement mechanisms are considered inadequate to implement legal provisions in the field of online hate speech and fake news. More and more public policy makers in Europe and elsewhere are contemplating and adopting various additional mechanisms to put the respective laws into effect.

In that context, the German venture appears to show model character, but is it really a good policy example? How does it hold up in comparison with other systems in the European Union? A comparative model analysis will reveal advantages and dangers so as to contribute to the shaping of a superior model for law enforcement in social networks.

Different laws and policy approaches currently in effect in the Europe will be described (2) before turning to the underlying question of delimitating the roles of public and private actors (3). Against that background, three models will be distinguished and evaluated with particular regard to dangers for the exercise of free speech (4). Finally, conclusions and propositions for a recommendable model for law enforcement in social networks honouring both private and public responsibilities will be presented (5).

2 Law Enforcement Strategies in Social Networks

Law enforcement has a servicing function in relation to the substantive law. Traditional law enforcement mechanisms are put into place by the state. More and more, alternatives are considered by policy makers in numerous fields of law.[12] With regard to illegal UGC on social networks, legal norms have been created and policy initiatives launched to complement criminal prosecution and civil law actions. Balkin characterised these informal control measures as new-school speech regulation rather than old-school speech regulation like penalties and injunctions directed at speakers and publishers.[13]

Following a short overview of the legal provisions to be enforced in the context of hate speech and fake news (2.1), the traditional law enforcement strategies of criminal prosecution and civil law actions will be scanned with particular regard to mechanisms to overcome online anonymity (2.2). These laws are complemented by EU law and policy. The elemental legal source within the European Union is the E-Commerce Directive of 2000 (2.3). More recently, the EU Commission has, however, favoured voluntary commitments by social networks (2.4). On a national level, the German and the Swedish regulation will be described (2.5 and 2.6) before briefly summarising the findings (2.7).

2.1 Enforceable Legal Provisions

Online content is illegal when it is contrary to the applicable legal order. In the context of fake news and hate speech, relevant legal provisions are mainly national criminal and civil law affording protection of honour and rights of personality. In addition to criminal prosecution, unlawful statements touching a person's honour, reputation or personality rights generally also trigger the civil liability of the infringer.

B.-J. Koops, 'Cybercrime Legislation in the Netherlands', in P.C. Reich (ed.), *Cybercrime and Security* (2005) 1, at 6.

6. D.M. Boyd and N.B. Ellison, 'Social Network Sites: Definition, History, and Scholarship', 13 *Journal of Computer-Mediated Communication* 210 (2007).
7. *See* 2.6.
8. B.-J. Koops, J. E. J. Prins & H. Hijmans, *ICT Law and Internationalisation* (2000), at 129.
9. *See* 2.5.
10. House of Commons, above n. 5, at 14.
11. Reporters Without Borders, 'Russian Bill is Copy-And-Paste of Germany's Hate Speech Law', published 19 July 2017, available at https://rsf.org/en/news/russian-bill-copy-and-paste-germanys-hate-speech-law.
12. *See* for competition law and ADR, J. Basedow, ‚Rechtsdurchsetzung und Streitbeilegung – Die Vielfalt von Durchsetzungsformen im Lichte von Zielkonflikten', *JZ* 1, at 5 ff. (2018).
13. J.M. Balkin, 'Old-School/New-School Speech Regulation', 127 *Harvard Law Review* 2296, at 2298 (2014).

Despite certain efforts,[14] fake news is not as such illegal in most countries. Regarding both hate speech and fake news, defamation and insult laws are relevant. A number of legal orders foresee a specific criminal provision for cases in which the fact supported by the speaker is false.[15] Prohibited behaviours in the context of hate speech vary widely, also among the Member States of the European Union.[16] International instruments such as the *EU Council Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law*,[17] the *UN Convention on the elimination of all forms of racial discrimination of 21 December 1965* and the *Council of Europe Additional Protocol to the Convention on cybercrime concerning the criminalization of acts of a racist and xenophobic nature committed through computer systems of 28 January 2003* have only had a very limited harmonising effect.

Under the aforementioned EU Framework Decision, hate speech is to be considered a criminal offence when it publicly encourages violence or hatred against a person or group of people because of race, colour, religion, descent or national or ethnic origin. Even so, public incitement to violence is only criminalised in some Member States when its manner is likely to disturb public order or public peace.[18] In addition, varying defamation and insult laws play a considerable role in the fight against hate speech when penalising collective defamation and insults of groups.

In some countries, mainly Common Law countries, a demise of criminal defamation and insult law could be observed.[19] In the context of fake news and hate speech, however, these provisions have (re-)gained importance. Online communication, especially on social networks, has made defamation and insult laws very topical.[20] While rules on the illegality of statements differ, the problem of how to put existing rules to effect in social networks contexts occurs in all legal orders.

2.2 Traditional Public Law Enforcement Mechanisms and Its Limits

Traditional public law enforcement encompasses the criminal prosecution of perpetrators (2.2.1) as well as civil legal protection afforded (2.2.2). In online contexts, their effectiveness is largely called into question by the relative anonymity provided to social network users so that tools helping to overcome that online anonymity are specifically taken into account (2.2.3).

2.2.1 Criminal Prosecution

Criminal prosecution presupposes not only personal jurisdiction over the accused, but generally also his presence at trial, which might prove difficult in international contexts with extradition treaties being limited. Criminal provisions are generally enforced by instituting proceedings in the proper court on behalf of the public. In that case, the public prosecutor somehow learns of potential illegal online activity, investigates *ex officio* and then brings charges. Especially concerning general defamation and insult laws, prosecution presupposes the active involvement of the affected individual. In numerous legal orders, such charges cannot be brought without the victim's consent.[21] Alternatively, victims can act as a private prosecutors, file a criminal suit and prove the relevant facts of the case without the public prosecutor's participation.[22]

The enforcement of general defamation and insult laws is consequently already limited as it largely depends on the victim's authorisation or even legal action. Insofar, law enforcement is left to the victim's discretion. Victims also have the option of choosing civil over criminal action, which might be preferable due to the lighter burden of proof in civil cases.[23] Criminal cases can also be combined with the corresponding civil ones in many legal orders.[24]

2.2.2 Civil Legal Protection

UGC on social networks can also trigger the civil liability of the infringer. In the civil law context, sanctions generally include injunctive relief and damages. Victims of untrue rumours disseminated on social networks, for example, have the demand injunctive relief and revocation from the infringer.[25] This right can be secured by means of interim injunctions. In a social media context, the concerned can thus demand the deletion of tweets, media or short postings. The further dissemination of false information can be prevented by an order to rectify false statements made. In some cases and countries, the victim also has general civil law claims against the platform operator, i.e. the social network provider. For example, under German law, the affected individual can request that the platform operator (temporarily) blocks the account of the infringer in exceptional cases.[26]

Civil (interim) legal protection generally depends on the active intervention of the victims. They have to issue takedown notices or institute civil legal proceedings.

14. *E.g.* U.S. *Honest Ads Bill* of 2017, 115th Congress, 1st session, S. 1989.
15. *E.g.* Germany, Greece and Switzerland.
16. Mandola Intermediate Report – Monitoring and Detecting Online Hate Speech in the Framework of Rights, Equality and Citizenship Programme of the EU Commission of 20 July 2016, at 9, available at http://mandola-project.eu/m/filer_public/7b/8f/7b8f3f88-2270-47ed-8791-8fbfb320b755/mandola-d21.pdf.
17. 2008/913/JHA of November 2008; follow-up to Joint Action 96/443/JHA of 15 July 1996.
18. Mandola Intermediate Report, above n. 16, at 10.
19. *See e.g.* UK Defamation Act 2013 (c 26); for an overview of U.S. States; *see* L.Y. Garfield, 'The Death of Slander', 17 *Columbia Journal of Law & the Arts* 17, at 53-54.
20. *See* for the U.S.A. A. J. Wagner and A. L. Fargo, 'Criminal Libel in the Land of the First Amendment', *Special Report for the International Press Institute*, at 27-28 (2015).
21. *See* S. Griffen, 'Defamation and Insult Laws in the OSCE Region: A Comparative Study', at 10 (2017), available at https://www.osce.org/fom/303181?download=true.
22. *E.g.* Russian Criminal Code Art. 128.1(1); German Criminal Procedural Code Section 374 para. 1, No. 2.
23. Griffen, above n. 21, at 11.
24. *Ibid*. at 10.
25. *E.g.,* German civil code Section 823 para. 1 Civil Code in conjunction with Art. 1 para. 1 and Art. 2 para. 1 Basic Law and Section 1004 Civil Code; Civil Code Section 823 para. 2 in conjunction with criminal law.
26. C. M. Giebel, Zivilrechtlicher Rechtsschutz gegen Cybermobbing in sozialen Netzwerken, *NJW* 977, 980 (2017).

Judicial legal protection can be costly, particularly if multiple jurisdictions are involved as it is likely regarding online UGC.

Victims usually also have a claim for damages if their personality rights were infringed. Damages are supposed to compensate the victim for any harm to his or her reputation or emotional well-being. Their amount differs considerably from legal order to legal order; the incentives for the victim to pursue such a civil legal action vary accordingly.

2.2.3 Mechanisms to Overcome Online Anonymity

The identification of the infringer as potential perpetrator and defendant is crucial for both criminal prosecution and civil legal protection.[27] Social media, however, offers a relative anonymity to its users. Commonly, identity verifications are not required. E-mail addresses are generally needed to register, but can in turn be easily created using false information. IP addresses associated with illegal postings can sometimes, but not always, be traced back to the actual user at the time in question. The anonymity provided is not absolute, as the infringer's identity can also be revealed in the course of investigations going off his social media contacts and information. In many cases, effective legal protection will, however, hinge on mechanisms to overcome that anonymity.

Insofar, the protection of personality rights lags considerably behind intellectual property law. The identification of the infringer can be a question of the applicable substantive or procedural law. By now, a number of legal orders know mechanisms to identify online users hiding behind a pseudonym or commenting anonymously. For example, in Germany, platform operators are now allowed to disclose details about users in cases of insult, defamation, incitement to violence and similar instances.[28] In contrast to copyright law[29] and despite proposals to that effect,[30] there is, however, no specific claim to information in that context.[31] If the applicable substantive law does not provide for a claim for information, there might be procedural court orders available to that end. In the famous UK Internet libel case *Motley Fool*, the service provider was ordered to reveal details about the user posting under a pseudonym under Section 10 of the Contempt of Court Act.[32]

The need for identification of the infringer also affects the ability to quickly move forward with the initiation of judicial protection measures, above all interim legal protection. Its effectiveness is correspondingly tied to the processing time at the competent court, with time being of the essence with the risks of quick uncontrolled proliferation of the personality right violations in online contexts.

2.3 The EU E-Commerce Directive

The basic EU rules on duties of social networks regarding illegal UGC on their platforms were already included in the E-Commerce Directive of 2000 (ECD).[33] The ECD aims to establish a coherent legal framework for the development of electronic commerce within the Single Market.[34] The ECD does not pertain to social networks specifically and concerns all types of illegal content.

Primarily, it regulates the role of information society service providers (ISPs) such as social networks. The ECD distinguishes between three types of services depending on the ISP's activities, i.e. mere conduit (Article 12 ECD), caching (Article 13 ECD) and hosting (Article 14 ECD). Social networks fall under the third category of hosting services, i.e. ISPs that store information by a recipient of the service. These ISPs are not liable for information stored at the request of a recipient on two conditions. Firstly, the ISP may not have actual knowledge of the illegal activity and secondly, the ISP has to act expeditiously to remove or to disable access to the information (Article 14, Recital 46 ECD).

According to Article 15 ECD, Member States shall not impose any obligation to monitor the information that they transmit or store or a general obligation to actively seek facts or circumstances indicating illegal activity on any type of ISP.[35] Member States may, however, establish specific requirements that must be fulfilled expeditiously prior to the removal or disabling of information (Recital 46 ECD) and monitoring obligations in specific cases (Recital 47 ECD). They may require hosting services to apply duties of care that can reasonably be expected from them in order to detect and prevent certain types of illegal activities (Recital 48 ECD). In summary, the ECD only prohibits a general obligation to monitor, while more specific monitoring obligations under national law are permissible.[36] Distinctive features of these two categories remain to be developed.[37]

Legal uncertainty exists regarding the delimitation of the types of ISPs and as to the definition of the relevant terms, such as 'expeditiously', which does not give any specification of a particular time frame in question. Recital 42 ECD clarifies that the exemptions from liability only extend to 'cases where the activity of the information society service provider is limited to the technical process of operating and giving access to a communication network'. It further specifies that this activity is of a mere technical, automatic and passive

27. Cf. R. Perry, and T. Zarsky, Who Should Be Liable for Online Anonymous Defamation?, *University of Chicago Law Review Dialogue* (2015) 162.
28. Section 14 para. 3-5 in conjunction with Section 15 German Teleservices Act and Section 1 III NetzDG.
29. German Copyright Law Section 101.
30. Statement of the German Federal Assembly on the 2nd amending law of the German Teleservices Act of 6 November 2015, BT-Drs. 18/6745.
31. G. Spindler, 'Rechtsdurchsetzung von Persönlichkeitsrechten', *GRUR* 365, at 372 (2018).
32. *Totalise Plc* v. *The Motley Fool Ltd. Anor* [2001] EWHC 706 (QB).
33. Directive 2000/31/EC.
34. EU Commission Press Release, Electronic commerce: Commission proposes legal framework, IP/98/999, Brussels, 18 November 1998.
35. *See also* Recital 47 ECD.
36. *Ibid*.
37. P. Van Eecke, 'Online Service Providers and Liability: A Plea for a Balanced Approach'48 *CMLR* 1455, at 1486-1487 (2011).

nature, thus implying that the ISP has neither knowledge of nor control over the information that is transmitted or stored.[38] Recital 46 spells out that the expeditious removal or disabling of access is in fact a precondition for the limitation of liability. Failing to comply with that obligation, ISPs are not in the safe harbour. The ECD has therefore led to the institution of takedown procedures for social networks.

According to Recital 49 ECD, Member States and the Commission are to encourage the drawing-up of voluntary codes of conduct. In line with this, the Commission has recently presented more targeted approaches aimed at hate speech and fake news.

2.4 EU Hate Speech Code of Conduct and Fake News Initiative

Both with regard to hate speech and to fake news, the EU Commission now works with the biggest social networks towards voluntary commitments without sanctions for non-compliance.

2.4.1 Hate Speech Code of Conduct

In order to combat illegal online hate speech, the European Commission and significant IT companies announced the *Code of Conduct on countering illegal hate speech online* in 2016. This code of conduct was agreed upon by Facebook, Microsoft, Twitter and YouTube. In 2018, Instagram, Google+ and Snapchat also publicly committed to it.[39]

The Hate Speech Code of Conduct relies on the signatory private companies to take the lead, as emphasised by the EU Commission.[40] It does not primarily aim at ensuring compliance with national laws. Social networks firstly test the content against their individual 'Rules or Community guidelines', which have to clarify that the promotion of incitement to violence and hateful conduct is prohibited.[41]

The review of UGC by the participating IT companies is limited to notified posts. Posts can be notified by other users, special 'trusted flaggers' that can use specific channels to alert the social networks and national law enforcement authorities that learned about that content. Upon notification, they examine the request for removal against their rules and community guidelines and where necessary national laws on hate speech transposing the Framework Decision 2008/913/JHA. To that purpose, they set up 'dedicated teams'.[42] The social networks pledged to assess 'the majority of valid notifications' in less than twenty-four hours after notification and remove or disable access to such content, if necessary.[43] Notification of law enforcement authorities and 'trusted flaggers' should be addressed more quickly than others.[44]

In March 2018, the Commission has published an additional *Recommendation on measures to effectively tackle illegal content online*.[45] It reiterates the importance of cooperation of social networks with state actors and further specifies them. Service providers are encouraged to take voluntary proactive measures beyond the notice-and-action mechanisms, including automated means.[46]

2.4.2 Fake News Initiative

In light of the fake information spread on social media in the run-up to the 2016 US presidential election, the European Parliament and Commission are particularly worried about fake news ahead of the 2019 EU election.[47] So far, it has tackled the problem by setting the Fake News Initiative into motion and threatening legislation if social network self-regulation does not prove sufficient. In April 2018, the European Commission gave online platforms the assignment to develop a common Code of Practice on Disinformation by July 2018.[48] This instrument of voluntary public commitment shall be prepared by a multi-stakeholder forum representing not only online platforms, but also the advertising industry and major advertisers. The Commission also urged social networks to promote voluntary online identification systems. A Commission report on the progress made shall be published by December 2018. It will include an evaluation as to whether further (legislative) action is warranted.[49]

The Commission has stressed that proactive measures taken by social networks – as they are encouraged by its fake news initiative – are without prejudice to Article 15 (1) ECD.[50] This also includes 'using automated means in certain cases',[51] which appears to refer to a voluntary monitoring with the help of available filtering and/or research software. According to the Commission, hosting service providers therefore do not risk losing their liability exemption under Article 14 ECD.[52]

2.5 The German Act to Improve the Enforcement of Rights on Social Networks

The recently adopted German Netzwerkdurchsetzungsgesetz (NetzDG) aims to raise the level of protection on social media.[53] The German legislator introduced this Act in 2017 specifically as action against hate speech and fake news following reports about the latter in the

38. Cases C-236/08 – 238/08, *Google France and others* v. *Louis Vuitton and others* [2010] ECR I-02417, Rec. 120.
39. European Commission, Daily News 7 May 2018, MEX/18/3723.
40. Hate Speech Code of Conduct at 2, available at http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=54300; Press release 'European Commission and IT Companies announce Code of Conduct on illegal online hate speech Brussels' of 31 May 2016, IP/16/1937.
41. *Ibid*.
42. *Ibid*.
43. *Ibid*.; European Commission Communication 'Tackling Illegal Content Online. Towards an enhanced responsibility of online platforms', COM (2017) 555 Final, 28 September 2017, para. 13.
44. European Commission, *ibid*.
45. Commission Recommendation of 1 March 2018 on measures to effectively tackle illegal content online (C [2018] 1177 final).
46. *Ibid*, at Rec. (24).
47. EPRS, Online disinformation and the EU's response, PE 620.230 – May 2018.
48. EU Commission Press Release, Tackling online disinformation: Commission proposes an EU-wide Code of Practice, IP/18/3370, Brussels, 26 April 2018.
49. *Ibid*.
50. Above *see* n. 45, Rec. 24.
51. *Ibid.*
52. European Commission, above n. 43, at para. 3.3.
53. R. Schütz, 'Regulierung in der digitalen Medienwelt', *MMR* 36 (2018).

course of the last U.S. Presidential Election. Its name – 'Act to Improve the Enforcement of Rights on Social Networks' – highlights the difficulty the German legislator perceived regarding law enforcement in online contexts and against globally active platform operators that do not have a bricks-and-mortar presence in the state's controlled territory. The NetzDG therefore creates a link to that territory the legislator can control by requiring every social media network to designate a domestic agent as point of contact for public authorities.

The Act did not introduce any new enforceable legal provisions. Instead, illegality within the meaning of the NetzDG is defined by referring to more than twenty criminal law provisions, including defamation and insult, public incitement to crime and hatred as well as propaganda and use of symbols of unconstitutional organisations. In principle, the NetzDG ascertains existing obligations in the framework of the notice-and-takedown procedures as instituted following the ECD. However, it adds further specifications regarding the self-control procedures of social networks and provides for sanctions in case of non-compliance.

The Act sets standards for the social network's complaint mechanism and decision-making. Under the NetzDG, social networks are obligated to institute a procedure for complaints regarding illegal content that allows for a timely deletion. The deadlines for removal depend on the obviousness of the content's illegality. Content that is 'clearly illegal' has to be blocked within twenty-four hours after receiving a complaint. If the illegality is less obvious, the social network has seven days to investigate and delete, with the deadline being extended in case of participation of an 'agency of regulated self-regulation'. These agencies are private outside institutions that were recognised by the Ministry of Justice according to guidelines set out in the NetzDG. Above all, its examiners have to be independent and possess the necessary expertise. Moreover, the agency of regulated self-regulation has to guarantee an examination within seven days and foresee rules of procedure and a complaint mechanism. In case of organisational and systematic failure to comply, social media networks may be fined up to fifty million EUR by the competent public authority. This includes a systematically false decision-making practice, but not a single failure to remove notified illegal UGC. Social networks receiving more than hundred complaints about illegal content in a calendar year are also obliged to publish biannual reports on these complaint procedures.

It is unclear how the NetzDG fits with the ECD.[54] In light of the number of issues, the German legislator at least risked a potential violation of ECD and other EU law principles, most notably the country-of-origin principle as mirrored in Article 3 ECD.[55] The German legislator applied a public policy derogation as criminal offences needed to be respected and the fight against hate speech made regulation necessary.[56] It can also be argued that the NetzDG imposes considerably higher standards on social networks than foreseen by the ECD.[57] While the NetzDG maintains the ECD's general liability and notification system, it sets rather precise deadlines for the deletion of illegal content, which begin with the receipt of the respective complaint.[58] In that regard, the German legislation could possibly exceed the Member States' margin of discretion. Especially in light of these EU law concerns, the NetzDG demonstrates the legislator's determination to combat illegal content like hate speech and fake news more efficiently. The means of choice for the German legislator is – not unlike the EU Commission's more recent approaches – imposing more responsibility on social networks.

2.6 The Swedish Act on Responsibility for Electronic Bulletin Boards

Sweden already regulated illegal content management on 'electronic bulletin boards' in 1998 with the Act on Responsibility for Electronic Bulletin Boards (EBB).[59] According to its Section 1, electronic bulletin boards are services for mediation of electronic messages, i.e. platforms where users can upload data, read news and exchange messages with other users.[60] The Act aims at establishing the provider's responsibility to remove messages that clearly constitute incitement, hate speech, child pornography, unlawful depiction of violence or messages where the posting user manifestly infringes on copyright.[61] The ECD was incorporated by the Act on Electronic Commerce and Information Society Services of 2000.[62]

Under the Swedish regime, owners and providers of Internet-based information services are responsible for illegal content on their systems.[63] UGC considered illegal under the EBB has to be removed by the service provider. According to Section 4 EBB, the service provider has to supervise the service to an extent that is reasonable considering the extent and objective of the service in order to fulfil its obligations to remove or block illegal content under Section 5 EBB. Service providers like social networks thus generally have an obligation to

54. Cf. W. Schulz, 'Regulating Intermediaries to Protect Privacy Online – The Case of the German NetzDG', in M. Albers and I. Sarlet (ed.), *Personality and Data Protection Rights on the Internet* (2018) 1, at 6 et seq., available at https://ssrn.com/abstract=3216572.
55. In support of a violation M. Liesching, 'Die Durchsetzung von Verfassungs- und Europarecht gegen das NetzDG', *MMR* 26, at 29 (2018); G. Spindler, 'Der Regierungsentwurf zum Netzwerkdurchsetzungsgesetz – europarechtswidrig?', *ZUM* 474, at 477 (2017).
56. Art. 3 lit a, no. i. ECD, Bundestag printed matter 18/12356, at 13-4.
57. Spindler, above n. 55, at 478; Liesching, above n. 55, at 29.
58. Section 2 para. 2, No. 2 and 3 NetzDG.
59. Swedish Code of Statutes 1998:112.
60. C. Kirchberger, *Cyber Law in Sweden* (2011), at 35.
61. S. Larsson, 'Metaphors, Law and Digital Phenomena: The Swedish Pirate Bay Court Case', *International Journal of Law and Information Technology* 370 (2013); B.-J. Koops, J. E. J. Prins & H. Hijmans, above n. 8, at 164.
62. Swedish Code of Statutes 2002:562.
63. G. Antonsson and A. Fernlund: Franchising: E-Commerce and Data Protection Issues in Sweden, 4 *Int'l J. Franchising L.* 26, at 26-7 (2006); M. Klang, *The APC European Internet Rights Project, Country Report — Sweden*, available at http://europe.rights.apc.org/c_rpt/sweden.html.

monitor its platforms.[64] Social networks do not fall under the explicit exemptions, as they were introduced to implement the ECD categories of mere conduit and caching.[65]

Removal obligations are limited to specific matters. Relevant illegality under Swedish law is defined in Section 5 with regard to Swedish criminal law provisions on the incitement of rebellion, agitation against a national ethnic group, child pornography crime, and unlawful depiction of violence as well as the infringement of copyrights. An intentional or negligent violation of this obligation is a criminal offence.[66]

Limitations to the general obligation to monitor are set by the law itself, as it stipulates that this obligation is limited to a reasonable extent. Consequently, not all UGC has to be checked under all circumstances. Periodical controls can be sufficient.[67] Service providers like social networks can also make use of notification procedures like user reporting functions and abuse boards, to which users can complain about illegal messages.[68] It is however not sufficient to generally limit the social network's activity to reaction to complaints.[69] How often the provider has to go through the content of the electronic bulletin board depends on the content of the service.[70] In particular, commercial services must check more regularly than private services.[71] For areas where illegal content is common, the provider of the area must check regularly and remove illegal content.[72] Hence, social network providers must maintain a (more) regular control if they learn of illegal UGC.[73]

2.7 Summary

In summary, traditional public law enforcement is increasingly complemented by additional mechanisms largely depending on social networks as intermediaries. These mechanisms range from voluntary self-commitment, code of conducts to negligence liability systems with or without fines to strict liability approaches with an obligation to monitor.

Law enforcement is traditionally seen as state function, albeit relying on the active intervention of the entitled parties. Illegal content is created and disseminated in multilateral constellations involving the infringer and perpetrator, the victim(s), social networks as intermediaries and other users that come into contact with prohibited forms of hate speech and fake news. Within this multi-player context, public and private responsibilities of the actors involved are to be marked down.

64. *See* T. Verbiest, G. Spindler and G.M. Riccio, Study on the Liability of Internet Intermediaries (November 12, 2007), available at http://dx.doi.org/10.2139/ssrn.2575069, p. 109; Klang, above n. 63.
65. Antonsson and Fernlund, above n. 63, at 27.
66. Verbiest et al., above n. 64.
67. *Ibid.*
68. J. Palme, *English Translation of the Swedish Law on Responsibilities for Internet Information Providers*, 3 June 1998, available at https://people.dsv.su.se/~jpalme/society/swedish-bbs-act.html.
69. Klang, above n. 63; Antonsson and Fernlund, above n. 63, at 27.
70. Palme, above n. 68, Comment to Art. 4.
71. *Ibid.*
72. *Ibid.*
73. Klang, above n. 63.

3 Private and Public Responsibilities

The Internet is governed by multiple, overlapping modalities including social norms, code, market and the law. Social media companies serve as intermediaries, who supply the environment enabling users to create and access UGC. Naturally, they are not public utilities, but private entities carrying out a business endeavour. While they are thus prone to implement market-oriented business strategies, it is the public policy makers' task to adequately safeguard the exercise of fundamental rights. At the same time, the individual social media user voluntarily joins and frequents social networks according to his habits. The task of preventing and combating hate speech and fake news could be attributed to all three groups of actors – social media users, social networks and public policy makers.[74]

Could social media users not simply be trusted to make their own choices, thus making any intervention from the other two actors expendable (3.1)? Why should law enforcement not be largely delegated to social network providers (3.2) and what are public non-disposable core responsibilities (3.3)? These questions will be answered in order to pave the way for a comparative model analysis against that background (4).

3.1 User Self-Censorship

It has been argued that commercially available filtering software can be applied by users to block sites on the basis of content, thus making (additional) governmental regulation unnecessary.[75] Individual users can customise these filters in accordance with their moral and social attitudes and by this means control their receptions.[76] Rather than a censorship by the state, users only censor themselves. Technological tools that allow the blocking of sites on the basis of content were especially developed to shield children from inappropriate content.[77] Shortcomings of these tools have however also been identified.[78] Like all technological tools, further development can certainly improve the overall software quality.

Even with enhanced technological tools, factual limits of hate speech would, however, be placed in hands of commercial interests.[79] Moreover, with the referral to commercially available filtering devices, hate speech remains accessible to all those that did not install adequate filtering software. The socially destabilising force of hate

74. Cf. for a new structure of speech regulation J.M. Balkin, 'Free Speech is a Triangle', Colum. L. Rev. 1, at 4 et seq (forthcoming 2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3186205.
75. R. Weintraub-Reiter, 'Hate Speech Over the Internet: A Traditional Constitutional Analysis or a New Cyber Constitution?', 8 *Boston University Public Interest Law Journal* 145, at 173 (1998).
76. *Ibid*.
77. *E.g.* CyberPatrol, NetNanny, SurfWatch, HateFilter.
78. M. Krantz, 'Censor's Sensibility: Are Web Filters Valuable Watchdogs or Just New Online Thought Police?', *Time Magazine*, 11 August 1997, 48.
79. *See* A. Tsesis, 'Hate in Cyberspace: Regulating Hate Speech on the Internet', 38 *San Diego Law Review* 817, at 867 (2001).

speech is therefore not fended off.[80] Compliance with laws lies in the discretion of users, thereby circumventing the ratio of hate speech laws. Particularly with regard to content whose illegality stems from incitement to violence, the danger lies primarily in reaching out to those recipients who might not be interested in blocking that very illegal content. Self-censorship by users is therefore clearly insufficient for protecting societal welfare and the individual rights at stake.[81]

3.2 Law Enforcement in the Hands of Social Networks – Why Not?

Social networks operate platforms for social traffic online. By creating these environments, they do not only render communication between users possible, but also shape it according to the platform design. Unlike telephone landlines, they do not only make a means of communication between a small number of communicators possible for a monetary consideration.[82] The success of business models of Facebook, Twitter, Instagram and the like is based on the creation of UGC in large volumes and at fast publishing rates. Social media has a magnifying effect for all ideas and opinions expressed, while at the same time offering a (relative) anonymity to the user creating content. It also favours the creation and organisation of groups on national and international levels, including extreme movements prone to generate illegal content.[83] Hence, the facilitation of the spread of illegal content is provoked by the business model itself. Why then not simply give the responsibility for the lawfulness of content to the social media operators?

As platform operators, these social networks like Twitter, Facebook and Instagram create the environment for user statements and naturally govern social interaction on their platforms. They certainly are in a good position to carry out control[84] – arguably in a better position than state regulators.[85] Social networks therefore seem to be the point of least cost at first glance.[86] Unlike national state governments, social networks are able to set rules on all markets they are active on (3.2.1). In addition, they can make more effective use of technology (3.2.2).[87] However, both of these apparent advantages are limited by practical considerations. Moreover, the application and interpretation of relevant (criminal) provisions by social networks hold considerable dangers for fundamental rights and thus (democratic) societies (3.2.3).

3.2.1 Social Media Policies and Terms of Use

Social media operators have long had policies against the use of hate speech as part of their corporate responsibility,[88] i.e. by reserving themselves the right to revoke accounts that are against their hate speech policy. They have contracts with their users and can unilaterally impose terms of use for their worldwide operations. These contracts generally contain provisions prohibiting users from creating content in violation of the law, especially defamatory, harassing, hateful, or racially or ethnically offensive content. For example, Facebook's terms of use require the users not to bully, intimidate or harass any user and not to post content that is hate speech, threatening, or pornographic; incites violence; or contains graphic or gratuitous violence.[89]

Such terms of use are, however, not effective, not even for the purposes of deterrence. Social media terms of use, much like any small print, are hardly actually read by the end users who manifest their consent merely by clicking a button in a pop-up window or a dialogue box. It thus cannot even be assumed that the terms of use create awareness amongst the users.[90] Besides, users willing to violate criminal law are likely to be willing to violate terms of use as well.

3.2.2 Use of Technology

Social networks could implement technological tools for the detection and blocking of hate speech and fake news more effectively than external state actors.[91] However, not only state actors, but also social networks are confronted with the large volume and high rate of publication of UGC. This renders thorough monitoring of content fairly difficult for those intermediaries as well.[92]

With regard to copyright infringements, filtering mechanisms employing digital fingerprinting, i.e. matching uploaded and protected works, have been successfully employed on a voluntary basis for years. The software 'Content ID' has been used by YouTube and Facebook to filter illegal extremist content.[93] Only after clearly extremist content has been identified, can a hash be created in order to compare this content via digital fingerprinting. While other filtering devices are tested, there is currently no appropriate technology that allows for an effective monitoring for illegal hate speech and fake news. Striking a fair balance between fundamental rights affected in specific cases at hand is not easily programmed.[94] Filtering tools would also have to take into account the specificities of the jurisdictions concerned.

80. Tsesis, above n. 79, at 866.
81. *Ibid*.
82. Cf. *See* T. Gillespie, 'Regulation of and by Platforms,' in J. Burgess, A. Marwick, T. Poell (eds.), *The Sage Handbook of Social Media* (2017), at 257-8; T. Gillespie, 'Platforms are Not Intermediaries', 2 *Georgetown Law Technology Review* 198 (2018).
83. *See* B. Perry and P. Olsson, 'Cyberhate: The Globalization of Hate', 18 *Information and Communications Technology Law* 185 (2009).
84. C.E. George and J. Scerri, 'Web 2.0 and User-Generated Content: Legal Challenges in the New Frontier', 2 *Journal of Information, Law and Technology* 1, at 10 (2007).
85. *Ibid*. at 18.
86. *See* Rec. 59 Copyright in the Information Society Directive.
87. George and Scerri, above n. 84.
88. K. Klonick, 'The New Governors: The People, Rules and Processes Governing Online Speech', 131 *Harvard Law Review* 1598, at 1626 (2018).
89. Facebook Terms of Use U.S., retrieved from Germany, available at https://www.facebook.com/terms.php (last visited 18 June 2018), Section 3.
90. George and Scerri, above n. 84, at 12.
91. *Ibid*., at 18.
92. *Ibid*., at 10.
93. O. Solon, 'Facebook, Twitter, Google and Microsoft Team up to Tackle Extremist Content', *The Guardian*, 6 December 2016.
94. *See* D. Burk and J. Cohen, 'Fair Use Infrastructure for Copyright Management Systems', *Georgetown Public Law Research Paper* (2000) 239731/2000 for 'fair use' in copyright law.

3.2.3 *Dangers of the Application and Interpretation of the Law by Private Entities*

Assigning responsibility for the lawfulness of UGC to social networks as intermediaries involves the application and interpretation of the relevant, mainly criminal, provisions. It is then left up to social media operators as private entities to draw the oftentimes thin line between legitimate exercise of the right to free speech and criminal conduct. This is namely – but not exclusively – due to the underlying importance of constitutional law. The interpretation of criminal legal norms safeguarding personal honour and dignity against factual claims, opinions and incitement to hatred and violence as well as their application to the individual case is strongly shaped by fundamental right considerations. The application and interpretation of the provisions of criminal law are to be carried out in light of the affected fundamental rights,[95] such as the protection of personal honour as part of the general right of personality, freedom of speech and expression and potentially artistic freedom.

For example, in German law, there is no general precedence of one over the other, which makes the determination of a statement's legality – both online and offline – particularly challenging. The German Federal Constitutional Court has underlined the general principle that certain contents of statements, especially regarding political views, shall not be sanctioned.[96] Nonetheless, there are limits to freedom of speech and freedom of the media, such as restrictions inherent in other fundamental rights, especially human dignity.[97] The German constitutional jurisprudence on that matter shows that sufficient consideration of freedom of speech and expression has proven consistently difficult even for judges of the ordinary jurisdiction with the federal constitutional court repeatedly overturning judgements.[98] The ECHR has also consistently stressed the overriding and essential nature of freedom of expression in a democratic society, while at the same time accepting and setting limits in case of incitement to hatred, discrimination and violence.[99] If the application and interpretation of relevant provisions is carried out by the competent state courts, a constitutional review by the competent constitutional authorities is secured. If these tasks are handed off to social networks, the participation of the concerned before the decision on the removal depends on the social networks' good will.

3.2.4 *Social Networks as Rational Market Players*

'Services have a moral duty to fight illegal behaviour online', David Cameron is quoted as stating in the context of child pornography.[100] Surely, the question of moral responsibility of social networks becomes more and more pressing in light of their developing role in society. It is linked to a number of ethical issues regarding both users and network administrators.[101] Notwithstanding that worthy discussion, social networks as private commercial entities do not serve public policy purposes or other altruistic interests. They are not directly bound by fundamental rights and by no means guardians of their protection. Reliance on private entities 'relegates governmental duties to private prejudices, incentives and priorities'.[102]

When evaluating law enforcement strategies, social networks have to be seen as rational market players acting in accordance with their interests. For example, obscene and violent material can negatively impact advertising revenue.[103] The platform operators' principal aim as businesses is economic gain. Hence, the incentive structure created for these economic social networks has to be analysed in order to determine the consequences for fundamental right protection. The premise in this context is that social networks will act to prevent the dissemination of illegal, but not legal content, if this outcome is in line with its own interests like the maximisation of profits and the reduction of risks. This is especially true as most cases of illegal content are not as easily identifiable and not as severe as child pornography. In less severe cases, both the moral scruples and the public relations issues are weaker and so are the incentives to combat such illegal content.

3.2.5 *Conclusions*

With all models delegating the responsibility of legal tests to the private entities that social networks are, the application and interpretation of legal norms is left to them and their agenda, even though this is an essential state task. As has been shown for Germany in an exemplary manner, this application and interpretation in consideration of the fundamental rights at stake is a rather complicated task that regularly leads to the repeal of judgements by constitutional bodies. As the applicable legal sources vary, the decentralised character of worldwide social networks is no advantage.

Social networks certainly have the potential to be technical chokepoints in the fight against hate speech and fake news. Social media policies and terms of use are, however, not effective tools to ensure the legality of UGC. While networks are well placed to implement detection technology and filtering devices, no such software currently exists with regard to hate speech and fake news. The successful technique of digital footprinting can only be used with regard to certain, severe cases. Generally, the determination as illegal presupposes the consideration of the context in the individual case and the affected fundamental rights. The application and interpretation of relevant provisions by social networks therefore hold risks for the exercise of fundamental

95. For Germany *see* Federal Constitutional Court, *NJW* 2943(1994); *NJW* 3303 (1995); D. Grimm, ‚Die Meinungsfreiheit in der Rechtsprechung des – Bundesverfassungsgerichts', *NJW* 1697, at 1701-02 (1995).
96. Federal Constitutional Court, *NJW* 257, 258 f. (1958).
97. Federal Constitutional Court, *NJW* 1303 (2003).
98. Federal Constitutional Court, *NJW* 2022 (2015); *NJW* 2643 (2016); *NJW* 1092 (2017).
99. *Belkacem* v. *Belgium*, Application no. 34367/14, ECHR, 27 June 2017.
100. R. Watts, 'David Cameron: Web Firms Have a "Moral Duty" to Wipe Out Indecent Images', *The Telegraph*, 20 July 2013.
101. M. Turculeţ, 'Ethical Issues Concerning Online Social Networks', 149 *Procedia, Social and Behavioral Sciences* 967 (2014).
102. Tsesis, above n. 79, at 868.
103. Klonick, above n. 88, at 1627.

rights. Their realisation depends on the particular law enforcement model in place and will thus be further examined below with regard to the models compared.[104] The protection of fundamental rights is not a task of private economic entities, but one of the public core responsibilities.

3.3 Public Core Responsibilities

State actors are the guardians of fundamental rights. They are bound by law to respect and safeguard fundamental rights. Hence, they cannot comprehensively delegate this underlying responsibility to private actors, as the enforcement of existing law by the state is the necessary counterpart to the state monopoly on the legitimate use of force. The restriction of the rights of the individuals depends on the state's empowerment with enforcement rights. When the state entrusts private actors with the enforcement of the law, its delegating mechanisms are to be analysed with regard to the fundamental rights ramifications and the effectiveness of enforcement. This is also true when responsibility is assigned to social networks and public policy is implemented by shaping their incentive structure.

4 Comparative Model Analysis against That Background

The spectre of possibilities to safeguard social networks against hate speech and fake news in addition to the traditional law enforcement mechanisms covers many different approaches. It ranges from a laissez-faire approach with user self-censorship all the way to active monitoring obligations of social networks. Naturally, there is a continuum between these different strategies; legal regimes like the ones referenced above[105] can fall anywhere along that continuum. Given the private and public core responsibilities identified above,[106] different models will be juxtaposed and evaluated.

Self-censorship has already been dismissed, as it does not effectively serve the purpose of fighting the dissemination of hate speech and fake news.[107] Keeping in mind the reservations regarding the delegation of law enforcement to social networks,[108] different schemes of network responsibility will be examined. For that purpose, three basic models shall be distinguished, namely a strict liability approach with an obligation to monitor (4.1), a negligence-based liability system with a notice-and-takedown mechanism (4.2) and voluntary commitments of social networks to code of conducts and the like (4.3).

4.1 Obligation to Monitor and Strict Liability

Proactive monitoring obligations are generally and increasingly used to impose a strict liability standard on Internet intermediaries such as social networks.[109] With a strict liability approach, social networks are held responsible for illegal content on their platforms even if they did not have any knowledge of the content concerned. The legal doctrine of strict liability makes a person or company responsible regardless of any negligence or fault on their part. It is conventionally applied when such persons engage in inherently dangerous activities. This can be said with regard to social networks as the business models they profit of favour the creation of (illegal) UGC.[110]

Obligations to monitor the UGC establish such a strict liability regime.[111] Compliance with general monitoring obligations proves tremendously difficult in light of the insufficient technical tools.[112] Smaller social networks and start-ups are pushed out due to the high operating costs related to the shielding against risks, thus cementing the market.[113] Innovation and competition are thus hindered by this strict approach, with economic exchange online not being furthered.[114]

This model of law enforcement in social networks creates strong incentives for social networks to block all potentially illegal content in order to avoid any liability. Content carrying the risk of provoking controversy is thus likely taken down pre-emptively or at the first complaint received. There is no significant economic advantage to hosting debatable UGC. Decisions are therefore not primarily made on the legality of the content. Content will readily be removed or blocked before any court involvement. Individual incentives for interventions vary largely and are not sufficient to safeguard the fundamental rights concerned.[115] There is also no incentive for social networks to carry out factual investigations first. This is all the more significant as illegality of UGC in the context of hate speech and fake news is only rarely evident.[116] Accordingly, monitoring obligations lead to incentives to overblock.[117] For that reason, the OSCE Special Rapporteurs on Freedom of Expression spoke out against the imposition of duties to monitor the legality of the activity taking place within the intermediaries' services.[118]

104. *See* 4.
105. *See* 2.2.-2.7.
106. *See* 3.
107. *See* 3.1.
108. *See* 3.2.
109. B. Kleinschmidt, 'An International Comparison of ISP's Liabilities for Unlawful Third Party Content', 18 *IJLIT* 332, at 346 (2010); P. Baistrocchi, 'Liability of Intermediary Service Providers in the EU Directive on Electronic Commerce', 19 *Santa Clara High Tech. L.J.* 111, at 114 (2002).
110. *See* 3.2.
111. Baistrocchi, above n. 109.
112. *See* 3.2.2.
113. *See* Baistrocchi, above n. 109; J. Hornik and C. Villa, 'An Economic Analysis of Liability of Hosting Services: Uncertainty and Incentives Online', 37 *Bruges European Economic Research Papers* 13 (2017) for all ISPs under the ECD.
114. *Ibid.*
115. *See* 2.2.
116. *See* 3.2.3.
117. *Delfi AS* v. *Estonia*, Application no. 64569/09, ECHR, 16 June 2015, Joint Dissenting Opinion of Judges Sajò and Tsotsoria § I.2.
118. Joint Declaration of the Three Special Rapporteurs for Freedom of Expression (2011) 2.b, available at www.oas.org/en/iachr/expression/showarticle.asp?artID=848.

The danger of overblocking leads to chilling effects for the exercise of fundamental rights.[119] Social networks are deterred from hosting content in legal grey areas, and users are discouraged from exercising their fundamental rights such as free speech on social networks in light of the expected quick removal of controversial content. Free speech and potentially also artistic freedom and freedom of the media are most restricted in strict liability systems with an obligation to monitor. This model represents a case of 'collateral censorship' that occurs when the state holds a private party – the social networks – liable for the speech of another private party – the user generating content – and the first private party also has the power to control access to B's speech.[120]

Swedish law foresees an obligation to monitor.[121] However, the social networks' duty to supervise under Swedish law is considerably relativised. Networks do not have to guarantee that their systems are clean.[122] The proactive duty to check for illegal content is limited to areas where UGC is more likely to occur on the basis of past experiences or context. For other areas, a notification system can be sufficient. The Swedish system thus combines the first model of an obligation to monitor with the second model of a notification system.

4.2 Notice-and-Takedown and Negligence-Based Liability Systems

The second model can be described as conditional safe harbour model. Social networks are protected in the safe harbour as long as they comply with the requirements for dealing with unlawful content on their platforms.

With that model, social networks as such have no general monitoring obligation. Their liability for illegal content disseminated via their facilities is limited. It depends on knowledge of the illegal content in question and compliance with duties to take down that content. Examples for notice-and-takedown and negligence-based liability systems are the ECD and the German NetzDG. According to both the ECD and the NetzDG, social network liability is excluded if upon obtaining knowledge or awareness of illegal content, the social media provider acts expeditiously to remove or to disable access to the information, with the German system defining more clear-cut deadlines than does the EU one.[123] Such systems based on knowledge or notice of illegal content mirror the lack of adequate monitoring software.

The rather nebulous concept of expeditious acting, however, risks blurring the lines of the social network's responsibility. In terms of legal certainty, the German model appears to be favourable at first sight as it clearly stipulates deadlines for the takedown. While it appears sensible to tie these deadlines to the time needed to properly assess the illegality of the content, the gradations according to the obviousness of illegality reintroduce elements of legal uncertainty and unpredictability. As a result of legal uncertainty, it is difficult for social networks to weigh how much to invest in the prevention of the publication of illegal UGC on their networks.[124] Legal uncertainty affects the social network's ability to determine a rational investment and an efficient targeted line of attack.[125]

Notice-and-takedown systems can protect the exercise of fundamental rights inasmuch as they drive social networks to actually test the legality of the content before removing or blocking it. Neither the ECD nor the NetzDG foresee specific mechanisms to ensure the test of legality; the tiered deadlines for removal, however, give room for adequate examination.

The option of involving a private outside institution (agency of regulated self-regulation) provided by the German NetzDG does not guarantee correct rulings on the legality of UGC. Even though the examiners' expertise has to be recognised by the Federal Office of Justice, they are part of a private institution offering their services to social networks. As such, their incentives are approximated to those of their clients. There is thus little to no[126] added value in comparison with mere in-house assessments by skilled jurists. Complaint mechanisms are confined to the agency; there is no integration into ordinary jurisdiction. Court reviews are – as with all decisions taken by social networks – limited to the period after the fact, i.e. the removal.

In contrast to the first model with a general obligation to monitor, the incentives to swiftly remove all questionable content are limited to the notified UGC with notice-and-takedown and negligence-based liability systems. They still entail dangers for fundamental rights with regard to the notified content because they cause incentives to overblock as well as considerable chilling effects.[127] These incentives are enhanced by the threat of considerable fines in the NetzDG. Social networks will readily remove content in order to minimise their risks, especially towards the end of the standardised deadlines. The NetzDG has therefore been described as bold gambit with fundamental rights.[128]

This danger is reduced, but far from banned by the limitation of fines to systematic failure rather than to the non-compliance in individual cases by the NetzDG. Standardised deletion upon call minimises the risks to

119. W. Seltzer, 'Free Speech Unmoored in Copyright's Safe Harbor: Chilling Effects of the DMCA on the First Amendment', *Harv J L &Tech* 171, at 175-6 (2010).
120. Balkin, above n. 13, at 2309.
121. *See* 2.6.
122. Koops, Prins & Hijmans, above n. 8, at 165.
123. *See* 2.3 and 2.5.
124. Relying *inter alia* on deterrence theory, Hornik and Villa, above n. 113, at 6.
125. *Ibid*., at 11.
126. M. Liesching, '§ 3', in Erbs/Kohlhaas/Liesching, *NetzDG* (2018), at § 3 Rec. 23.
127. J. Urban and L. Quilter, 'Efficient Process or "Chilling Effects"? Takedown Notices, Under Section 512 of the Digital Millennium Copyright Act', 22 *Santa Clara Tech. L. J.* 621 (2006).
128. E. Douek, 'Germany's Bold Gambit to Prevent Online Hate Crimes and Fake News Takes Effect', published 31 October 2017, available at https://www.lawfareblog.com/germanys-bold-gambit-prevent-online-hate-crimes-and-fake-news-takes-effect.

be fined or prosecuted.[129] The danger actually manifested itself only ninety-six hours after the NetzDG's entry into force, when Twitter blocked the account of a German satirical magazine. The magazine had parodied a far-right politician whose social media accounts were blocked earlier that week due to anti-Muslim posts.[130] There are no data as to how much legal content has been removed and how much illegal content kept.[131] Consequently, the proportionality of measures like the German notification and fining system is hard to assess because of the lack of (reliable) data. According to press reports, Facebook performed 100,000 deletions in Germany in the month of August 2016 alone.[132] Data pertaining to the removal of copyright infringing content support an over-removal of content by Internet hosting providers under a notice-and-takedown system.[133]

4.3 Voluntary Commitments – Code of Conducts and the Like

Voluntary commitments to comply with a code of conduct appear like paper tigers, especially against strict liability or negligence-based systems with severe fines for non-compliance. It must, however, not be forgotten that every deletion of a legal upload, post or tweet violates freedom of speech and expression and possibly also freedom of the media and other fundamental rights. According to the third evaluation of the EU Hate Speech Code of Conduct, whose results were published in January 2018, the signatory IT companies removed on average 70 per cent of illegal hate speech notified to them by non-governmental organisations (NGOs) and public bodies participating in the evaluation.[134] For that reason, EU Commissioner for Justice, Consumers and Gender Equality Jourová found the code of conduct a valuable tool to tackle illegal content quickly and efficiently.[135] The European Commission expressed its conviction that the code of conduct will not lead to censorship, as it does not oblige the signatory companies to take down content that does not count as illegal hate speech.[136]

Against that background, it needs to be reiterated that none of the models and examples presented obliges social networks to take down legal content. As long as non-compliance with voluntary commitments does not lead to any liability or sanction, there is certainly less incentive to overblock than with strict or negligence-based liability systems. Nonetheless, considerable incentives to delete not only illegal but also legal content exist.

As social networks firstly test the content against their individual 'Rules or Community guidelines' according to the Code of Conduct, restrictions on free speech and other fundamental rights are detached from legal prerequisites. The code does consequently not safeguard existing laws that strive to balance free speech and rights of third parties.[137] If policies are significantly stricter than the applicable state law, free speech is unduly limited by deleting legal, albeit undesirable, statements. With social networks increasingly under fire for hate speech and fake news dissemination on their platforms, there is substantial public pressure to act. They can document their efforts with a media-effective signature of a code of conduct and the publication of the percentage of quickly deleted notified content. The figures published by social networks have been recently taken into account by numerous state actors.[138] They also play a substantial role for the businesses' public image. When the image of a company like Facebook or Twitter suffers, this can easily translate to financial loss.

Voluntary commitments gradually include more and more proactive duties.[139] The ECD principle that there is no general obligation to monitor is called into question by the voluntary frameworks set up at EU level. The Commission explicitly demands proactive monitoring: 'Online platforms should, in light of their central role and capabilities and their associated responsibilities, adopt effective proactive measures to detect and remove illegal content online and not only limit themselves to reacting to notices which they receive.'[140] This imposes de facto monitoring obligations[141] with the corresponding dangers for the exercise of fundamental rights. These duties clearly surpass the scope of a notice-and-takedown system, as they also apply to non-notified content. With regard to notified UGC, an overreliance on trusted flaggers is to be feared. Social networks must not refrain from any legal test in cases of notifications from this group of users and institutions.

The encouragement to proactively deploy filtering devices, for example, by the EU fake news initiative, also holds risks for a lawful application of relevant provisions. Fully automated deletion or suspension of content can be particularly effective and deserves support in circumstances that leave little doubt about the illegality of the material, for example, in cases of child pornography. Filtering without an additional case-by-case review equals deletion without any legal test and is therefore

129. Liesching, above n. 55, at 30.
130. Titanic Magazin, 'Twitter sperrt TITANIC wegen Beatrix-von-Storch-Parodie', 3 January 2018, available at www.titanic-magazin.de/news/twitter-sperrt-titanic-wegen-beatrix-von-storch-parodie-9376/.
131. German Federal Ministry of Justice and Consumer Protection, Answer to Written question from André Hunko, No. 10/19 of 6 October 2016, at 1.
132. Zeit Online, 'Facebook nennt erstmals Zahl entfernter Hasskommentare', 26 September 2016, available at https://www.zeit.de/digital/2016-09/hasskommentare-facebook-heiko-maas-richard-allan.
133. A. Marsoof, 'Notice and Takedown: A Copyright Perspective', 5 *Queen Mary J. of Intell. Prop*. 183 (2015); D. Seng, 'The State of the Discordant Union: An Empirical Analysis of DMCA Takedown Notices', 18 *Va. J. L. & Tech.* 369 (2014).
134. Press release, 'Countering Illegal Hate Speech Online – Commission Initiative Shows Continued Improvement, Further Platforms Join' of 19 January 2018, IP/18/261.
135. *Ibid.*
136. Tweet @EU_Justice, Twitter, 07:11 – 19 January 2018.
137. *See* EDRi. 'EDRi and Access Now Withdraw from the EU Commission IT Forum Discussions', *EDRi*, 16 May 2016.
138. *See* 4.2 and 4.3; for the UK *see* House of Commons, above n. 5, at 13.
139. *See* for the fake news initiative 2.4.2.
140. European Commission, above n. 43, at para. 10.
141. G. Frosio, 'The Death of 'No Monitoring Obligations', *CEIPI Research Paper* No. 2017-15, 1, at 25 (2017).

hazardous to the exercise of fundamental rights.[142] Filtering especially leads to significant chilling effects.[143] While voluntary commitments might be paper tigers with regard to their enforcement against the social network's will, they show their teeth when it comes to the endangerment of fundamental right exercise online.

4.4 Conclusions

Social networks have gained a considerable amount of control in areas with high relevance for the enjoyment of fundamental rights like free speech and right of personality. With regard to different models of social network responsibilities, it has been shown that all three of them harbour risks for the safeguard of fundamental rights concerned, especially for the exercise of free speech. All these models delegate the application and interpretation to private entities to an extent endangering the lawful interpretation and application of the criminal provisions penalising hate speech and fake news.

The German negligence-based liability system cannot be recommended as international policy example because of these dangers flowing from its incentive structure. For the same reason, an obligation of social networks to autonomously monitor the content on their platforms has to be dismissed. Voluntary commitments to code of conducts can help integrate social networks in the fight against fake news and hate speech, but not without creating an – albeit mitigated – incentive to overblock UGC.

5 Shaping a Superior Model for Law Enforcement in Social Networks

So far, public policy makers in Europe have largely reacted to the challenge of regulating UGC on social networks with far-reaching delegations of law enforcement to social networks. Territorial governments should realise their regulating potential (5.1). In light of all the above, a multi-player solution with stronger public engagement is favoured (5.2).

5.1 Regulating Potential of Public Policy Makers

The somewhat extraspatial character of the Internet does not mean that online activities shall remain unencumbered by government regulations. The approach that cyberspace 'exists, in effect everywhere, nowhere in particular and only in the Net'[144] and that the Internet is 'not subject to the same laws of reality as any other electromagnetic process'[145] is outdated. Geographically based governmental authority is not inapplicable because of a certain non-physical nature of 'the Internet'. Transmission of online content occurs through physical processes in specific jurisdictions by means of physical infrastructure and processes. It has effect on 'real people and real places'.[146]

Online activities indeed make jurisdictional limits visible, as online content is generally accessible beyond borders. States have personal jurisdiction over Internet users depending on their situation.[147] In case of cross-border crimes, more than one jurisdiction can apply to a single act. While the applicable private law can be determined by the rules on conflict of laws and international civil jurisdiction is established in accordance with international procedural law, this is a significant challenge in practice. Therefore, a further harmonisation and unification of law and policy both in the area of private international law and in the area of substantive laws on fake news, hate speech and other defamatory and illegal content would greatly benefit effective traditional law enforcement.[148] In that context, the level of (international) regulation has to be chosen with particular regard to legal cultures in the participating states, especially the concept of free speech and its limits.

The development of online communication through social media, which has *inter alia* physical, psychological, and cultural effects,[149] brings about major changes for law enforcement. State actors both on national and EU level have extensively criticised social networks for their failure to effectively address fake news, hate speech and defamatory UGC. In spite of this, the adjustment of the law to such major changes is a governmental task rather than a private one. State actors have to meet the regulatory needs created and (re-)evaluate law enforcement strategies in place with regard to the new challenges and actors. The specificities of social networks cannot justify a comprehensive delegation of law enforcement to social networks. State actors need to ensure that the policies they put into place produce a fair balance of rights of personality and honour and free speech rather than legal vacuums.

5.2 Multi-Player Solutions

Many players are involved in the sculpting of social network environments.[150] A superior model for law enforcement on these platforms must therefore not neglect their roles, above all the power relationship between international social media companies and public policy makers, for now mostly nation state governments. State responsibilities can be extended and assumed in cooperation with social networks, whose business models justify their participation in the costs of

142. *See* 4.2; for copyright S. Kulk and F.J.Z. Borgesius, 'Filtering for Copyright Enforcement in Europe after the Sabam Cases', 34 *EIPR* 791 (2012); E. Psychogiopoulou, 'Copyright Enforcement, Human Rights Protection and the Responsibilities of Internet Service Providers After Scarlet', 38 *EIPR* 552, at 555 (2012).
143. Frosio, above n. 141, at 27.
144. D.R. Johnson and D.G. Post, 'Law and Borders, The Rise of Law in Cyberspace', 48 *Stanford Law Review* 1367, at 1375 (1996); *see also* Barlow, above n. 4.
145. M. Wertheim, *The Pearly Gates of Cyberspace* (1999), at 228.
146. *See* Tsesis, above n. 79, at 864.
147. *But see* Johnson and Post, above n. 144, at 1375.
148. N. Alkiviadou, 'Regulating Internet Hate – A Flying Pig?', 7 *JIPITEC* 216, at 217 (2017).
149. *See* Tsesis, above n. 79, at 864.
150. *See* 3.

combating hate speech and fake news. Such multi-player solutions can combine the advantages of the strategic placement of social networks as points of control, while defending law enforcement and the exercise of fundamental rights as basic state task.

Propositions for a superior model of law enforcement can build upon existing concepts. The German NetzDG system already incorporates external assessors for non-obvious cases of UGC legality. While the NetzDG system relegates them to the role of in-house counsel, it shows that a cooperation with an external assessment body is possible. A similar cooperation could be envisioned as private-public partnership. Decisions on the legality of UGC could then be taken by ordinary judges. They possess the necessary expertise and enjoy independence and impartiality. In contrast to private (outside) institutions, their incentives are detached from the ones influencing social networks to overblock content. They would apply the law of the particular circumstances of the case and the fundamental rights affected; their decisions would be subject to review within the ordinary judicial system as well as constitutional review.

Within that proposed scheme, notifications regarding questionable UGC would thus be forwarded to public institutions responsible for the decisions on the take-down of questionable tweets, uploads and other UGC. A timely evaluation could be guaranteed just like swift judicial rulings are provided in the framework of interim legal protection. The referral to the competent judges can happen just as quickly as an in-house transmission. As well as in other contexts, specialised judges can rule within hours or days on the legality of the content, provided sufficient human resources are in place. Such a state intervention obviously requires the attribution of considerable government resources. Costs for this model of law enforcement would be incurred by the state rather than by social platforms as private entities. However, in light of the benefits drawn from the business models, social network responsibility can also be expressed in financial contributions to such a public-private partnership model. The overall cost for law enforcement in social networks would not change. Both public and private investments are worth making in light of the relevance of both social media in today's society and free speech as well as rights of personality in democratic state systems.